UFS·UV

UNIVERSITY OF THE **FREE STATE**
UNIVERSITEIT VAN DIE **VRYSTAAT**
YUNIVESITHI YA **FREISTATA**

Extending the Reach of Sequential Regression Multiple Imputation

by

Michael Johan von Maltitz

(2001029061)

A thesis submitted in fulfilment of the degree of

Doctor of Philosophy

in the

Faculty of Natural and Agricultural Sciences

Department of Mathematical Statistics and Actuarial Science

June 2015

Promoter: Prof. A. J. van der Merwe

Co-Promoters: Prof. T. E. Raghunathan & Prof. R. Schall

Declaration

I hereby declare that this thesis submitted for the Philosophae Doctor (PhD) degree at the University of the Free State is my own original work, and has not been submitted at another university/faculty for any reward, recognition, or degree. Every effort has been made to clearly reference the contributions of others within this thesis. I furthermore cede copyright of the thesis in favour of the University of the Free State.

This thesis was completed under the guidance of:

- Professors A.J. van der Merwe and R. Schall from the University of the Free State, Bloemfontein, South Africa; and
- Prof. T.E. Raghunathan from the University of Michigan, Ann Arbor, USA.

Michael J. von Maltitz

June, 2015

Acknowledgements

I would like to acknowledge the following people who helped me get through this seven-year adventure:

- First and foremost, I would like to thank my supervisors for their valuable input.
    - I would not have been able to venture into the Bayesian mathematics behind multiple imputation, sequential regression multiple imputation, the skew Student's t-distribution (including the priors on this distribution's degrees of freedom), and the calibrated posterior predictive p-value if it weren't for Prof. Abrie van der Merwe's tireless efforts. I would also like to thank Prof. Piet Groenewald for his commentary and programming associated with Prof. Abrie's mathematics.
    - I would like to thank Prof. Robert Schall for his input in editing this thesis, as well his advice on writing and reporting techniques for the publishable projects that this thesis is comprised of.
    - I would like to thank Prof. Trivellore Raghunathan for the valuable input he offered during his visit to Bloemfontein. The structure and direction of this thesis was guided by his expert input, and as a result, I feel that this work is truly a contribution to the research field that Prof. Raghunathan himself helped to build.
- I want to thank my colleague, Mr Sean van der Merwe, for all his input into the programming aspects of this thesis. When I would sit for hours stuck on a problem in one of my simulation algorithms, Sean would pop his head round my door, figure out what it was I was trying to accomplish, and solve my problems often enough in record time. You're a programming life-saver.
- I would like to thank my wife, Adelheid, for her faith in me. Even though this endeavour has stretched over seven years, she has never given up believing that this thesis was possible. Her encouragement over the last two years has been phenomenal, and without her, I surely would not have had the will to carry on with my research when it seemed to hit a dead-end between years four and five. Thank you, Love,

for making my life easy enough to be able to handle my non-research work and this enormous endeavour simultaneously.

- I am thankful to my parents for also having faith in me, and never even once thinking I might not be able to complete this thesis, no matter how many times over the last decade I told them "the research isn't going so well; I'm struggling with my latest simulation program".
- Finally, I would like to thank God for my brain, and the curiosity it has blessed me with. My brain has let me enjoy this PhD journey immensely; I look forward to continuing research in this field for as long as I am able.

Abstract

The purpose of this thesis is twofold. *Firstly*, it reviews a significant portion of literature concerning multiple imputation and, in particular, sequential regression multiple imputation, and summarises this information, thereby allowing a reader to gain in-depth knowledge of this research field. *Secondly*, the thesis delves into one particular novel topic in sequential regression multiple imputation. The latter objective, of course, is not truly possible without the former, since the deeper the review of multiple imputation, the more likely it will be to identify and solve pressing concerns in the sequential regression multiple imputation subfield.

The literature review will show that there is room in imputation research for work on a robust model for the sequential regression multiple imputation algorithm. This thesis pays particular attention to this robust model, formulating its estimation procedure within the context of sequential regression multiple imputation of continuous data, attempting to discover a statistic that would show when to use the robust model over the regular Normal specification, and then implementing the robust model in another estimation algorithm that might allow for better imputation of ordinal data.

This thesis contributes to ‘extending the reach of sequential regression multiple imputation’ in two ways. *Firstly*, it is my wish for users of public data sets, particularly in South Africa, to become familiar with the (now internationally standard) topics presented in the first half of this thesis. The only way to start publicising sequential regression multiple imputation in South Africa is to lay out the evidence for and against this procedure in a logical manner, so that any reader of this thesis might be able to understand the procedures for analysing multiply imputed data, or tackle one of the many research problems uncovered in this text. In this way, this thesis will extend the reach of sequential regression multiple imputation to many more South African researchers. *Secondly*, by working on a new robust model for use in the sequential regression multiple imputation algorithm, this thesis strengthens the sequential regression multiple imputation algorithm by extending its reach to incomplete data that is not necessarily Normally distributed, be it due to heavy tails, or inherent skewness, or both.

Key Terms

Incomplete data; multiple imputation; sequential regression multiple imputation; robust Bayesian regression model; skew Student t-distribution

Abstract (Afrikaans)

Die doel van hierdie tesis is tweeledig. *Eerstens*, gee dit 'n oorsig oor 'n beduidende gedeelte van die literatuur oor toerekening, en in die besonder, opeenvolgende regressie veelvuldige toerekening, en som hierdie inligting op, waardeur n leser in-diepte kennis van die navorsingsveld kan kry. *Tweedens*, die tesis vors 'n bepaalde nuwe onderwerp na in opeenvolgende regressie veelvuldige toerekening. Die laasgenoemde doelwit is natuurlik nie werklik moontlik sonder die voormalige nie, want hoe deegliker die oorsig oor veelvuldige toerekening, hoe meer waarskynlik sal dit wees om belangrike onderwerpe in die opeenvolgende regressie veelvuldige toerekening area te identifiseer en op te los.

Die literatuuroorsig sal wys dat daar ruimte in die navorsingsgebied oor toerekening is vir werk oor 'n robuuste model vir die opeenvolgende regressie veelvuldige toerekening algoritme. Hierdie tesis bestee besondere aandag aan hierdie robuuste model, naamlik die formulering van sy beramingsprosedure binne die konteks van opeenvolgende regressie veelvuldige toerekening van deurlopende data, en die tesis poog om 'n statistiek te vind wat aanwys wanneer die robuuste model moet gebruik word eerder as die gewone Normale spesifikasie; daarna word die robuuste model geimplementeer in 'n ander beramingsalgoritme wat moontlik ordinale data beter kan toereken.

Hierdie tesis dra by tot die 'uitbreiding van die aanreik van opeenvolgende regressie veelvuldige toerekening' op twee maniere. *Eerstens*, dit is my wens dat gebruikers van openbare data stelle, veral in Suid-Afrika, vertroud raak met die onderwerpe (wat nou die internasionale standaard is) wat in die eerste helfte van hierdie tesis hersien is. Die enigste manier om opeenvolgende regressie veelvuldige toerekening in Suid-Afrika bekend te stel is om sy voor- en nadele op 'n logiese manier uit te lê, sodat enige leser van hierdie tesis in staat kan wees om die prosedures vir die ontleding van vermeerderde toegerekende data te verstaan, of poging kan maak om een van die vele navorsingsprobleme wat in hierdie teks voorgestel is op te los. Op hierdie manier sal die tesis die rykwydte van opeenvolgende regressie veelvuldige toerekening uitbrei na baie meer Suid-Afrikaanse navorsers. *Tweedens*, deur te werk op 'n nuwe robuuste model vir gebruik in die opeenvolgende regressie veelvuldige toerekening algoritme, verbeter hierdie tesis die opeenvolgende regressie veelvuldige toerekening algoritme deur die uitbreiding van sy aanreik oor onvolledige data wat nie noodwendig Normaal versprei is, of dit nou te danke is aan swaar sterte van die verdeling, of innerlike skeefheid daarvan, of albei.

Description of content

The background and rationale for this thesis are given in Chapter 1. In Chapter 2 this thesis reviews multiple imputation, from its origin to several recent advances. The main advance in question, sequential regression multiple imputation, is reviewed in Chapter 3. The sequential regression multiple imputation algorithm's development, processes and recent advances are discussed in detail within that chapter. A new robust model for the sequential regression multiple imputation process is introduced and tested in Chapter 4. The question of sequential regression multiple imputation evaluation is then discussed in Chapter 5, with the goal of identifying when the new robust model should be used instead of the traditional Normal model. An additional application in sequential regression multiple imputation for the robust model introduced in this thesis is then evaluated in Chapter 6. Chapter 7 will conclude this thesis. In this thesis, Chapters 1 to 3 represent a review of previous literature, while the chapters thereafter contain new, original work.

List of acronyms

- ABB – approximate Bayesian bootstrap
- BB – Bayesian bootstrap
- CC – complete-case
- CSR – cubic spline regression
- ERD – expanded residual draw
- EM – expectation maximisation
- FCS – fully conditional specification
- GAMLSS – generalised additive models for location, scale and shape
- GLM – generalised linear model
- GOF – goodness of fit
- HD – hot-deck
- INC – incomplete (data)
- KS – Kolmogorov-Smirnov, as in the KS test
- LOWESS – locally weighted scatterplot smoothing
- LM – linear model
- LRD – local residual draw
- MAD – mean absolute deviation
- MAR – missing at random; one of the three MDMs
- MARMID – MAR MDM with more missing data in the centre of the distribution
- MARRIGHT – MAR MDM with more missing data in larger values
- MARTAIL – MAR MDM with more missing data in the tails of the distribution
- MCAR – missing completely at random; one of the three MDMs
- MCMC – Markov Chain Monte Carlo
- MDM – missing data mechanism, *i.e.* MCAR, MAR or MNAR
- MI – multiple imputation
- MICE – multiple imputation through chained equations
- ML – maximum likelihood
- MN – multivariate Normal
- MNAR – missing not at random; one of the three MDMs

- MSE – mean squared error
- NMV – Normal method, adjusting for mean and variance
- OLS – ordinary least squares
- PM – predictive model
- PMM – predictive mean matching
- PS – propensity score
- QQ – quantile-quantile, as in QQ plots
- $RBIAS$ – relative bias
- RMSE – root of the mean squared error
- ROC – receiver operating characteristic, as in ROC curve
- $RRMSE$ – root of the relative mean squared error
- SI – single imputation
- SIR – sampling importance resampling
- S-HD – single hot-deck
- SRMI – sequential regression multiple imputation

Mathematical notations

- $Y^{(N)} = N \times p$ (population) matrix of partially observed outcome variables
- $Y_{inc}^{(N)}$ (population) matrix of outcome variables included in the survey
- $Y_{obs}^{(N)}$ (population) matrix of outcome variables that are observed
- $X^{(N)} = N \times q$ (population) matrix of fully observed covariates
- $I^{(N)} = N \times p$ (population) indicator matrix for inclusion of $Y^{(N)}$ in survey
- $R^{(N)} = N \times p$ (population) indicator matrix for response on $Y^{(N)}$. In $R^{(N)}$ there are ones in the positions of the missing data entries of the complete data matrix, and zeroes elsewhere
- $R_{inc}^{(N)}$ (population) indicator matrix for response on $Y^{(N)}$, but only that part of $R^{(N)}$ for the data that is included in the survey
- $Y = n \times p$ matrix of the incomplete part of a data set
- Y_j j^{th} variable of the data set, Y
- Y_{-j} The entire Y data set excluding the j^{th} variable of the data set
- $Y_{com} = \left\{X^{(N)}, Y_{inc}^{(N)}, I^{(N)}\right\} = n \times p$ matrix representing the complete (but not observed) version of the incomplete Y
- $Y_{obs} = \left\{X^{(N)}, Y_{obs}^{(N)}, I^{(N)}, R_{inc}^{(N)} =\right\} n \times p$ matrix of the observed part of Y
- y^{obs}, a fully-observed dependent variable (Chapter 5)
- $Y_{mis} = n \times p$ matrix of the missing part of Y
- $X = n \times q$ matrix of fully observed covariates in a data set
- $X_{obs} = n \times q$ matrix, is that part of X that corresponds to Y_{obs}
- $X_{mis} = n \times q$ matrix, is that part of X that corresponds to Y_{mis}
- n_{obs} is the number of cases observed in the variable of interest
- n_{mis} is the number of cases missing in the variable of interest
- $R \equiv \{r_{ij}\} = n \times p$ indicator matrix of response on Y. In R there are ones in the positions of the missing data entries of the complete data matrix, and zeroes elsewhere. The distribution of R, known as the missing data mechanism (or MDM following the acronyms in this section section), is $\Pr(R|Y_{com}, \theta)$, where θ is a vector of unknown parameters.

Contents

List of Figures

Chapter 1

Introduction

One of the major issues associated with large surveys is that of non-response or lost data — missing survey data that is almost always multivariate in nature. Moreover, the main problematic issue regarding missing data is that most data analysis procedures are not designed to handle them, leading to analyses that conclude invalid and inefficient inferences about a population (Schafer & Graham 2002). Many economic analyses use either complete-case analysis or a simple method of imputing missing data, such as single imputations. Even if simple single imputations are accurate, they almost certainly do not capture the imputer's inherent uncertainty in the guesswork involved. This is one of the reasons for the development of multiple imputation. The science of multiple imputation has evolved in such a way as to be able to remove the onus of imputing from the analyst (Meng 1994), so that public-use data sets can be prepared by imputation experts and offered to experts in the analysis arena without substantial loss in estimation validity and/or efficiency, provided the guidelines for the use of multiply imputed data are followed. Bayesian statistics, in particular, seem to be particularly well-suited for imputation (Meng 1994), since the unknown values can be modelled directly given the known and explicit model parameters.

1.1 Background: Incomplete Data

It is important to lay the groundwork under the concept of incomplete data and the non-imputation and imputation procedures used to solve incomplete data problems, so that any reader not familiar with the topic can be guided into the field before being bombarded with the technical aspects presented in Chapters 2 and 3. In essence, several topics need to be introduced, namely, the concept of incomplete data and how data becomes incomplete, how incomplete data is dealt with by statisticians.

1.1.1 What is incomplete data?

Incomplete data refers to any data set that contains missing values within one or more of the variables within that data set. If the missing data points are found only within a single variable, then the problem is univariate, but, as mentioned before, most often missing values are spread across several variables within a data set. These missing values are often the result of subjects not responding to certain questions in a survey questionnaire in the cross-sectional case, or subjects dropping out of a study in the longitudinal case. However, missing values can also be the result of several other factors, for example, anomalous responses that are deleted.

In the statistical world, it is assumed that some random process causes data to become missing. This process is known as the missing data mechanism. In brief, there are three mechanisms by which data is said to be missing — 'missing at random' (MAR), 'missing completely at random' (MCAR), or 'missing not at random' (MNAR). In the MAR mechanism, the distribution of positions of the missing data entries is assumed to be independent of the missing data in the analysis, or $\Pr(R|Y_{com},\theta) = \Pr(R|Y_{obs},\theta)$, where R is the missing data mechanism, Y_{com} is the theoretical complete data set, θ is the unknown parameter of the data, and Y_{obs} is the observed part of the data. In the case of MCAR, a special version of the MAR mechanism, the positions of the missing data entries are assumed to be independent of all of the variables in the analysis, *i.e.* $\Pr(R|Y_{com},\theta) = \Pr(R|\theta)$, using the same notations as before. Of course, this implies that the missing entries are entirely independently randomly missing. In the last case, the

MNAR missing data mechanism, the positions of the missing data entries are assumed to be at least dependent on data that is missing from the data set, or, more basically, the distribution of missingness is not MAR. Once again using the same notations, this means that for MNAR, $\Pr(R|Y_{com}, \theta) \neq \Pr(R|Y_{obs}, \theta)$.

1.1.2 Patterns of missing data

The missing data in an incomplete data set can form one of several different patterns. These can be categorised into two main groups, namely, a monotone pattern and a general, or non-monotone pattern.

A monotone pattern exists if the variables can be ordered such that, for each observation in the data set, all previous variables are observed if a later variable is observed (*i.e.*, if a variable is observed for a particular row or case, all variables to the left of that variable are also observed for that case). This pattern often exists in longitudinal studies when patients drop out and do not return to the study.

Any pattern that is not a monotone form is a general pattern, or non-monotone pattern. Specific types of general patterns do exist, however. One example is the file matching pattern, in which a set of incomplete variables only has a single observed quantity per observation.

1.2 Review of the methods of handling incomplete data

1.2.1 Complete-case analysis on incomplete data

In complete-case analysis, only cases containing values for each of the variables in the analysis are retained in the analysis procedures. This can raise the problem of serious bias in the analysis if the data is originally incomplete (Little & Rubin 2002), including problems relating to invalid and/or inefficient estimates. For example, suppose that a survey is taken over rural and urban households, both poor and non-poor, and is designed to measure the amount of total years of formal education of the head of the household and the monthly income of the household. One would expect a positive relationship between income and education a priori. Moreover, for urban households, education may fluctuate over a wide interval from poorly educated to well-educated (being closer to a wider range of institutions able to offer a formal education). For rural households, however, one would expect formal education to be more of a luxury, and so values for this variable would be much lower in these households. Additionally, for rural households, where income is traditionally lower and more difficult to determine, the probability of non-response in the income variable may be higher. A complete-case analysis of all the household data will then possibly drop a large number of rural observations (due to missing income) and the strength of the positive relationship between education and income could be understated. Hence, a bias can exist in the analysis. Additionally, loss of sampled units increases variance and therefore leads to inefficient estimates. One must note, however, that these possible biases may not always exist in complete-case analysis, but rather that the extent of bias will depend on the mechanism by which data is deemed to be missing. Particularly, if the data is missing completely at random (MCAR), then there will be no bias in complete-case analysis of multivariate data with missing entries (Schafer 2003). This is logical, since if data is missing completely at random, any incomplete cases dropped from the complete-case analysis can be thought of as sampled units dropped in a second random stage of sampling. However, even if this is the case, the resulting inferences from this list-wise deletion may be inefficient, since the sample size is reduced, essentially unnecessarily.

To overcome the possible biases in complete-case analysis, many methods of dealing with incomplete data have been suggested. These methods are divided into two camps — the non-imputation procedures, and the imputation procedures. The former is introduced in the following subsection, while the latter is again divided into two main fields, namely single imputation and multiple imputation. These are introduced in Subsections 1.2.3 and 1.2.5.

1.2.2 Non-imputation procedures

The non-imputation methods of handling incomplete data include available-case analysis, weighting (or re-weighting) procedures, indicator methods, and model-based procedures.

Available-case analysis. Available-case analysis estimates different parameters of interest using different subsets of the data set, basically creating estimates of interest according to the data that is available. While using all the available data is sensible, analytical procedures are difficult to perform under these circumstances (Schafer & Graham 2002).

Re-weighting. The re-weighting procedure drops incomplete cases and assigns weights to the remaining observations (determined by additional or auxiliary variables) so that the remaining cases more accurately reflect the distribution of the complete data. In this way generalised estimating equations can be modified to provide valid inferences when the missing data mechanism is MCAR or MAR (Kenward & Carpenter 2007). However, "[w]eighting can eliminate bias due to differential response related to the variables used to model the response probabilities, but it cannot correct for biases related to variables that are unused or unmeasured" (Schafer & Graham 2002, p.157). In other words, if the probability of response is determined by unmeasured variables, which is entirely possible, then this method becomes less attractive.

Additionally, re-weighting complete observations so that they are representative of the population sampled implies calculating weights which are generated from estimated probabilities of non-response. These estimated probabilities are inferred from the data or from auxiliary variables, but the overall weighted analysis often does not include an uncertainty component due to the estimation of these probabilities from the data.

Indicator method. In the indicator method, summarised by Brand (1998), for each incomplete independent variable x_j, the regression term $\beta_j x_j$ can simply be replaced by $\beta_{0j} x_j (1 - R_j) + \beta_j R_j x_j$, where R_j is the response indicator of x_j. This procedure simply adjusts the intercept when the value is missing, and as, such, can lead to biased estimates under a number of conditions. A better replacement would be $\beta_{0j} x_j (1 - R_j) + R_j \beta_j x_j +$

$\sum_{k \in mis; k \neq j} R_j(1 - R_k)\beta_{jk}x_j$, but this method increases the number of parameters greatly, and, therefore, may not be more efficient than list-wise deletion (Brand 1998).

Maximum likelihood. Rubin (1976) introduces this maximum likelihood estimation procedure that integrates out missing data. Schafer & Graham (2002, p.162) mention that, "[u]nder MAR, the marginal distribution of the observed data... provides the correct likelihood for the unknown parameters [of the data,] θ, provided that the model for the complete data is realistic". Thus, even the maximum likelihood (ML) method of imputation can suffer from serious drawbacks — a lack of robustness in estimates when the model deviates from the fully parametric model assumed for the complete data, and the fact that the ML method needs a large sample for ML estimates to be approximately unbiased and Normally distributed (Schafer & Graham 2002).

In essence, however, the maximum likelihood procedure for incomplete data follows a simple ideal that itself has merit enough to warrant a basic understanding of the process. Brand (1998, pp.36–37) summarises the process well:

> "[I]mpute the missing data entries on the basis of an initial estimate $\theta^{(0)}$ or θ; re-estimate θ from the completed data $\theta^{(1)}$; use $\theta^{(1)}$ to re-impute the missing data entries; use the re-completed data to re-estimate θ by $\theta^{(2)}$; and so on; iterate this until $\theta^{(t)}$ converges. Each iteration of [this so-called] EM [algorithm] consists of an E-step (Expectation) and an M-step (Maximi[s]ation). In the E-step, the expected complete data log-likelihood $Q(\theta|\theta^{(t)}) = E[l(\theta|y)|y_{obs}, \theta = \theta^{(t)}]$ is estimated from the current estimate $\theta^{(t)}$. In the M-step, a new estimate $\theta^{(t+1)}$ is found, which maximi[s]es the expected complete data likelihood $Q(\theta|\theta^{(t)})$ from the previous E-step for θ. In fact, in the E-step, not the missing data y_{mis} but functions of y_{mis} on which the complete data log-likelihood $l(\theta|y)$ depends are estimated... For special statistical models within the [E]xponential family, such as the multivariate [N]ormal model, the M-step is similar to the MLE for complete data."

Additional drawbacks mentioned by Brand (1998) include the fact that convergence of EM can be very slow in cases where there is a large proportion of missing data, that convergence to a global maximum is not guaranteed, that standard errors and correlation matrices of point estimates are not directly available from EM and their calculation can be complicated, that ML is designed for large samples and has limitations for small samples, and that EM requires statistical expertise.

Extending the ML method by including priors to formulate posteriors alleviates the inconvenience of the large samples being required. However, these posteriors may be extremely complex and may require numerical integration or Monte Carlo techniques in order to solve them, similar to the final non-imputation method presented below.

Integration. One could attempt to integrate out the missing data within an incomplete data set, in a way that is summarised by Carpenter & Kenward (2007). Suppose that we divide the incomplete data set into outcome variables, Y, and covariates, X. Let R be the missing data mechanism. Our initial model is $f(Y, X, R) = f(Y, X)\,\Pr(R|Y, X)$. If the missing data mechanism is MCAR or MAR, then it is ignorable[1], and if the overall analysis model and the model for missingness share no parameter space, we can integrate over the missing observation outcomes and covariates, Y_{mis} and X_{mis}, as follows:

$$\begin{aligned} & f\left(Y_{obs}, X_{obs}\right) \\ &\quad = f\left(Y_{obs}|X_{obs}\right) f\left(X_{obs}\right) \\ &\quad = \int\!\!\int f\left(Y_{obs}, Y_{mis}|X_{obs}, X_{mis}\right) f\left(X_{mis}|X_{obs}\right) f\left(X_{obs}\right) dX_{mis} dY_{mis}, \end{aligned}$$

where Y_{obs} and X_{obs} are the observed parts of the outcome and covariate matrices, respectively.

This integration is often analytically intractable, and therefore many methods have been developed and applied to tackle the problem, including the expectation-maximisation (EM) algorithm, Monte Carlo Newton Raphson and Monte Carlo likelihood, mean score methods, and fully Bayesian methods based on Markov Chain Monte Carlo (MCMC) modelling (Kenward & Carpenter 2007).

[1] The concept of ignorability is fully discussed in Chapter 2.

1.2.3 Single imputation before complete-case analysis

Alternatively, if complete-case analysis methods are to be used on an data set that is originally incomplete, data can be filled in by several single imputation procedures, including substitution, cold-deck imputation, unconditional and conditional mean (or mean/mode) substitution, imputation from unconditional distributions or (single) hot-deck imputation, and imputation from conditional distributions. The term 'single' in the concept, 'single imputation methods' implies imputing only one value for each missing datum.

Substitution. Substitution, occurring at the fieldwork stage of a survey, substitutes non-respondents with respondents not originally selected for interview. Again, possible bias may exist in parameters drawn from analysis if the non-respondents differ systematically from the respondents.

Cold-decking. Cold-deck imputation substitutes missing values with values from outside the current data set, such as a previous wave of the current survey. As with substitution, possible bias may exist due to a systematic difference between non-respondents and the respondents from which the imputed values are taken.

Unconditional mean substitution. Unconditional mean substitution simply replaces a missing value in a variable with the mean of the available data for that variable. While the means of the variables will be preserved by this process, the standard errors will be reduced, leading to parameter estimates that seem more significant than they actually are. A variation of mean substitution is mean/mode substitution. The difference between these two methods lies in their handling of categorical variables. For mean substitution the mean of the corresponding indicator variables created from a categorical variable is used, whereas in mean/mode substitution the mode of the categorical variables are used for imputations.

Conditional mean substitution. Conditional mean substitution regresses the complete part of a variable on other variables and predicts values for the incomplete part of that variable. The missing values are imputed using the fitted values from the regression model. This method is not recommended for analysis of covariances or correlations, as the strengths of the relationships between the imputation-filled variables and the rest of the data set are overstated.

Hot-decking. Imputation from unconditional distributions (hot-decking) chooses a value for the missing entries in a variable from the observed values of that variable. In this case bias in analysis on the completed data is still possible, but it is more likely to occur in regressions equations based on the completed data than in measures of central tendency (Saunders, Morrow-Howell, Spitznagel, Doré, Proctor & Pescarino 2006). Ardington, Lam, Leibbrandt & Welch (2006) also correctly point out that observed outliers or anomalies can affect the analyses more than they should be allowed to do so since any outlier or anomaly has a chance of being drawn to replace a missing value.

Imputation from conditional distributions. Imputation from conditional distributions implies simulating a draw from the distribution $\Pr(Y_{mis}|Y_{obs},\theta) = \Pr(Y_{obs},Y_{mis},\theta)\,/\,\Pr(Y_{obs}|\theta)$, where θ is again the unknown parameter of the data. Since θ is unknown, an estimate of θ, $\hat{\theta}$, must be made from Y_{obs}, after which a draw can be made from $\Pr(Y_{mis}|Y_{obs},\hat{\theta})$. This method requires a correctly specified model for $\Pr(Y_{mis}|Y_{obs},\theta)$, but if this is the case it will produce "nearly unbiased estimates for many population quantities under [the] MAR [mechanism]" (Schafer & Graham 2002, p. 159). For more on these procedures, both imputation and non-imputation, see Little & Rubin (2002) and Schafer & Graham (2002).

1.2.4 Requirements for good imputations

Incomplete data problems generally require a solution that has the following capabilities, according to Rubin (1987, p. 11). *Firstly*, it should be possible to utilise standard complete-data analysis methods on the data sets that have been filled in. *Secondly*, the imputation technique and adjustments to the follow-up analysis should yield valid inferences that produce both estimates that adjust for observed differences between respondents and non-respondents and standard errors of these estimates that reflect the reduced sample size and an adjustment for observed differences between respondents and non-respondents. *Finally*, the multiple imputation technique should display the sensitivity of inferences to various plausible models for nonresponse. These are the guidelines that will be used to judge imputation methods within this thesis. As Meng (1994, p. 538) puts it, "[f]rom an inferential point of view, perhaps the most fundamental reason for imputation is that a data collector's assessment and information about the data, both observed and unobserved, can be incorporated into the imputations."

The single imputation methods mentioned in this section have the advantage of allowing existing complete-case analysis methods to be used on the filled-in data set. Additionally, the imputer's knowledge can be incorporated into the imputation procedure. The drawbacks, however, are that the complete-data methods that will be used assume the imputed values are known. This means that inferences based on the data are systematically sharper than they should be, and quantities based on variability (e.g. correlations) can be biased. Moreover, if the nonresponse mechanisms are not understood, no accommodation is being made for the uncertainty of not knowing which nonresponse models for imputation are appropriate. In essence, the single imputation procedures only sufficiently adhere to one of the three properties needed from a solution to incomplete data problems (Rubin 1987).

1.2.5 Multiple Imputation

Multiple imputation covers a broad category of methods of imputation that impute several plausible values for each missing value in a data set. Rubin (1978) mentions that the interest in multiple imputation may have grown for three reasons:

1. Surveys seemed to be suffering more and more from nonresponse;

2. There was a growing awareness that the existing standard methods of handling nonresponse were unsatisfactory; and,

3. Both mathematically and computationally, this topic was proving to be a rich statistical research area.

From the start of the development of his methods (which, in time, have been proved to be the fundamental groundwork for this entire research area), Rubin's (1978, p. 20) objective was to build "statistically sound tools for handling nonresponse in general purpose surveys", and to "be concerned with both *theoretical appropriateness* and *practical utility*" [emphasis added]. From these statements we can see that the aim of multiple imputation was two-fold from the beginning: statistical appropriateness *and* tractability, in particular, for application in survey nonresponse. Rubin continues his explanation of the guidelines under which he developed his methods, writing that "handling nonresponse must mean displaying how different the answers from the surveys might have been if the nonrespondents had responded" (Rubin 1978, p. 20). At this early stage in the development of multiple imputation, Rubin already knew that key to this research area would be to separate the analysis task from the imputation task. This is evident when he writes, "in multipurpose surveys, some form of imputation is just about the only practically possible method for handling nonresponse, because the data set will be used to address many questions now and in the future. Remodelling the missing data process each time a new question is to be asked of the data base seems to be impractical, while creating an imputed data set is quite practical" (Rubin 1978, p. 21).

Multiple imputation is viewed as a flexible alternative to likelihood methods for a range of incomplete data problems (Schafer & Graham 2002), as well as a range of nonresponse models. As in single imputation, the knowledge of the imputer can be incorporated into the imputation procedure, and, once the imputations have been completed, complete-case analysis procedures can be used to analyse the data. However, the primary advantage of multiple imputation is the inflation of uncertainty in the analysis estimates. In essence, multiple imputation covers a class of methods that impute several plausible values for a single missing data entry. Once the missing values have been imputed, several completed data sets are left to be analysed by complete-case methods. A simple set of rules is then used to combine the estimates from the separate analyses of the several data sets, and the uncertainty of these estimates is then formed from the sample variation as well as variation in the imputed values themselves. So the estimates derived from multiple imputation adjust for observed differences between respondents and nonrespondents and the standard errors of these estimates reflect the reduced sample size and an adjustment for observed differences between respondents and nonrespondents. Hence multiple imputation methods technically adhere to all three of the guidelines set out by Rubin (1987).

Multiple imputation can easily be linked to Bayesian statistics, as the imputed values for a single missing data entry can be draws from the posterior predictive distribution for the non-missing data. Additionally, in the quest for parsimonious Bayesian algorithms for multiple imputation, several advances in methodology have been made since Rubin's (1978)'s fundamental multiple imputation concepts were established. This thesis is primarily focused on one particular multiple imputation evolution, that being sequential regression multiple imputation. Multiple imputation in general, will be covered in Chapter 2, while the more recent adaptation, sequential regression multiple imputation, will be discussed in Chapter 3.

The drawbacks of multiple imputation, according to Rubin (1987), are as follows: that more work is required to produce multiple imputations rather than single imputations; that more space is needed to store multiply-imputed data sets; and that more work is needed to analyse multiply-imputed data sets. However, with the advance in computing power in modern times, these three disadvantages pale in comparison to the advantages of applying multiple imputations in solving incomplete-data problems, even if we choose to create a large number of multiply-imputed data sets for each incomplete-data problem.

1.3 Objectives

With the groundwork laid concerning the traditional measures of handling incomplete data, the objectives of this thesis should be brought to light. The methods for handling incomplete data introduced above clearly suggest that the only sensible way to produce complete data from incomplete data in a variety of contexts, is to use a form of multiple imputation.

The *first* objective of this thesis, therefore, is to review the literature on multiple imputation, and in particular, the later adaptations of multiple imputation, to be able to provide a primer for any reader on the topic of multiple imputation, and to identify the key research areas in which additional work will be beneficial.

The *second* objective, which can only be completed after a thorough review of the literature on multiple imputation, is to contribute to one of the identified key research areas within multiple imputation, by attempting to solve a set of related unsolved problems under the topic of multiple imputation. The key research area under consideration is sequential regression multiple imputation.

Both thesis objectives seek to "extend the reach of sequential regression multiple imputation". By thoroughly reviewing the evolution of multiple imputation to sequential regression multiple imputation, the reader will be familiarised with the methodology, benefits, and uses of multiple imputation in general, and sequential regression multiple imputation in particular. The reader can use this thesis as a primer on multiple imputation and sequential regression multiple imputation, in order to contribute to these fields of research, or simply just to know how to handle multiple imputed data. The latter is particularly important in South Africa, where the crude, non-multiple-imputation measures mentioned in this chapter are still being used extensively. The second objective of this thesis seeks to "extend the reach of sequential regression multiple imputation", but this time, in a more literal manner. As will be shown in the literature review, one area of research within sequential regression multiple imputation (and the focus area of this thesis) is the use of robust alternatives to the standard models. These robust alternatives allow sequential regression multiple imputation to reach a wider array of incomplete data models. Moreover, these robust models allow for incomplete data with longer distribution tails than are usually controlled for, thereby literally "extending the reach of sequential regression multiple imputation".

Chapter 2

Multiple Imputation

2.1 Introduction

Multiple imputation (MI) was first proposed by Donald Rubin in the 1970's as one solution to survey nonresponse problems. Rubin (1978) suggested that guidelines be established for imputers to be able to follow, rather than having imputers create *ad hoc* measures to solve the nonresponse problem every time it arose. Rubin (1978) also mentions that a goal of, or the plan for MI, was that the imputed values reflect the variation within an imputation model as well as sensitivity to different imputation models, and that the analysis of the resultant multiply-imputed data be viewed as simulating predictive distributions of desired summary statistics under imputation models. The entire process behind MI and analysis is then divided into three areas, namely the modelling task (specifying a hypothetical joint distribution), the imputation task (deriving a predictive posterior distribution for the incomplete variable(s)), and the analysis task (estimating parameters of interest from the completed data). Restrictions inherently exist in MI problems, namely statistical appropriateness and tractability or practicability. Additionally, tractability can be emphasised in the way that MI was engineered to split the imputation task from the analysis task.

The important early advances in this field of research will be reviewed in detail in Section 2.2, while the more recent advances in the field will be reviewed in Section 2.3.

2.2 Early Multiple Imputation

The early years in the development of MI saw some profound results. The first problem that had to be dealt with was the the confounding process causing the missing data to appear the way it does, since its inclusion in the modelling procedure made matters more complicated. Research into the implications behind ignoring this MDM, and the situations in which this was an acceptable practice, laid the groundwork into opening up the MI research field to many new advances, including the standard multiple data set estimate combining rules, analyses of the validity and efficiency of the estimates generated from these rules, re-weighted combining rules, and, the primary topic of interest in this thesis sequential regression multiple imputation, amongst others.

2.2.1 Ignorable processes for causing missing data

In general, Rubin's (1976) work shows that one may ignore the process that causes the missing data if the missing data are missing at random and the observed data are observed at random. These two conditions together imply that the data is missing completely at random (MCAR) (Little & Rubin 2002). In a Bayesian context, these requirements are slightly adjusted, as will be explained below.

Rubin (1976) defines missing data to be missing at random if, for each possible value of the parameter ϕ of the MDM, $g_\phi(\tilde{R}|\tilde{Y})$, the conditional probability of the observed pattern of missing data, given the missing data and the observed data, is the same for all possible values of the missing data. The tilde in the formulations represents the observed matrices. So no matter what the missing values are or could be, we will still see the same pattern of missing data if the data is MAR. Rubin (1976) defines data to be observed at random if, for each possible value of the missing data and ϕ, the conditional probability of the observed pattern of missing data, given the missing data and the observed data, *i.e.* $g_\phi(\tilde{R}|Y)$, is the same for all possible values of the observed data. Rubin (1976) also defines the parameter for the MDM, ϕ, and the parameter for the data, θ, as distinct if their joint parameter space factorises into a space for ϕ and a space for θ, and when prior distributions are specified for these two parameters, if these parameters are independent.

Zhang (2003) elaborates, explaining that the parameters are distinct if:

1. From a frequentist perspective, the joint parameter space of (θ, ϕ) is the Cartesian cross-product of the parameter spaces for θ and ϕ;

2. From a Bayesian perspective, the joint prior distribution of (θ, ϕ) can be factored into independent marginal prior distributions for θ and ϕ.

Rubin's (1976) objective was to use $\tilde{Y}$, or alternatively, $\tilde{R}$ and $\tilde{Y}_{obs}$ to make inferences about θ. This essentially implies ignoring the process that causes missing data by fixing the random variable R at the observed pattern of missing data $\tilde{R}$ and assuming that the values of the observed data, $\tilde{Y}_{obs}$, arose from the marginal density of the random variable Y_{obs}:

$$\int f_\theta \left(Y\right) dY_{mis} \tag{2.1}$$

The question was whether this process would imply valid inferences about θ. In fact, fixing the random variable R at the observed pattern of missing data $\tilde{R}$, the sampling distribution of a statistic based on the observed data, $S(\tilde{Y}_{obs})$, is the conditional density of Y_{obs} given $R = \tilde{R}$:

$$\int f_\theta \left(Y\right) \frac{g_\phi\left(\tilde{R}|Y\right)}{k_{\theta,\phi}\left(\tilde{R}\right)} dY_{mis} \tag{2.2}$$

where $k_{\theta,\phi}(\tilde{R}) = \int f_\theta(Y) g_\phi(\tilde{R}|Y)$, the marginal probability that $R = \tilde{R}$. This means that the correct sampling distribution of $S(\tilde{Y})$ depends in general not only on the fixed hypothesised f_θ, but also on the fixed hypothesised g_ϕ. However, Rubin (1976) goes on to mention that if the missing data is MCAR, $g_\phi(\tilde{R}|Y)$ takes on the same value for all Y. Hence, $k_{\theta,\phi}(\tilde{R}) = g_\phi(\tilde{R}|Y)$, and thus the distribution of every statistic under the density (2.1) is the same as under the density (2.2).

Moreover, the sampling distribution of $S(\tilde{Y})$ under f_θ calculated by ignoring the MDM equals the correct conditional sampling distribution of $S(\tilde{Y})$ given $\tilde{R}$ under $f_\theta g_\phi$ for every $S(\tilde{Y})$, if and only if $E_{Y_{mis}}[g_\phi(\tilde{R}|Y)|\tilde{R}, Y_{obs}, \theta, \phi] = k_{\theta,\phi}(\tilde{R}) > 1$, and, the sampling distribution of $S(\tilde{Y})$ under f_θ calculated by ignoring the MDM equals the correct *un*conditional sampling distribution of $S(\tilde{Y})$ given $\tilde{R}$ under $f_\theta g_\phi$ for every $S(\tilde{Y})$, if and only if $g_\phi(\tilde{R}|Y) = 1$. For more information concerning these two theorems, see Rubin (1976, p. 585).

In essence, ignoring the MDM and making inferences about the underlying parameter of the data, θ, means comparing the estimator from the observed data (given the missing data pattern) to the estimator from the marginal distribution of the observed data (ignoring the MDM). As soon as one fixes which data are missing, the sampling distribution of the observed data follows the conditional density of the observed data given the pattern of missing data that was observed. Rubin (1976) shows that this conditional density is the same as the marginal density (ignoring the missing data mechanism) if the missing data are MAR and the observed data are observed at random. However, the resulting inferences are conditional on the pattern of missing data that was observed; the densities are equal to the correct sampling density if the pattern of missing data is assumed to be the same regardless of the parameter of the process causing the data to be missing. So if the MDM is ignorable, valid inferences about the population parameter can be made from unconditional distributions of the observed data.

In Rubin's (1976) paper, this process is also analysed in a Bayesian context (as well as a direct likelihood approach, which will not be revisited here). In a Bayesian context, θ and ϕ are random variables whose marginal distribution is specified by the product of the priors, $p(\theta)p(\phi|\theta)$. If we ignore the MDM, we choose $p(\theta)$ and assume that $\tilde{Y}_{obs}$ arises from density (2.1). So the posterior distribution of θ ignoring the MDM is proportional to:

$$p\left(\theta\right)\int f_\theta\left(\tilde{Y}\right)dY_{mis} \tag{2.3}$$

However, we are fixing $R = \tilde{R}$ without conditioning on it in posterior (2.3). The correct conditional posterior distribution is indeed proportional to:

$$p\left(\theta\right)p\left(\phi|\theta\right)\int f_\theta\left(\tilde{Y}\right)g_\phi\left(\tilde{R}|\tilde{Y}\right)dY_{mis} \tag{2.4}$$

However, if the data are MAR, and θ and ϕ are distinct, then:

$$\begin{aligned} p(\theta)\, p(\phi|\theta) \int f_\theta\left(\tilde{Y}\right) g_\phi\left(\tilde{R}|\tilde{Y}\right) dY_{mis} = \\ \left[p(\theta) \int f_\theta\left(\tilde{Y}\right) dY_{mis}\right] \left[p(\phi) g_\phi\left(\tilde{R}|\tilde{Y}\right) dY_{mis}\right] \end{aligned}$$

So, the posterior distribution of θ ignoring the process that causes missing data equals the correct posterior distribution of θ, and the posterior distributions of θ and ϕ are independent. Rubin (1976) also shows that the posterior distribution of θ ignoring the process that causes missing data equals the correct posterior distribution of θ if and only if $E_{\phi,Y_{mis}}\left[g_\phi\left(\tilde{R}|\tilde{Y}\right)|\tilde{R},\tilde{Y}_{obs},\theta\right]$ takes on a constant positive value, since the posterior distribution of θ is proportional to posterior (2.4) integrated over ϕ, *i.e.*

$$\begin{aligned} &\left[p(\theta)\, f_\theta\left(\tilde{Y}\right) dY_{mis}\right] \int E_{Y_{mis}}\left[g_\phi\left(\tilde{R}|\tilde{Y}\right)|\tilde{R},\tilde{Y}_{obs},\theta,\phi\right] p(\phi|\theta)\, d\phi \\ =\ &\left[p(\theta)\, f_\theta\left(\tilde{Y}\right) dY_{mis}\right] E_{\phi,Y_{mis}}\left[g_\phi\left(\tilde{R}|\tilde{Y}\right)|\tilde{R},\tilde{Y}_{obs},\theta\right] \\ \propto\ & p(\theta) \int f_\theta\left(\tilde{Y}\right) dY_{mis} \end{aligned}$$

To summarise the Bayesian aspect of this research, if the data are MAR and the parameters of the MDM and the overall data are distinct, the joint posterior distribution of the parameters of the data and the MDM is the same as the correct posterior of the parameter of the data. These distributions are the correct posterior distributions for the parameter of the data if and only if the conditional expectation of the MDM, given the pattern of missing data, the observed data and the underlying parameter for the data, is a constant positive value. Note that the missing data need not be MAR *and* the observed data be observed at random (*i.e.* we need not have missing data that is MCAR), but rather only that the data are MAR and the parameters of the missing data process and the overall data are distinct.

This work has boiled down to a very useful fact: "When response is unrelated to values of missing variables within subgroups defined by observed covariates, the non-response is called ignorable" (Glynn, Laird & Rubin 1993, p. 984). Rubin (1978, p. 21) simplifies this idea even more, stating that "when mechanisms used to sample units and record data are known (possibly probabilistic) functions of recorded values, the mechanisms are said to be ignorable." It is more important to note, however, that if the MDM is ignorable and is ignored, the inferences based on the observed data are valid inferences concerning the original population parameter.

2.2.2 Nonignorable process for causing missing data

Having reviewed the case of ignorable MDMs, it seems justified at least to glance at the problems that are inherent in nonignorable MDMs. The topic of nonignorable MDMs has come into focus recently, and as such, studies relating to this research field are reviewed in Section 2.3.

The main problem associated with assuming a certain MDM is that a data set that is being filled-in through MI cannot provide evidence to suggest that the MDM is nonignorable. Nonignorable mechanisms, are, by the very definition of nonignorability, mechanisms in which missingness is related to that which is unobserved. As Kenward & Carpenter (2007) put it, the choice between a MAR and a MNAR MDM is untestable. A researcher could rather choose several mechanisms for their problem, and analyse results under each mechanism as a sensitivity analysis. Thus, the problem of the choice between MAR and MNAR mechanisms becomes a sensitivity analysis.

Alternatively, a data collector may be able to make a possibly nonignorable mechanism less so, as mentioned by Rubin (1978). Rubin's suggestion is for survey data collectors to collect supplementary information from survey respondents, either information that is hypothesised to be correlated with data that might be prone to nonignorable nonresponse, or simply supplementary information that might help to explain the nonresponse. In this way, the nonignorable nonresponse may be able to be modelled by these additional covariates, in essence transforming the nonignorable process into an ignorable one.

The drawback of this process, of course, is that it needs to be completed at the survey data collection stage, which is not likely to occur properly in practice. However, the process raises the possibility that if one believes a MDM might be nonignorable, one could theoretically increase the number of covariates in the imputation procedure (since surveys often have multitudes of questions that are answered) to increase the variation of the incomplete variable that is explained by the rest of the covariates, thereby approximating an ignorable MDM.[1]

2.2.3 Analysing multiple imputations

Once multiple data sets have been imputed from the same starting point, inferences on the data sets can be combined using a simple set of rules as originally defined by Rubin (1987), and explained below. More insight on the formation of these rules will be given in Subsection (2.2.4).

Suppose that Q is a scalar population quantity to be estimated from the sample data taken in a survey, and that an estimate of this quantity, $\hat{Q}$ and standard error $\sqrt{U}$ could be easily calculated if Y_{mis} were available. In MI, Y_{mis} is replaced by $m > 1$ simulated versions, $Y_{mis}^{(1)}, Y_{mis}^{(2)}, \ldots, Y_{mis}^{(m)}$, leading to m estimates and their respective standard errors, $\left(\hat{Q}_j, \sqrt{U_j}\right), j = 1, \ldots, m$. An overall estimate for Q is

$$\bar{Q} = \frac{1}{m} \sum_{j=1}^{m} \hat{Q}_j \tag{2.5}$$

with a standard error of $\sqrt{T}$, where

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B \tag{2.6}$$

$$\bar{U} = \frac{1}{m} \sum_{j=1}^{m} U_j \tag{2.7}$$

and

$$B = \frac{1}{m-1} \sum_{j=1}^{m} \left(\hat{Q}_j - \bar{Q}\right)^2 . \tag{2.8}$$

We have that T is the total variance of the estimator, while $\bar{U}$ is the regular within-

[1] This suggestion leads one to believe that imputers could possibly view R^2 statistics within an imputation process as measures of the ignorability of a MDM. Perhaps this could be a viable avenue of research, although it is beyond the scope of this thesis.

imputation variance, and B is the between-imputation variance.

Note that if all imputations are made under the same MDM, the variability measured by B stems from the inability of the observed data to predict the missing data under the given MDM. In contrast if the MDM is varied across the m imputed data sets, B would describe a measure of the sensitivity of the MI process to the choice of the MDM (Rubin 1978). This idea has not received much attention over the years. It is most common to fix the MDM and impute several data sets, rather than to impute under differing MDMs and incorporate the variation from these changing MDMs into B.

If $\frac{(\hat{Q}-Q)}{\sqrt{\bar{U}}}$ is approximately $N(0,1)$ with complete data, for example as is assumed to be the case in many regression contexts, then, in the imputed case, according to Rubin & Schenker (1986):

$$\frac{\left(\hat{Q}-Q\right)}{\sqrt{T}} \sim t_v \tag{2.9}$$

where:

$$v = (m-1)\left[1+\left(\frac{m}{m+1}\right)\frac{\bar{U}}{B}\right]^2, \tag{2.10}$$

or

$$v = (m-1)\left(1+\frac{1}{r}\right)^2 \tag{2.11}$$

and

$$r = \left(1+\frac{1}{m}\right)\frac{B}{\bar{U}} \tag{2.12}$$

It is worth repeating that Equations (2.9)-(2.14) are for a complete data analysis that is based on the Normal distribution.

The latter, r, is the relative increase in variance due to nonresponse (Schafer & Graham 2002). The degrees of freedom vary from $(m-1)$ to ∞ according to the rate of missing information in the data set. According to Rubin (1987), the rate of missing information is given by:

$$\gamma = \frac{r + \frac{2}{v+3}}{r+1}, \tag{2.13}$$

where r is given above. The estimated rate of missing information, $\hat{\gamma}$, is approximately $r/(r+1)$. Through simple rearranging, this term can be written as the form for the rate of missing information given by Little & Rubin (2002):

$$(1 + \frac{1}{m})\frac{B}{T} \tag{2.14}$$

The derivation of this estimate is given in Subsection 2.2.4. As B approaches T, this expression shows that the rate of missing information becomes large. This is expected, as in this case the total variation of the estimate is made up mostly of between-imputation variance, meaning that the imputation model is imputing wildly different imputations from one imputed data set to the next. This can happen if the imputation model is guessing imputations based on little known information.

Schafer & Graham (2002) also note that with large degrees of freedom (or alternatively when the variation in the estimates between imputations is small compared to the overall variation), there is not much that can be gained from increasing m, the number of imputed data sets.

Additionally, Schenker, Raghunathan, Chiu, Makuc, Zhang & Cohen (2006) show that when the rate of missing information is low, point estimates from MI vary little from those obtained through single imputation. This is naturally due to a small value of B that arises when there is little missing information. These authors also mention that the rate of missing information is regularly less than the proportion of nonresponse, due to the predictive power of other variables within the incomplete data set.

Barnard & Rubin (1999) provide a further refinement to the expression for the degrees of freedom when the completed data sets are based on limited degrees of freedom, say, v_{com} (when there are no missing values). In this case, v is replaced by v^*, given by

$$v^* = \left(v^{-1} + \hat{v}_{obs}^{-1}\right)^{-1}, \tag{2.15}$$

where

$$\hat{v}_{obs} = (1 - \hat{\gamma}) \left(\frac{v_{com} + 1}{v_{com} + 3} \right) v_{com}. \tag{2.16}$$

This modified degrees of freedom increases monotonically in v_{com}, is always less than or equal to v_{com}, and is equal to the original degrees of freedom, v, when v_{com} is infinite.

In order to determine the number of imputed data sets that should be created, Rubin (1987) provides a measure of efficiency, measured in standard errors, and based on the rate of missing information, γ, or at least $\hat{\gamma}$. It is given by:

$$\lambda = \left(1 + \frac{\gamma}{m} \right)^{-\frac{1}{2}} \tag{2.17}$$

This measure essentially compares the size of the standard error after m imputations with the size of the standard error after an infinite number of imputations.

Although the number of data sets that should be completed is often debated, a small number of completed data sets, say, between 10 and 20, often suffices in order to obtain precise estimates (assuming that the fraction of missing information is not extreme). According to Little & Rubin (2002, p. 209),

> "In those cases where inference from the complete-data posterior distribution is based on multivariate [N]ormality (or the multivariate t), posterior moments of θ can be reliably estimated from a surprisingly small number, D, of draws of the missing data Y_{mis} (e.g., $D = 2$–10), if the fraction of missing information is not too large."

It should be noted, however, that recent arguments against modest m. Zhou & Reiter (2010) show in a simulation study that if an analyst intends on doing Bayesian analysis on multiply imputed data, a large number of imputed data sets should be created. Bayesians should not use average posterior quantiles of the resulting statistics, as the regular combining rules might suggest, but rather use the approach of Gelman, Carlin, Stern & Rubin (2004, p. 520) that obtains quantiles based on pooled statistic estimates from a large number of completed data sets.

2.2.4 Proper, valid multiple imputation

A large field of research is summarised by van Buuren, Boshuizen & Knook (1999, p.682) when they explain that if the complete data model leads to valid inferences in the absence of non-response and if the imputation procedure is proper with respect to the non-response mechanism, then MI yields valid inferences. It is, therefore, critical to expand on the definition and description of proper MI. One should note, however, that the expansion presented in this section is limited to work already deemed well-known in MI research. The repetition of this content within this thesis is necessary for a reader to be able to follow the evolution of MI without having to refer to the progression of works from which the definitions, formulae and proofs are obtained.

Nielson (2003) explains that proper MI methods are those for which the regular combining rules yield a consistently asymptotically Normal estimator of the unknown parameter and a weakly unbiased estimator of its asymptotic variance (given by a combination of the average of the complete data variance estimators and the empirical variance of the multiple estimators) in sufficiently regular models. He also explains that this means that Bayesian MI is proper when the model used for the imputations and the model used for the analysis are compatible. One instance when this is the case is when the complete data estimator is the complete data MLE.

Infinite repetitions

Rubin (1987) defines proper MI methods as first examining the theoretical situation where the number of imputed data sets is infinite, and then deriving the asymptotic distributions for the case of finite m, and thereby deriving the regular MI combining rules:

Definition: Let $\bar{Q}_\infty$ be the average of the estimators calculated over an infinite number of imputation-filled data sets and let B_∞ be the between-imputation variance of the estimators calculated over an infinite number of imputation-filled data sets. A MI procedure is **proper** for the set of complete-data statistics $\{\hat{Q}, Y\}$ if three conditions are satisfied:

1. Treating $(X^{(N)}, Y^{(N)}, I^{(N)})$ as fixed, under the posited response mechanism, the $m = \infty$ MI procedure provides randomisation-valid inferences for the complete-

data statistic $\hat{Q} = \hat{Q}(X^{(N)}, Y_{inc}^{(N)}, I^{(N)})$ based on the statistics $\bar{Q}_\infty$ and B_∞:

$$\left(\bar{Q}_\infty | X^{(N)}, Y^{(N)}, I^{(N)}\right) \sim N\left(\hat{Q}, B\right) \tag{2.18}$$

$$\left(B_\infty | X^{(N)}, Y^{(N)}, I^{(N)}\right) \sim (B, \ll B), \tag{2.19}$$

where $B = B\left(X^{(N)}, Y_{inc}^{(N)}, I^{(N)}\right)$ is defined by

$$B = V\left(\bar{Q}_\infty | X^{(N)}, Y^{(N)}, I^{(N)}\right), \tag{2.20}$$

and $D \sim (E, \ll F)$ means that the distribution of D tends to be centred at E with each component having variability substantially less than each positive component of F.

2. Treating $(X^{(N)}, Y^{(N)}, I^{(N)})$ as fixed, under the posited response mechanism, the $m = \infty$ imputation estimate of the complete-data statistic $U = U(X^{(N)}, Y_{inc}^{(N)}, I^{(N)})$, that is, $\bar{U}_\infty$, is centred at U with variability of a lower order than that of $\bar{Q}_\infty$:

$$\left(\bar{U}_\infty | X^{(N)}, Y^{(N)}, I^{(N)}\right) \sim (U, \ll B). \tag{2.21}$$

3. Treating $(X^{(N)}, Y^{(N)})$ as fixed, over repeated samples the variability of B is of lower order than that of $\hat{Q}$:

$$\left(B | X^{(N)}, Y^{(N)}\right) \sim (B_0, \ll U_0), \tag{2.22}$$

where $B_0 = B_0\left(X^{(N)}, Y^{(N)}\right)$ is defined by

$$B_0 = E\left(B | X^{(N)}, Y^{(N)}\right), \tag{2.23}$$

and $U_0 = U_0\left(X^{(N)}, Y^{(N)}\right) = E\left(U | X^{(N)}, Y^{(N)}\right) = V\left(\hat{Q} | X^{(N)}, Y^{(N)}\right)$ is fixed by the true value of $X^{(N)}$ and $Y^{(N)}$.

The underlying random variable in (2.18)–(2.21) is $R^{(N)}$ with distribution specified by the response mechanism (or MDM) $\Pr(R^{(N)} | X^{(N)}, Y^{(N)})$, whereas the underlying random variable in (2.22) and (2.23) is $I^{(N)}$ with distribution $\Pr(I^{(N)}|Y^{(N)}, X^{(N)}) = \Pr(I^{(N)}|X^{(N)})$.

Rubin (1987) proves that when proper MI methods are coupled with randomisation-validity of the complete-data inference, the randomisation-validity of the infinite-m repeated-imputation inference is implied.

Definition: Randomisation-validity of the complete-data inference means that

$$\left(\hat{Q}|X^{(N)}, Y^{(N)}\right) \sim N\left(Q, U_0\right) \tag{2.24}$$

and

$$\left(U|X^{(N)}, Y^{(N)}\right) \sim \left(U_0, \ll U_0\right), \tag{2.25}$$

where $Q = Q\left(X^{(N)}, Y^{(N)}\right)$ and $U_0 = U_0\left(X^{(N)}, Y^{(N)}\right)$ are fixed by the true value of $X^{(N)}$ and $Y^{(N)}$, and the underlying random variable in (2.24) and (2.25) is $I^{(N)}$ with distribution $\Pr(I^{(N)}|Y^{(N)}, X^{(N)}) = \Pr(I^{(N)}|X^{(N)})$.

Definition: Randomisation-validity of the infinite-m repeated imputation inference means that

$$\left(\bar{Q}_\infty|X^{(N)}, Y^{(N)}\right) \sim N\left(Q, T_0\right) \tag{2.26}$$

and

$$\left(T_\infty|X^{(N)}, Y^{(N)}\right) \sim \left(T_0, \ll T_0\right), \tag{2.27}$$

where $Q = Q\left(X^{(N)}, Y^{(N)}\right)$ and $T_0 = T_0\left(X^{(N)}, Y^{(N)}\right) = V\left(\bar{Q}_\infty|X^{(N)}, Y^{(N)}\right)$ are fixed by the true value of $X^{(N)}$ and $Y^{(N)}$, $T_\infty = \bar{U}_\infty + B_\infty$ and $\bar{Q}_\infty$, $\bar{U}_\infty$ and B_∞ are functions of $(X^{(N)}, Y_{obs}^{(N)}, R_{inc}^{(N)}, I^{(N)}) = (Y_{obs})$. The random variable in (2.26) and (2.27) is $(R^{(N)}, I^{(N)})$.

If the last definition holds, then the repeated-imputation inference is randomisation-valid, meaning that the 95% interval estimate given by $\bar{Q}_\infty \pm 1.96 T_\infty^{1/2}$ will be a 95% confidence interval, and if $Q = Q_0$, then the p-value given by

$$\begin{aligned} p - \text{value}\left(Q_0 | X^{(N)}, Y_{obs}^{(N)}, R_{inc}^{(N)}, I^{(N)}\right) = \\ \Pr\left\{\chi^2_{\dim(Q)} > \left(Q_0 - \bar{Q}_\infty\right) T_\infty^{-1} \left(Q_0 - \bar{Q}_\infty\right)'\right\} \end{aligned}$$

will be uniformly distributed on $(0, 1)$. Hence the importance of a randomisation-valid complete-data estimation procedure, and a proper MI procedure.

Theorem 1 (From Rubin, 1987, Result 4.1) *If the complete-data inference is randomisation-valid and the MI procedure is proper, then the infinite-m repeated-imputation inference is randomisation-valid under the posited response mechanism.*

Proof: Note that (2.24) and (2.18) imply that

$$\left(\bar{Q}_\infty | X^{(N)}, Y^{(N)}\right) \sim N\left(Q, U_0 + E\left(B | X^{(N)}, Y^{(N)}\right)\right), \tag{2.28}$$

which, by (2.22) gives

$$\left(\bar{Q}_\infty | X^{(N)}, Y^{(N)}\right) \sim N\left(Q, U_0 + B_0\right). \tag{2.29}$$

Next note that by (2.19) and (2.21)

$$\left(\bar{U}_\infty + B_\infty | X^{(N)}, Y^{(N)}, I^{(N)}\right) \sim \left(U + B, \ll 2B\right), \tag{2.30}$$

which, by (2.25) and (2.22)

$$\left(\bar{U}_\infty + B_\infty | X^{(N)}, Y^{(N)}\right) \sim \left(U_0 + B_0, \ll \left(2B_0 + 2U_0\right)\right), \tag{2.31}$$

as required.

■

Rubin (1996) provides an alternative way of proving this theorem, by proving the following distribution for inferences on Q:

$$\left(Q - \bar{Q}_\infty\right) \sim N\left(0, T_\infty\right)$$

This distribution is due to the following:

$$\begin{aligned}
E\left(\bar{Q}_\infty | X^{(N)}, Y^{(N)}\right) &= E\left[E\left(\bar{Q}_\infty | X^{(N)}, Y^{(N)}, I^{(N)}\right) | X^{(N)}, Y^{(N)}\right] \\
&= E\left(\hat{Q} | X^{(N)}, Y^{(N)}\right) \\
&= Q
\end{aligned}$$

and

$$\begin{aligned}
& E\left(T_\infty | X^{(N)}, Y^{(N)}\right) \\
&= E\left(\bar{U}_\infty | X^{(N)}, Y^{(N)}\right) + E\left(B_\infty | X^{(N)}, Y^{(N)}\right) \\
&= E\left[E\left(\bar{U}_\infty | X^{(N)}, Y^{(N)}, I^{(N)}\right) | X^{(N)}, Y^{(N)}\right] \\
&\quad + E\left[E\left(B_\infty | X^{(N)}, Y^{(N)}, I^{(N)}\right) | X^{(N)}, Y^{(N)}\right] \\
&= E\left(U | X^{(N)}, Y^{(N)}\right) \\
&\quad + E\left[V\left(\bar{Q}_\infty | X^{(N)}, Y^{(N)}, I^{(N)}\right) | X^{(N)}, Y^{(N)}\right] \\
&= V\left(\hat{Q} | X^{(N)}, Y^{(N)}\right) \\
&\quad + E\left[V\left(\bar{Q}_\infty | X^{(N)}, Y^{(N)}, I^{(N)}\right) | X^{(N)}, Y^{(N)}\right] \\
&= V\left[E\left(\bar{Q}_\infty | X^{(N)}, Y^{(N)}, I^{(N)}\right) | X^{(N)}, Y^{(N)}\right] \\
&\quad + E\left[V\left(\bar{Q}_\infty | X^{(N)}, Y^{(N)}, I^{(N)}\right) | X^{(N)}, Y^{(N)}\right] \\
&= V\left(\bar{Q}_\infty | X^{(N)}, Y^{(N)}\right)
\end{aligned}$$

Rubin (1987, p. 125–126) concludes that, "[i]f imputations are drawn to approximate repetitions from a Bayesian posterior distribution of Y_{mis} under the posited response mechanism and an appropriate model for the data…", *i.e.* that $(\bar{Q}_\infty, \bar{U}_\infty)$, the posterior mean of $(\hat{Q}, U)$ are unbiased for $(\hat{Q}, U)$ under the posited response mechanism:

$$E\left(\bar{Q}_\infty | X^{(N)}, Y^{(N)}, I^{(N)}\right) \doteq \hat{Q} \tag{2.32}$$

and

$$E\left(\bar{U}_\infty | X^{(N)}, Y^{(N)}, I^{(N)}\right) \doteq U \tag{2.33}$$

then, "...in large samples the imputation method is proper." Although exceptions can be found, this may serve as a guideline to proper Bayesian imputation. Rubin's (1987, p. 126–127) message is summarised in three steps:

1. Draw imputations following the Bayesian paradigm as repetitions from a Bayesian posterior distribution of the missing values under the chosen model for nonresponse and data, or an approximation to this posterior distribution that incorporates between-imputation variability.
2. Choose models of nonresponse appropriate for the posited response mechanism.
3. Choose models for the data that are appropriate for the complete-data statistics likely to be used — if the model for the data is correct, then the model is appropriate for all complete-data statistics.

Brand (1998, p. 87) summarises the conditions for proper and valid multiple imputations in the case of an infinite number of imputed data sets. For proper imputations:

$$\bar{Q}_\infty \sim N\left(\hat{Q}, B\right), \tag{2.34}$$

$$\bar{U}_\infty \approx U, \tag{2.35}$$

$$B_\infty \approx B, \tag{2.36}$$

$$B \approx B_0. \tag{2.37}$$

For valid imputations:

$$\bar{Q}_\infty \sim N\left(Q, T_0\right), \tag{2.38}$$

$$\bar{T}_\infty \approx T_0. \tag{2.39}$$

Nielson (2003) argues that multiple imputation is not proper when the imputation models and the analysis models are the same. Rubin (2003*a*) responds by mentioning that it is not necessary to multiply impute if one analyses the complete data according to the models known. If a researcher wishes to model the completed data using a different model to that of the imputation process, then MI on the original accepted models is useful. In this case, if one has an accepted model for the complete data (different models for each variable) and the mechanism that creates the missing data, one effectively has the posterior distribution(s).

Finite-m asymptotic distributions

The derivations in this section are taken from Section 4.5 from Rubin (1987).

Suppose that m draws of Y_{mis} are taken from its posterior distribution, given as $\Pr(Y_{mis}|X^{(N)}, Y_{obs}^{(N)}, R_{inc}^{(N)}, I^{(N)})$, and let $\mathbf{S}_m = \hat{Q}_{*l}, U_{*l}, l = 1, \ldots, m$ be the set of associated completed-data statistics $\hat{Q} = E(Q|X^{(N)}, Y_{inc}^{(N)}, R_{inc}^{(N)}, I^{(N)})$ and $U = V(Q|X^{(N)}, Y_{inc}^{(N)}, R_{inc}^{(N)}, I^{(N)})$ evaluated on each of the m filled-in data sets. Rubin (1987, p. 88–89) shows that it is reasonable to assume that

$$\left(\hat{Q}_{*l}|X^{(N)}, Y^{(N)}, I^{(N)}, R^{(N)}\right) \sim N\left(\bar{Q}_\infty, B_\infty\right) \tag{2.40}$$

and

$$\left(U_{*l}|X^{(N)}, Y^{(N)}, I^{(N)}, R^{(N)}\right) \sim \left(\bar{U}_\infty, \ll B_\infty\right), \tag{2.41}$$

where all the $\hat{Q}_{*l}$ and U_{*l} are mutually independent given $X^{(N)}$, $Y^{(N)}$, $I^{(N)}$, and $R^{(N)}$. When the imputations are repetitions drawn from the posterior distribution of Y_{mis}, under specified models for the data, $\Pr(X^{(N)}, Y^{(N)})$, and the response mechanism, $\Pr(R^{(N)}|X^{(N)}, Y^{(N)})$, it can be argued that (2.40) and (2.41) will hold regardless of the models' correctness. Furthermore, Rubin (1987) writes that evidence exists suggesting that these distributional forms will hold asymptotically for a wide variety of approximately Bayesian imputation models. These assumptions, as well as the assumptions of proper imputation methods and valid complete-data inferences, are used in the following derivations.

The derivation for the conditional distribution of $\bar{Q}_m$ follows three steps:

1. Average over the multiple estimates given $(X^{(N)}, Y^{(N)}, I^{(N)}, R^{(N)})$, assuming the

asymptotic validity of the sampling distribution $\mathbf{S}_m$. From (2.40) and (2.41) we have that

$$\left(\bar{Q}_m | X^{(N)}, Y^{(N)}, I^{(N)}, R^{(N)}\right) \sim N\left(\bar{Q}_\infty, B_\infty / m\right), \tag{2.42}$$

$$\left(\bar{U}_m | X^{(N)}, Y^{(N)}, I^{(N)}, R^{(N)}\right) \sim \left(\bar{U}_\infty, \ll B_\infty / m\right), \tag{2.43}$$

and

$$\left((m-1) B_m B_\infty^{-1} | X^{(N)}, Y^{(N)}, I^{(N)}, R^{(N)}\right) \sim \chi^2_{m-1} \tag{2.44}$$

where these random variables are mutually independent given $X^{(N)}$, $Y^{(N)}$, $I^{(N)}$, and $R^{(N)}$.

2. Average over $R^{(N)}$ given $(X^{(N)}, Y^{(N)}, I^{(N)})$, assuming the the imputation method is proper under the posited response mechanism. The expressions (2.18), (2.19), (2.42), and (2.44) imply that

$$\left(\bar{Q}_m | X^{(N)}, Y^{(N)}, I^{(N)}\right) \sim N\left(\hat{Q}, (1 + 1/m) B\right), \tag{2.45}$$

and

$$\left((m-1) B_m B^{-1} | X^{(N)}, Y^{(N)}, I^{(N)}\right) \sim \chi^2_{m-1}; \tag{2.46}$$

The expressions (2.21), (2.19), and (2.43) imply that

$$\left(\bar{U}_m | X^{(N)}, Y^{(N)}, I^{(N)}\right) \sim \left(U, \ll (1 + 1/m) B\right). \tag{2.47}$$

The three random variables in expressions (2.45)–(2.47) are mutually independent given $X^{(N)}$, $Y^{(N)}$, and $I^{(N)}$.

3. Average over $I^{(N)}$ given $(X^{(N)}, Y^{(N)})$, assuming the complete-data inference is randomisation-valid under the specified unconfounding sampling mechanism. Expressions (2.24),

(2.25), (2.22), and (2.45) imply that

$$\left(\bar{Q}_m|X^{(N)}, Y^{(N)}\right) \sim N\left(Q, U_0 + (1 + 1/m)\, B_0\right); \tag{2.48}$$

expressions (2.48) and (2.46) imply that

$$\left((m-1)\, B_m B_0^{-1}|X^{(N)}, Y^{(N)}\right) \sim \chi^2_{m-1}; \tag{2.49}$$

and expressions (2.25), (2.22), and (2.47) imply that

$$\left(\bar{U}_m|X^{(N)}, Y^{(N)}\right) \sim \left(U_0, \ll \left(U_0 + (1 + 1/m)\, B_0\right)\right); \tag{2.50}$$

where the three random variables in (2.48)–(2.50) are mutually independent given $X^{(N)}$ and $Y^{(N)}$.

Expressions (2.48)–(2.50), which comprise the asymptotic sampling distribution of $\mathbf{S}_m$ given $X^{(N)}$ and $Y^{(N)}$ for proper MI methods, imply that

$$E\left(T_m|X^{(N)}, Y^{(N)}\right) = V\left(\bar{Q}_m|X^{(N)}, Y^{(N)}\right) \tag{2.51}$$

and

$$V\left(T_m|X^{(N)}, Y^{(N)}\right) = 2\left(1 + 1/m\right)^2 B_0^2/\left(m-1\right) \tag{2.52}$$

which together imply that

$$\frac{\left[2E\left(T_m|X^{(N)}, Y^{(N)}\right)\right]^2}{V\left(T_m|X^{(N)}, Y^{(N)}\right)} = (m-1)\left\{1 + \left[\left(1 + \frac{1}{m}\right)\frac{B_0}{U_0}\right]^{-1}\right\}^2 \tag{2.53}$$

The reason for this is that for any multiple of a χ^2 random variable, twice the squared expectation divided by the variance gives the degrees of freedom.

These expressions are the derivations for the expressions given for the general combining rules in Subsection 2.2.3. Expressions (2.48), (2.51), and (2.53) provide a random-response randomisation-based justification for using a t reference distribution for $(\bar{Q}_m - Q)$ with a squared scale T_m,

$$T_m = \bar{U}_m + \left(1 + \frac{1}{m}\right) B_m, \tag{2.54}$$

and degrees of freedom

$$v = (m-1)\left(1 + \frac{1}{r_m}\right)^2, \tag{2.55}$$

where

$$r_m = \left(1 + \frac{1}{m}\right) \frac{B_m}{U_m}, \tag{2.56}$$

which estimates $(1 + 1/m)B_0/U_0$. The v in Equation (2.55) is the estimate of the v given as the degrees of freedom in Equation (2.53).

For multicomponent, or k-dimensional Q, these results are easily extended. Expressions (2.48) and (2.50) still hold, but B_m's distribution must be modified from the χ^2 to the k-dimensional Wishart with $m-1$ degrees of freedom:

$$\left((m-1)\, B_m | X^{(N)}, Y^{(N)}\right) \sim \text{Wishart}_k\,(m-1, B_0) \tag{2.57}$$

or, equivalently,

$$(m-1)\, B_m = \sum_{i=1}^{m-1} Z_i Z_i', \text{ where } Z_i \sim N\,(0, B_0)\,. \tag{2.58}$$

Note that the three random variables $\bar{Q}_m$, B_m, and $\bar{U}_m$ are still mutually independent.

Returning to the asymptotic t distribution of the finite-m statistics, we have that

$$(Q|\mathbf{S}_m) \sim t_v\left(\bar{Q}_m, T_m\right), \tag{2.59}$$

where v is as given in expression (2.55). The negative average second derivative of the logarithm of this posterior distribution, or, the information about Q in this posterior distribution, is

$$\gamma_m = \frac{\left[\bar{U}_m^{-1} - \frac{(v+1)}{(v+3)} T_m^{-1}\right]}{\bar{U}_m^{-1}}. \tag{2.60}$$

Using expressions (2.54) and (2.56), this can be rewritten as

$$\gamma_m = \frac{r_m + \frac{2}{v+3}}{r_m + 1}. \tag{2.61}$$

For notational simplicity one can drop the subscript, and the result given as Equation (2.17) is obtained. It is important to note that this quantity measures the extent of the effect of missing information for a univariate Q, or for single component of a multivariate Q. Indeed, as Brand (1998, p. 74) puts it, it measures "the loss in precision due to missing data."

One can also measure the gain in precision from MI when compared with complete-case analysis on incomplete data. Similar to Equation 2.60, if $\tilde{U}$ is the variance of the parameter estimate $\hat{Q}$ when $\hat{Q}$ is obtained from complete-case analysis, then Brand (1998, p. 74) gives the fraction of information gained about Q from MI above that obtained by the complete-case analysis as

$$\zeta_m = \frac{\frac{v+1}{v+3} T_m^{-1} - \tilde{U}^{-1}}{\tilde{U}^{-1}}. \tag{2.62}$$

Brand (1998, p. 47) summarises the link between infinite iterations and the finite-m case when he mentions the following:

> "In the theoretical situation that m is infinite, the empirical probability distribution obtained from the imputations is exactly equal to the probability distribution representing the missing data entries as uncertain values, so that the multiple imputation results are identical to the results from the analytical incorporation of the extra uncertainty due to missing data. For modest m, the multiple imputation results are an approximation of analytical incorporation, so that the pooling procedures also reflect extra uncertainty due to the simulation error."

Keep in mind, however, as Zhou & Reiter (2010) mention, that Bayesian analysts should indeed prefer a larger number of multiply imputed data sets, as mentioned previously in this chapter.

2.2.5 Congeniality and efficiency in multiple imputation

The two main topics discussed in this subsection, congeniality and efficiency, go hand-in-hand. When an analysis model is congenial to the imputation procedure, it can be argued that the estimators are inefficient.

Congeniality

Suppose that we have an analysis procedure Ψ. From a non-Bayesian perspective, the analyst's complete-data procedure is summarised by

$$\Psi_{com} = [\hat{Q}(X^{(N)}, Y_{inc}^{(N)}, I^{(N)}), U(X^{(N)}, Y_{inc}^{(N)}, I^{(N)})],$$

where $\hat{Q}(X^{(N)}, Y_{inc}^{(N)}, I^{(N)}) = \hat{Q}(Y_{com})$ is an estimator of Q, with associated variance estimator $U(X^{(N)}, Y_{inc}^{(N)}, I^{(N)}) = U(Y_{com})$. The procedure Ψ_{com} depends on $I^{(N)}$, but not on $R_{inc}^{(N)}$, the population indicator matrix for response on $Y^{(N)}$ for the data that is included in the survey. This shows that different survey designs are accommodated, but once a design is chosen, the response behaviour itself carries no information about Q when the MDM is ignorable (see Rubin 1987, Meng 1994). This complete-data procedure can include any method of completing the data, for example, MI. By comparison, let any incomplete-data procedure be

$$\Psi_{obs} = [\hat{Q}(X^{(N)}, Y_{obs}^{(N)}, I^{(N)}, R_{inc}^{(N)}), U(X^{(N)}, Y_{obs}^{(N)}, I^{(N)}, R_{inc}^{(N)})],$$

where $\hat{Q}(X^{(N)}, Y_{obs}^{(N)}, I^{(N)}, R_{inc}^{(N)}) = \hat{Q}(Y_{obs})$ is the incomplete-data estimator with associated variance estimator $U(X^{(N)}, Y_{obs}^{(N)}, I^{(N)}, R_{inc}^{(N)}) = U(Y_{obs})$. With a switch from finite-population Bayesian methodology to superpopulation Bayesian methodology (and hence a switch from estimating Q to estimating θ), without much loss of generality, Meng (1994,

p. 543) provides the following definition for a model being congenial to an analysis procedure:

Definition: A Bayesian model f is said to be (second-moment) *congenial* to the analysis procedure $\Psi \equiv \{\Psi_{obs}; \Psi_{com}\}$ for given $Y_{obs} = X^{(N)}, Y_{obs}^{(N)}, I^{(N)}, R_{inc}^{(N)}$ if the following hold:

1. The posterior mean and variance of θ under f given the incomplete data are asymptotically the same as the estimate and variance from the analyst's incomplete-data procedure Ψ_{obs}, that is

$$\left[\hat{\theta}\left(Y_{obs}\right), U\left(Y_{obs}\right)\right] \simeq \left[E_f\left(\theta|Y_{obs}\right), V_f\left(\theta|Y_{obs}\right)\right] \tag{2.63}$$

2. The posterior mean and variance of θ under f given the complete data are asymptotically the same as the estimate and variance from the analyst's complete-data procedure Ψ_{obs}, that is

$$\left[\hat{\theta}\left(Y_{com}\right), U\left(Y_{com}\right)\right] \simeq \left[E_f\left(\theta|Y_{com}\right), V_f\left(\theta|Y_{com}\right)\right] \tag{2.64}$$

for any possible $Y_{inc} = (Y_{obs}, Y_{mis})$ with Y_{obs} fixed. In this definition, $a \simeq b$ means that the differences between corresponding elements of a and b are negligible compared to the leading terms (e.g., $a/b \rightarrow 1$ when a and b are scalars) as the sample size of Y_{obs} becomes large, or, in other words, it denotes finite-sample asymptotic equivalency with respect to the size of the observed data. Also note that E_f and V_f are the posterior mean and variance with respect to f.

Any procedure that can't be embedded into a Bayesian model is one that is not congenial. This is rarely the case, as most analysis procedures are related to at least a simple Bayesian context, for example, a common procedure model being Normal model with a non-informative prior. Meng (1994, p. 543) immediately follows this definition with a definition relating imputation models to the analysis procedures.

Definition: The analysis procedure Ψ is said to be congenial to the imputation model $g(Y_{mis}|Y_{obs}, A)$ (where A refers to any information that the imputer has above and beyond that which is given to the analyser) if one can find an f such that f is congenial to Ψ and the posterior predictive density for Y_{mis} derived under f is identical to the imputation model, *i.e.* $f(Y_{mis}|Y_{obs}) = g(Y_{mis}|Y_{obs}, A)$, for all possible Y_{mis}.

Meng's (1994) second definition, given above, implies that imputation models congenial to analysis procedures essentially make use of the same model or distribution. It is easy to see that uncongeniality is then usually the norm in the standard setting for analysis of MI — that being public use of data that was multiply imputed 'behind the scenes' as it were. Most often, the analyst will not have access to any extra information built into the imputation model, and so will not know the imputation model, leading to uncongeniality. Moreover, imputers and analysers may have different goals in their work, and may thus use different models, also introducing uncongeniality. Finally, often several models are considered for imputation or analysis, inevitably introducing uncongeniality.

Even with these points in mind one still might wish to aim for congeniality. However, the dual use of the same model (which would assist in obtaining congeniality) is criticised by Nielson (2003), stating that when the imputation model and the analysis model are identical, the imputer has unwittingly increased inefficiency in the analyser's work. This criticism will be reviewed in the subsequent subsection. Moreover, as Schafer (2003) points out, MI under congeniality implies the same inferences as originating from a maximum likelihood setting, weakening the backing behind the evolution of MI. In light of these criticism, however, one should keep in mind a question that should be posed: How important is congeniality for the MI framework?

This question is an important one indeed, since the justification for the regular combining rules given in Subsection 2.2.3 is most straightforward when the analyst's procedure is congenial to the imputation model (Meng 1994). In this case we have *inferential congeniality*, meaning that the incomplete-data inference Ψ_{obs} asymptotically agrees with the infinite-m repeated-imputation inference, which is denoted Ψ_{∞}. Meng (1994, p. 544) shows how inferential congeniality arises if f is a Bayesian model congenial to Ψ, and g

is the imputation model, using the two definitions given above:

$$\begin{aligned}
\bar{\theta}_\infty &= E_g\left[\hat{\theta}\left(Y_{com}\right)|Y_{obs}, A\right] \\
&\simeq E_g\left[E_f\left[\theta|Y_{com}\right]|Y_{obs}, A\right] \\
&= E_f\left[E_f\left[\theta|Y_{com}\right]|Y_{obs}\right] \\
&= E_f\left[\theta|Y_{obs}\right] \\
&\simeq \theta\,(\hat{Y}_{obs}).
\end{aligned}$$

Also:

$$\begin{aligned}
\bar{U}_\infty &\equiv \lim_{m\to\infty} \bar{U}_m \simeq E_f\left[V_f\left[\theta|Y_{com}\right]|Y_{obs}\right], \\
B_\infty &\equiv \lim_{m\to\infty} B_m \simeq V_f\left[E_f\left[\theta|Y_{com}\right]|Y_{obs}\right].
\end{aligned}$$

And, so:

$$T_\infty = \bar{U}_\infty + B_\infty \simeq V_f\left[\theta|Y_{obs}\right] \simeq U\left(Y_{obs}\right).$$

These equations imply that inferential uncongeniality means that $\Psi_\infty \simeq \Psi_{obs}$. Meng's (1994) paper looks to find the answer to the question that immediately arises following these proofs: in the presence of inferential uncongeniality, which procedure, Ψ_{obs} or Ψ_∞, provides better statistical inference?

Meng's (1994) third definition reiterates Rubin's definition for proper MI (albeit a 'weaker' version than Rubin's). For the sake of completeness, it is restated in this thesis using Meng's (1994) notation.

Definition: An imputation model g is said to be *second-moment proper* for Ψ_{com} if the following three conditions are satisfied:

1. $\bar{\theta}_\infty$ and $\hat{\theta}(Y_{com})$ have the same expectation,

$$E\left[\bar{\theta}_\infty|X^{(N)}, Y^{(N)}\right] \simeq E\left[\hat{\theta}\left(Y_{com}\right)|X^{(N)}, Y^{(N)}\right]$$

2. $\bar{U}_\infty$ estimates the variance of $\hat{\theta}(Y_{com})$,

$$E\left[\bar{U}_\infty | X^{(N)}, Y^{(N)}\right] \simeq V\left[\hat{\theta}\left(Y_{com}\right) | X^{(N)}, Y^{(N)}\right]$$

3. B_∞ estimates the variance of $\bar{\theta}_\infty - \hat{\theta}(Y_{com})$,

$$E\left[B_\infty | X^{(N)}, Y^{(N)}\right] \simeq B\left[\left(\bar{\theta}_\infty - \hat{\theta}\left(Y_{com}\right)\right)^2 | X^{(N)}, Y^{(N)}\right].$$

This weaker version of the definition allows an imputation model to be *second-moment proper* while still being uncongenial to the analysis procedure. This definition also defines the conditions for validity of an imputation model, but does not describe its efficiency.

The remainder of Meng's (1994) paper connects congeniality with efficiency in MI estimators, particularly the variance decomposition into within- and between-imputation variances.

Measures of efficiency

Rubin (1987, Section 4.5) derives measures to calculate the asymptotic efficiency of $\bar{Q}_m$ relative to $\bar{Q}_\infty$. Given that $T_0 = T_0(X^{(N)}, Y^{(N)}) = V(\bar{Q}_\infty | X^{(N)}, Y^{(N)}) = U_0 + B_0$, and that the variance of $\bar{Q}_m$, given $X^{(N)}$ and $Y^{(N)}$, is $U_0 + (1 + 1/m)B_0$, from expression (2.48) (for the multicomponent Q), the variance-covariance matrix of $\bar{Q}_m$ relative to that of $\bar{Q}_\infty$ is

$$T_0^{-1/2}\left(T_0 + \frac{B_0}{m}\right)T_0^{-1/2} \tag{2.65}$$

This means that the efficiency of $\bar{Q}_m$ relative to $\bar{Q}_\infty$ in units of standard deviation is, from (2.65),

$$\lambda_0 = \left(I + \frac{\gamma_0}{m}\right)^{-1/2}, \tag{2.66}$$

where $\gamma_0 = T_0^{-1/2} B_0 T_0^{-1/2}$, the eigenvalues of B_0 with respect to T_0, or the population fractions of missing information. The largest fraction corresponds to the lowest relative efficiency for a linear combination of the components of Q, and the smallest fraction corresponds to the highest relative efficiency for such a linear combination. Since $T_0 = U_0 + B_0$, the eigenvalues of B_0 with respect to T_0 are one minus the eigenvalues of U_0 with respect to T_0. The formula for γ given in Subsection 2.2.3 follows from the asymptotic version of expression (2.66) for finite m.

Efficiency in estimators

With these basic ideas, one can explore the wealth of information arguing for and against the level of efficiency in the regular MI combining rules. The definitions provided by Meng (1994), and given below, follow directly after the definitions given in the previous paragraphs.

If the imputer's model makes more assumptions than an analysis procedure, and these assumptions are incorrect, invalid or inconsistent inferences follow. If these assumptions are correct, however, and the analyst does not know and include this information in the analysis procedure, then the completed-data analysis estimates will be *superefficient*, or more efficient than the estimates based only on the observed data. Superefficiency does not necessarily mean the estimators are the most efficient. Variance estimates are still inflated, a fact that has been criticised (see Fay 1992, Nielson 2003). However, while the estimates from the analysis not incorporating as much information as the imputation model are inefficient (and possibly inconsistent), the confidence intervals from Ψ_∞ will still be preferred to those from Ψ_{obs} since they have at least the same coverage, but are narrower. So in summary, Meng (1994, p. 547) writes:

> "[I]n the presence of uncongeniality, it is vital to recogni[s]e that disagreement between the repeated-imputation analysis and the (best possible) incomplete-data analysis does not automatically invalidate the repeated-imputation inference. Quite to the contrary, (substantial) disagreements between these two analyses often raise questions about the incomplete-data analysis, because it may suffer from serious nonresponse biases (as well as inefficiency) when the analyst has less information about the nonresponse mechanism than the imputer has."

On the other hand, if the imputation model is more general than the analysis procedure and the imputation model is valid, the inferences are automatically valid, without the analyst needing to identify the MDM. Meng (1994) argues that it is therefore very important for the imputer to correctly model the MDM. It must be noted that the opposite view is also well supported; Schafer (2003) notes that assuming the MAR mechanism when the data may have a more complex MDM is at least a step in the right direction, and should be done rather than not impute at all — more effort can thus be placed on modelling the data correctly, which may have stronger consequences than mis-modelling the MDM. In any case, a more general imputation model than analysis model often means larger standard errors on the MI estimates (Schafer 2003).

Meng (1994) provides a definition allowing one to quantify whether or not the imputations from a particular model are necessary. Note that the following definition does not need a congenial model for the analysis procedure, nor a correctly specified MDM.

Definition: An imputation model g is said to be *better* (than the analyst's congenial imputation model) for $\hat{\theta}(Y_{com})$ if

$$\begin{aligned} & E\left[\left(\bar{\theta}_{\infty} - \hat{\theta}\left(Y_{com}\right)\right)^2 |X^{(N)}, Y^{(N)}\right] \\ \leq \; & E\left[\left(\hat{\theta}\left(Y_{obs}\right) - \hat{\theta}\left(Y_{com}\right)\right)^2 |X^{(N)}, Y^{(N)}\right]. \end{aligned} \tag{2.67}$$

Meng (1994) then defines self-efficiency as follows (subsequently, Meng & Romero (2003) update the definition to the form that is presented later in this Subsection):

Definition: Let W_c be a data set, and let W_o be a subset of W_c created by a selection mechanism. A statistical estimation procedure $\hat{\theta}(.)$ for θ is said to be *self-efficient* (with respect to the selection mechanism) if there is no $\lambda \in (-\infty, +\infty)$ such that the mean-squared error of $\lambda\hat{\theta}(W_o) + (1-\lambda)\hat{\theta}(W_c)$ is less than that of $\hat{\theta}(W_c)$.

If the estimation procedure is self-efficient, then the variance decomposition formula $T_{\infty} = \bar{U}_{\infty} + B_{\infty}$ is correct from the Bayesian as well as the likelihood and randomisation perspectives, even if there is uncongeniality, as long as one does not assume that, "in the absence of missing data, the imputer has extra information to improve the efficiency of the analyst's self-efficient estimator" (Meng 1994, p. 549). Even if the imputer does have extra information, decomposition still provides a conservative estimate of the sampling variance of the repeated-imputation estimator, and this conservative estimator is still less than the sampling variance of the analyst's incomplete-data estimator, as long as the imputation model is *better* in terms of the definition above. As discussed in this Subsection, it would be unwise to assume an analyst would choose to use a non-self-efficient estimation procedure.

The decomposition will be conservative as long as there exists no negative λ that makes $\lambda\hat{\theta}(Y_{obs}) + (1-\lambda)\hat{\theta}(Y_{com})$ more efficient than $\hat{\theta}(Y_{com})$. This is common in practice, and leads us to Meng's (1994) next definition:

Definition: The imputation model g will be called *information regular* for estimating θ using the self-efficient estimator $\hat{\theta}(\cdot)$, if there is no negative λ such that the mean-squared error of $\lambda\bar{\theta}_{\infty} + (1-\lambda)\hat{\theta}(Y_{com})$ is less than that of $\hat{\theta}(Y_{com})$.

All of Meng's (1994) definitions (and the lemmas he provides in the paper to link them) lead to a set of conclusions (see Meng 1994, p. 551). If (i) the analyst's complete-data estimator $\hat{\theta}(Y_{com})$ is *self-efficient*, (ii) the imputer's model is *information regular* for estimating θ using $\hat{\theta}(Y_{com})$, and (iii) the imputer's model is *second-moment proper* with respect to the analyst's complete-data procedure Ψ_{com}, then the following hold:

1. The repeated-imputation estimator is consistent for θ, and is at least as efficient as the analyst's incomplete-data estimator.

2. For any nominal level, the corresponding repeated-imputation confidence interval has at least nominal coverage, but has at most the same width as the confidence interval from the analyst's incomplete-data procedure with the same nominal coverage.

It is important to note that information concerning the MDM is unimportant, and thus the analyst's incomplete-data estimator is allowed to be inconsistent. The conditions given above are basic guidelines for good practice in MI. The strictest condition is that of the imputer's model having to be second-moment proper.

Uncongeniality

The extensive work of Meng (1994) defines and analyses the advantages and disadvantages of congeniality, while Schafer (2003) provides insight into the benefits of uncongeniality. Summarising the findings of Collins, Schafer & Kam (2001), Schafer (2003, p. 25–26) writes, "Overall, there are potentially important gains and small risks associated with auxiliary variables in MI." This means that extra information used by the imputer and not available to the analyst (which means an imputation model uncongenial to the analysis model),

- improves the imputations generated if these auxiliary variables are correlated to the incomplete variable being imputed or are correlated with the MDM, and
- will not confound the imputation model seriously if they are not correlated with the incomplete variable being used.

These statements are related to the concept of superefficiency, as mentioned before. It must be noted that these statements also do not include auxiliary covariates that are functions of R, making the estimation consistent with a MNAR MDM, but possibly biased under MAR (Schafer 2003).

Increasing estimator efficiency

Meng (1994) continues his study into MI efficiencies by offering a method of combining imputations from different MDMs in order to obtain more efficient estimators, based on weights calculated using sampling importance resampling (SIR) methods. There might be cases when multiply imputed data sets are provided, but an analyst wishes to analyse the data completed through a different MDM. The method Meng (1994) offers could cater for this situation. In fact, a study by Carpenter & Kenward (2007), reviewed later in this chapter, uses Meng's methods to investigate the underlying MDMs.

The importance weights generated in Meng's (1994) algorithm provide a common way for adjusting draws from an incorrect model. Let $f(Y_{mis}|Y_{obs}, A)$ be an imputation model that the investigator desire to use (this model would not depend on A for an analyst), while $g(Y_{mis}|Y_{obs}, A)$ is the imputation model that was used for the existing imputations.

Then let the importance ratio be

$$\tau\left(Y_{mis}\right)=\frac{f\left(Y_{mis}|Y_{obs},A\right)}{g\left(Y_{mis}|Y_{obs},A\right)}C, \tag{2.68}$$

for some arbitrary positive constant C that does not depend on Y_{mis}.

Now suppose that the m parameter estimates and their associated variances from the multiply-imputed data sets are available, as well as the m values $\tau_j \equiv \tau(Y_{mis}^{(j)}), j = 1,\ldots,m$. Let

$$w_j=\frac{\tau_j}{\frac{1}{m}\sum_{j=1}^{m}\tau_j}\equiv\frac{\tau_j}{\bar{\tau}}. \tag{2.69}$$

Then Meng (1994) proposes that the weighted repeated-imputations estimator be

$$\bar{\theta}_w=\frac{1}{m}\sum_{j=1}^{m}w_j\hat{\theta}_j \tag{2.70}$$

which reduces to the regular combining rule if f and g are congenial. The new weighted combination for the variance is

$$T_w=\bar{U}_w+\left(1+\frac{1}{m}\right)B_w, \tag{2.71}$$

where

$$\bar{U}_w=\sum_{j=1}^{m}w_jU_j, \tag{2.72}$$

and

$$B_w=\sum_{j=1}^{m}w_j\left(\hat{\theta}_j-\bar{\theta}_w\right)\left(\hat{\theta}_j-\bar{\theta}_w\right)'. \tag{2.73}$$

Although this T_w provides a congenial variance associated with $\bar{\theta}_w$ as the number of imputations tends to infinity, for finite m it ignores the extra variability caused by the weighting scheme. The remedy provided by Meng (1994) refers to his earlier work (see Meng 1993), and suggests that the formula for T_w be adjusted to incorporate the sampling variance of the weights, s_w^2, *i.e.*,

$$\tilde{T}_w = \bar{U}_w + \left(1 + \frac{1 + s_w^2}{m}\right) B_w, \tag{2.74}$$

where

$$s_w^2 = \frac{1}{m-1} \sum_{j=1}^{m} \left(w_j - 1\right)^2 . \tag{2.75}$$

With large m, T_w and $\tilde{T}_w$ are, therefore, equivalent.

The main advantage of these weighted estimates is for uncongenial imputation and analysis models. If the imputations are made under one model, the imputer can provide importance weights for analyses under other models.

Even if the original imputations are proper with respect to the analysis model, a weighted mixture of the weighted combining rules and the unweighted, regular combining rules can be more efficient under certain circumstances, lending strength to analyses of this type, and providing the reason for this paragraph's title.

Some criticism

The most well-known criticism of estimators from MI is that they are inefficient (Fay 1992, Nielson 2003). Fay (1992) also suggests that estimators for subdomains in the incomplete data are worse under MI than under a correct design-based approach. Multiple imputation cannot capture the covariance between the estimators from the subgroups in the way a design-based approach can, although the variances under MI are larger than those under the design-based approach.

Nielson (2003) suggests that unless one uses an efficient complete data estimator, the variance estimator derived from the regular combining rules will be asymptotically biased. This bias may be upwards, leading to inefficient but correct inference, or downwards, leading to incorrect inference. Moreover, even if the complete data estimator is efficient, the MI procedure is inefficient, since the inconsistency of the variance estimator leads to weaker tests than a consistent estimator would. Rubin (2003*a*) responds by suggesting that even in the absence of missing data, an unbiased point estimate of sampling variance is only an intermediary goal — a secondary aim after the quality of the interval estimate for sampling variance. To clarify, he explains that the approximately unbiased estimation of legitimate parameter, *i.e.* population variance, is a legitimate goal. However, the bias of the estimated sampling variance associated with an estimate of a population quantity is not directly relevant, except to be a rough guide to show what might happen in large samples. To reiterate, what matters most is the coverage of the resulting interval estimates — an area that MI analyses are particularly strong in, despite the cases where point estimates are biased.

Nielson (2003) writes that a consistent estimator can always be constructed (although probably with added complexity), and even a more efficient estimator could be found based on the same amount of simulation work. However, it should be noted that introducing inefficiencies through simulation is the only way that data constructors can separate themselves from the data analysers, as should be the case with the construction and analysis of large survey data sets, and as is the case in the context for which Rubin's regular combining rules were created. Nielson's (2003) final suggestion for an approximately Bayesian imputation method is as follows:

1. Find the observed data MLE, $\hat{\theta}$.
2. Find a consistent estimator of the Fisher Information, $\hat{I}$.
3. Draw $\tilde{\theta}_{nj}$ from $N\left(\hat{\theta}_n, \hat{I}^{-1}/n\right)$.
4. Construct multiple imputations by drawing $\tilde{X}_{ij}$ from the conditional distribution of X_i given Y_i using $\tilde{\theta}_{nj}$ as the parameter.

This will give proper imputations when the complete data estimator is the MLE. In summary, Nielson (2003) suggests that not all Bayesian MI is proper, and even when it is proper, it is inefficient (Meng & Romero 2003). Rubin (2003*a*) clarifies that the Bayesian MI may be inefficient when the complete-data analysis is inefficient.

In Meng & Romero's (2003) discussion of Nielson's (2003) paper, the authors discuss self-efficiency once more, and the regular variance combining rule given by Rubin (1987). Meng & Romero (2003) believe that the complete-data estimator need only be self-efficient, as discussed in the practical guidelines given by Meng (1994) and listed above. If the complete data analysis method is self-efficient, the method cannot "improve upon itself by applying the same procedure to a part of the same data" (Meng & Romero 2003, p. 608). To recap, in mathematical terms Meng (1994) defines self-efficiency as follows, and Meng & Romero (2003) add to the definition in terms of selection by MDM:

Definition: Let W_c be a data set, and let W_o be a subset of W_c created by a selection mechanism (or *specified* missing-data mechanism). A statistical estimation procedure $\hat{\theta}(.)$ for θ is said to be *self-efficient* (with respect to the selection mechanism) if there is no $\lambda \in (-\infty, +\infty)$ such that the mean-squared error of $\lambda\hat{\theta}(W_o) + (1 - \lambda)\hat{\theta}(W_c)$ is less than that of $\hat{\theta}(W_c)$.

A researcher must have self-efficiency of their estimation procedure so that they cannot provide more efficient estimators by using less data. With this condition being met, the researcher is one step closer to providing valid inferences when using the regular MI combining rules.

Rubin (2003*a*) believes that the main idea is not to get a perfect variance estimate, but rather to obtain a variance estimator confidence interval that has nominal coverage, and a good average width. In summary, Meng & Romero (2003) suggest that it may be useful to sacrifice some efficiency for tractability (if the resulting method is still valid). Rubin (2003*a*) goes so far as to say that even some validity could be lost (for example, with large fractions of missing information), the benefits from using MI outweigh these efficiency and validity losses.

2.2.6 The beauty of multiple imputation

There are multiple sources of uncertainty in MI. Rubin (2003*a*) points out that these often complement each other to make MI "self-correcting" for approximately valid statistical inference. Rubin (2003*a*) lists these three forms of uncertainty:

1. There is almost always uncertainty in choosing the correct imputation model and MDM (ignorable or nonignorable)

2. Even with complete knowledge of the form of an imputation model governed by unknown parameters, there is uncertainty in the parameters' values used to create the imputations.

3. Given both the imputation model and its parameters, there is residual uncertainty to be reflected when drawing imputed values

MI can reflect all of these uncertainties: the first, by drawing under different imputation models, the second, by randomly drawing parameters from their posterior distributions and thereby attempting to make the MI "proper" or "confidence proper" (see Rubin 1976, Rubin 1996), and the third, by randomly drawing imputed values from their predictive distribution, given the fixed parameters drawn previously.

Zhang (2003, pp. 581, 584) lists the three uncertainties in a slightly different way:

1. Uncertainty due to modelling the joint distribution of the response variables and the missingness indicators, *i.e.* the uncertainty from $P(Y, R|\theta, \phi)$.

2. Uncertainty due to the sampling from a given imputation model assuming that the observed data and the values of the model parameters are known, *i.e.* the uncertainty from $P(Y_{mis}|Y_{obs}, \theta)$

3. Uncertainty about the values of the model parameters; the uncertainty for selecting the imputation model; *i.e.* uncertainty from $P(\theta|Y_{obs})$.

According to Rubin (2003*a*) one can also use MI to investigate changes in the completed-data inference resulting from changes in the assumed process for creating missing data, when there is a desire for such sensitivity analysis, for example, testing whether the missing data mechanism is MCAR or MAR, since, in the former case, imputation will not change the complete-case analysis results, while in the latter case, results may change significantly.

Rubin (2003*a*) believes that the combining rules for multiply imputed data from a sensible but imperfect model will lead to slightly conservative inferences (coverage slightly larger than nominal). In other words, the MI and combining process is self-correcting with the result that imperfect MI tends to be confidence proper.

From Nielson's (2003) paper, MI is self-correcting. This is because the between-to-within variance ratio is upwardly biased, leading to small degrees of freedom attached to the resulting inference, and thus valid MI confidence intervals with a realistic number of imputations even when the MI variance estimate is downwardly biased. So, the parameter estimates have distributions that are heavy-tailed, leading to more uncertainty even if the estimate itself is too low.

Another advantage of MI is concerned with the ignorability of the MDM. Rubin (2003*a*) suggests that researchers should not be held up by the belief that their MDM is nonignorable. The primary concern should be to build a MI procedure around an ignorable MDM that builds the relationships between variables, and to then test the sensitivity to a nonignorable model. This is because one cannot determine whether the mechanism is, in fact, nonignorable, by the very definition of nonignorability.

It can also be shown that the MI analysis and maximum likelihood estimation techniques often produce similar results (in large samples and with diffuse priors) if their distributional assumptions are equivalent (Schafer 2003). This would seem to make MI an unnecessary evolution. However, the critical point to remember is that MI allows the separation of the data collection and analysis tasks, which remains a particularly high priority in the large-sample survey data sets that are typically made available today. One alternative to MI, the design-based approach (Fay 1992), places significant burden on the data analyser, rather than allow a specialised imputer handle the missing data problem. So, in essence the MI procedure can be entirely modular (van Buuren 2007), splitting the imputation and analysis tasks between the two parties.

Meng (1994) suggests that even with inferential uncongeniality, one can have superior repeated-imputation inference in terms of validity and efficiency. Schafer (2003) also argues that uncongeniality between imputation and analysis models may even be a good thing, as opposed to how it was originally viewed, *i.e.* as a characteristic detrimental to the imputation and analysis procedures (Fay 1992). Schafer (2003) mentions that there is much practical evidence to suggest that under uncongeniality imputation and analysis has fared rather better than expected. Additionally, with uncongenial models of imputation and analysis, if the imputer's model is reasonably accurate, inferences with serious non-response bias are avoided, if not eliminated completely.

It is well documented that, while the rules given by Meng (1994) concerning congeniality and efficiency are profound guidelines for correct MI procedures, it suffices to have an imputation model that is more general than the analysis model in order to obtain good, valid MI results, even regardless of the MDM (see, for example Rubin 1978, Rubin 2003*a*, Schafer 2003, Zhang 2003). The importance of this generality is two-fold: to link the analysis and imputation procedures (adhering better to the guidelines and rules given in this chapter) and to make any (possibly) nonignorable MDM more likely to be able to be ignored. One of the arguments for a more general imputation model than the analysis model is made by Zhang (2003), pointing out that if variables are used in the analyses but not in the imputation model then the correlations between these omitted variables and the imputed variables will be biased towards zero. Thus, it is important to recognise this uncongeniality, not to invalidate the repeated-imputation estimators, but to ensure correct interpretation of the conclusions from MI inferences since these inferences may include information the imputer has that the analyst does not.

However, even if an imputer does not feel at ease embracing uncongeniality between analysis and imputation models, there is a way out. If an analyst wants inferential congeniality (maybe to take account of a variable that was not used in the imputation model that the analyst wanted in the model), then the reweighted combining rules first mentioned by Meng (1994) can be used to steer towards a better estimator.

So, in essence, MI has been derived for the specific Normal case, with specific rules for both imputation and analysis procedures in order to obtain perfectly unbiased, valid, efficient estimates in the analysis. However, these rules can be relaxed in numerous ways without greatly affecting unbiasedness, validity, and efficiency. If the restrictions are relaxed in any one of several ways, approximately unbiased and generally valid estimates are still obtained, these estimates being more efficient than those based on the incomplete data. The process becomes entirely modular, allowing the imputer and analyser to be separate entities. The precise science that would yield perfect results can be reduced to an art form that, when incorporated into scientific analyses, makes the answers from those analyses more scientifically agreeable than naïve results.

2.3 Recent Advances in Multiple Imputation

2.3.1 Other multiple imputation procedures

The MI method developed by Rubin was first implemented assuming a multivariate Normal distribution for the data. Several other MI procedures besides the regular multivariate Normal (MN) method presented in this chapter have been proposed. For MI from discrete data, these procedures include the well known hot-deck (HD) imputation method, variations on the HD procedure as explained by Ambler, Omar & Royston (2007), the Bayesian Bootstrap (BB), proposed by Rubin (1981), and the Approximate Bayesian Bootstrap (ABB), proposed by Rubin & Schenker (1986). One method for imputation from continuous data is the Normal method that adjusts for uncertainty in the mean and variance (NMV), reviewed by Rubin & Schenker (1986). These and the fully Normal imputation method are reviewed by Rubin & Schenker (1986), and are found to yield similar results (in the context of coverage intervals) in the presence of non-Normality, with results improving as m was increased. The best intervals are obtained from the NMV method.

Other methods for continuous imputation on montone patterns of missingness are the propensity score (PS) method, proposed by Lavori, Dawson & Shera (1995), and the predictive model (PM) method by Little & Rubin (2002), both reviewed by Zhang (2003). These methods are reviewed in detail in the subsequent subsection.

Discrete imputation

- **Hot-deck imputation (HD)**. In this MI method, the imputed values for Y_{mis} are drawn with replacement from Y_{obs}, where each element of Y_{obs} has equal probability of being drawn. Unfortunately, since the parameter of the data, θ, is not drawn from its own posterior distribution, and draws from a predictive posterior distribution conditional on this θ are not made, the HD method underestimates uncertainty. Moreover, observed outliers will have a greater influence on the post-imputation analyses, since these outliers will form part of the donor pool for missing values (Ardington et al. 2006).

- **Hot-deck imputation by covariate pattern ($\mathbf{HD}_{\mathrm{CP}}$).** This modification of the HD procedure matches fully observed categorical variables (or categorised continuous variables) for observations with missing values. These matches form the donor pool from which imputations are drawn. Note that only predictors are imputed this way.

- **Hot-deck imputation by observation ($\mathbf{HD}_{\mathrm{obs}}$).** In this modification of the $\mathrm{HD}_{\mathrm{CP}}$ procedure, the entire observation is replaced by a fully observed match, with the matching made once more through fully observed categorical variables (or categorised continuous variables). Once more, only predictors are replaced this way.

- **Hot-deck imputation including outcome ($\mathbf{HD}_{\mathrm{Y}}$).** Two variations of this modification of the $\mathrm{HD}_{\mathrm{CP}}$ procedure exist: one where the outcome variable is the only variable used to find matches from which imputed values are drawn, and the other, where both the outcome and predictors are used to find matches from which imputed values are drawn.

- **Bayesian Bootstrap (BB).** This method improves on the HD method by incorporating uncertainty in θ. Suppose that each element of the population takes one of the values $d_1, \ldots, d_K$ with probabilities $\theta_1, \ldots, \theta_K$, respectively. If the improper Dirichlet prior with density $\propto \prod_{k=1}^{K} \theta_k^{-1}$ is placed on the vector $\theta = (\theta_1, \ldots, \theta_K)$, then the posterior distribution of θ is the Dirichlet distribution with density $\propto \prod_{k=1}^{K} \theta_k^{q_k - 1}$ and K-dimensional mean vector $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_K)$ having components given by $\hat{\theta}_k = q_k / n_{obs}$, where $q_k =$ (number of times d_k appears in Y_{obs}). The BB MI method first draws θ^* from this posterior for θ. Then the components of Y_{mis} are independently drawn from among $d_1, \ldots, d_K$ using the probabilities in θ^*.

- **Approximate Bayesian Bootstrap (ABB).** This method is more computationally direct than the BB method. First draw n_{obs} observations with replacement from Y_{obs}. Then from this sample, draw n_{mis} observations with replacement as imputations. In this way, rather than drawing θ from a Dirichlet posterior distribution, this method draws θ from a scaled multinomial distribution. The distributions for θ in this method have the same means and correlations as in the BB method, but have $(n_{obs} + 1)/n_{obs}$ times the variances. This method is approximately equivalent to choosing the values of θ for the conditional posterior predictive distribution $P(Y_{mis}|Y_{obs}, \theta)$ from the observed data posterior distribution $P(\theta|Y_{obs})$. Note that the former HD procedure modifications can also be combined with the ABB procedure, as is done in Ambler et al. (2007). Also note that a detailed study on MDMs is made by Siddique & Belin (2008), in which the authors make extensive use of various adaptations of the ABB procedure. This study will be reviewed in Subsection 2.3.3.

Continuous imputation

- **Multivariate Normal (MN)**. The MN method assumes that all variables in the imputation model follow a joint multivariate Normal distribution. This is one of the most commonly used MI models. Implementation often uses a Bayesian MCMC approach, which appropriately allows for uncertainty in the estimation of the model parameters as well as uncertainty in the imputation draws themselves. The development of this method is attributed to Schafer (1997). The MN method first calculates parameter estimates for the joint probability distribution using the expectation-maximisation (EM) algorithm. The Gibbs sampler is then used to iteratively sample values for the missing data given the observed data and the current estimates, then new parameter estimates given the observed data and currently sampled values, and the process repeats itself. Schafer (1997) describes methods to determine the number of iterations that should be completed between each data set drawn as one of the m imputed data sets.

- **Normal method with uncertainty in mean and variance (NMV)**. In this method the observed data can influence the Normal shape of the distribution of the imputed values for Y_{mis}. Let $\bar{Y}_{obs}$ be the mean of Y_{obs} and let s^2_{obs} be the variance of Y_{obs}. First draw (μ^*, σ^2_*) from the posterior distribution of (μ, σ^2); *i.e.* σ^2_* is drawn from $(n_{obs} - 1)s^2_1/\chi^2_{n_{obs}-1}$, and then μ^* is drawn from $N(\bar{Y}_{obs}, \sigma^2_*/n_{obs})$. Then the components of a n_{mis}-dimensional vector $X = (X_1, \ldots, X_{n_{mis}})$ are drawn with replacement from Y_{obs}. Under repeated draws from Y_{obs}, each $Z_i = (X_i - \bar{Y}_{obs})/[(n_{obs} - 1)s^2_1/n_{obs}]^{1/2}$ has expected value 0 and variance 1. Finally, the n_{mis}

components of Y_{mis} are set equal to $\mu^* + \sigma_* Z_i, i = 1, \ldots, n_{mis}$. In this way the shape of the Normal distribution is influenced by the observed data.

- **Cubic spline regression method (CSR).** This method is utilised in the study by Faris, Ghali, Brant, Norris, Galbraith & Knudtson (2002). Cubic spline regressions, conditional on all other variables from the completely observed part of a data set, are used to generate imputations. This method does not assume the variables in the analysis conform to a joint distribution. Following the conditional modelling, errors for imputations are obtained by bootstrapping errors from the completely observed data, using the ABB method.

- **Markov Chain Monte Carlo method (MCMC).** This method, reviewed by Zhang (2003), is also known as the conditional modelling method. The method encompasses the basic idea behind sequential regression multiple imputation (SRMI), a major topic of this thesis, so the MCMC method will be reviewed in more detail in Chapter 3.

2.3.2 Imputation on monotone patterns

Imputation procedures for missing data with a monotone pattern deserve some attention. Valid, proper MI in such data sets can be formed by stringing together as many univariate imputation schemes as there are incomplete variables, as long as the pattern is monotone-distinct (see pp. 174–178 in Rubin 1987). A monotone pattern is monotone-distinct if the priors on $\theta_i, i = 1, \ldots, p$ are *a priori* independent in the following factorisation:

$$\begin{aligned} f\left(Y_i|X_i,\theta\right) &= f_1\left(Y_{i1}|X_i,\theta_1\right) f_2\left(Y_{i2}|X_i,Y_{i1}\theta_2\right)\ldots \\ & \quad \ldots f_p\left(Y_{ip}|X_i,Y_{i1},\ldots,Y_{i,p-1}\theta_p\right) \end{aligned} \tag{2.76}$$

i.e., if $\Pr(\theta) = \prod_{j=1}^{p} \Pr(\theta_j)$.

Another method for monotone missingness is the propensity score method, which is different to this univariate-chain predictive model type of method, and does not require the assumption that the pattern be monotone-distinct.

- **Propensity score method (PS)**. In this method, summarised by Zhang (2003), missingness is predicted using a linear logistic regression model. The conditional probability of observing y_{ij}, given the history, $y_{i1}, \ldots, y_{i,j-1}$, can be called a propensity score s_{ij}. That is, $s_{ij} = \Pr(r_{ij} = 0 | y_{i1}, \ldots, y_{i,j-1})$, where the r_{ij} are the elements of the response indicator matrix R. Since the missing data have a monotone pattern, the propensity score can be modelled by:

$$\log\left(\frac{s_{ij}}{1 - s_{ij}}\right) = \beta_0 + \beta_1 y_{i1} + \cdots + \beta_{j-1} y_{i,j-1}, \tag{2.77}$$

  where $\beta_0, \beta_1, \ldots, \beta_{j-1}$ are the regression coefficients. Once these coefficients are estimated from the observed r_{ij} for the response variable Y_j and the complete data for the covariates $Y_1, \ldots, Y_{j-1}$, each observation can be assigned an estimated propensity score,

$$\hat{s}_{ij} = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 y_{i1} + \cdots + \hat{\beta}_{j-1} y_{i,j-1}\right)}{1 + \exp\left(\hat{\beta}_0 + \hat{\beta}_1 y_{i1} + \cdots + \hat{\beta}_{j-1} y_{i,j-1}\right)}, \tag{2.78}$$

  and then all observations are stratified in q strata based on the quantiles of these estimated scores. Within each stratum a donor pool is created by applying the ABB method to the observed cases within each stratum, in order to reflect additional uncertainty about the posterior distribution of the underlying parameters. Each Y_{mis} is then imputed from a draw from the applicable donor pool. Multiple imputations can be made by creating conditionally independent donor pools and imputing a single value from each donor pool. It is important to note that this method cannot preserve any correlations that might exist between the actual history data, $(Y_1, \ldots, Y_{j-1})$, and Y_j. Only the missingness indicators in R and the history are being considered; the joint distribution of Y_j and $(Y_1, \ldots, Y_{j-1})$ is not being modelled.

- **Predictive model method (PM)**. This method, reviewed by Zhang (2003), is also meant for data with a monotone pattern of missingness. In these cases, the joint observed data likelihood function can be factored into the independent observed data likelihood functions:

$$L\left(\theta_1,\ldots,\theta_p|Y_{obs}\right)=\prod_{j=1}^{p}L\left(\theta_j|Y_{obs}\right)$$

where

$$L\left(\theta_j|Y_{obs}\right)\propto\prod_{i=1}^{n_j}P\left(y_{ij}|y_{i1},\ldots,y_{i,j-1},\theta_j\right),$$

and $P(Y_j|Y_1,\ldots,Y_{j-1},\theta_j)$ is the conditional distribution of Y_j given $Y_1,Y_2,\ldots,Y_{j-1}$ and θ_j is the conditional distribution parameters.

If a multivariate Normal is assumed for the response variables $Y_1,\ldots,Y_p$, then $L(\theta_j|Y_{obs})$ becomes a linear regression of Y_j on $Y_1,\ldots,Y_{j-1}$ using the first $n_{obs,j}$ observations, and the conditional distribution parameters θ_j become the regression coefficients and the residual variance. The missing values of Y_j can be imputed from the predicted values from this regression, given the observed values of $Y_1,\ldots,Y_{j-1}$ and the simulated regression parameters which are randomly drawn from from their observed data posterior distributions (created using uninformative priors). In this way, the extra uncertainty concerning the regression parameters is reflected in the imputations.

This method is closely related to SRMI, discussed in Chapter 3. To uncover the relationship, the computational aspects of the multivariate Normal-based PM procedure will be discussed below.

Let X_{obs} and X_{mis} be the rows of the data matrix X corresponding to $Y_{obs,j}$ and $Y_{mis,j}$ respectively. The probability model of Y_j given $Y_1,\ldots,Y_{j-1}$ is univariate Normal, $Y_j\sim N_1(\mu_j,\sigma_j^2)$, where $\mu_j=\beta_o+\beta_1Y_1+\cdots+\beta_{j-1}Y_{j-1}$, the observed data likelihood function of the regression parameters $\theta_j=(\beta_0,\beta_1,\ldots,\beta_{j-1},\sigma_j^2)$ is:

$$L\left(\mu_j,\sigma_j^2|Y_{obs}\right)\propto\sigma_j^{-n_{obs,j}}\exp\left[-\frac{1}{2\sigma_j^2}\sum_{i=1}^{n_{obs,j}}\left(y_{ij}-X_{obs(i)}\beta\right)^2\right] \tag{2.79}$$

where $\beta=(\beta_0,\beta_1,\ldots,\beta_{j-1})$ is a vector of regression coefficients. A non-informative prior for (β, σ_j^2) is assumed, *i.e.* there is a flat prior on β and $\pi(\sigma_j^2) \propto \sigma_j^{-2}$. After some manipulation, the observed data posterior distribution of (β, σ_j^2) can be expressed as:

$$\sigma_j^{-p/2} \exp\left[-\frac{1}{2\sigma_j^2}\left(\beta - \hat{\beta}\right)' \left(X'_{obs}X_{obs}\right)\left(\beta - \hat{\beta}\right)\right] \times$$
$$\sigma_j^{-\left(n_{obs,j}-p\right)/2-1} \exp\left[-\frac{1}{2\sigma_j^2}\sum_{i=1}^{n_{obs,j}}\left(y_{ij} - X_{obs(i)}\hat{\beta}\right)^2\right]$$

which is the product of a multivariate Normal and a scaled inverted chi-square distribution:

$$\beta | Y_{obs}, \sigma_j^2 \sim N\left(\hat{\beta}, \sigma_j^2\left(X'_{obs}X_{obs}\right)'\right), \tag{2.80}$$

and

$$\sigma_j^2 | Y_{obs} \sim \left(Y_{obs,j} - X_{obs}\hat{\beta}\right)' \left(Y_{obs,j} - X_{obs}\hat{\beta}\right) \chi^{-2}_{n_{obs,j}-p}, \tag{2.81}$$

where $\hat{\beta} = (X'_{obs}X_{obs})^{-1}X'_{obs}Y_{obs,j}$ is the MLE of β from the observed data likelihood function and $Y_{obs,j} - X_{obs}\hat{\beta}$ is the residual vector.

The values of (β, σ_j^2) can be simulated from their observed data posteriors (2.80) and (2.81) by drawing u from $\chi_{n_{obs,j}-p}$, then letting

$$\sigma_*^2 = \frac{\left(Y_{obs,j} - X_{obs}\hat{\beta}\right)' \left(Y_{obs,j} - X_{obs}\hat{\beta}\right)}{u}. \tag{2.82}$$

Then draw β^* from (2.80) given σ_*^2.

After a random draw from (β^*, σ_*^2) has been taken from their observed data posterior distribution, the missing values of Y_j, $Y_{mis,j}$, are imputed by independent random draws from the conditional predictive distribution $N_1(X_{mis}\beta^*, \sigma_*^2)$.

To obtain m sets of multiply-imputed data, m conditionally independent random draws are taken from the observed data posteriors (2.80) and (2.81), say $(\beta^{*(t)}, \sigma^2_{*(t)})$, $t = 1, \ldots, m$. For each of these sets of regression parameter draws the $Y_{mis,j}$ are imputed by independent random draws from $N_1(X_{mis}\beta^{*(t)}, \sigma^2_{*(t)})$.[2]

In essence, using the same procedure as in the PM method with the univariate Normal distributions, a monotone pattern of missing data can be imputed from a single run through a Gibbs sampler, with any set of univariate distribution models, as long as the pattern is monotone-distinct. The imputations are made in a univariate fashion, variable by variable, in the following manner:

$$\begin{aligned} P\left(Y_{mis,1}|X,\theta_1\right) \\ P\left(Y_{mis,2}|X,Y_1^*\theta_2\right) \\ \cdots \\ P\left(Y_{mis,k}|X,Y_1^*,\ldots,Y_{k-1}^*\theta_k\right) \end{aligned}$$

where k is the number of incomplete variables in the data set, and ,Y_j^* stands for the jth completed variable, as before (van Buuren 2007).

A multitude of imputation methods is available for univariate missing data, all of which can be used in this chain of univariate distributions. For a comprehensive list of methods and sources, see van Buuren (2007, p. 226). The sequence can be replicated m times to create m multiply imputed data sets.

2.3.3 Research into the mechanisms behind missing data

Nonresponse is a difficult and complicated problem especially because there is never any hard evidence in the data set itself to contradict relevant aspects of the assumptions on the MDM (Rubin 1978).

After seeing many practical MI examples Rubin (2003*a*) states that the MDM may not be the most important aspect to focus on in incomplete-data problems. From his experience, he suggests that the most important procedure to follow, at a minimum, is to model the missing data using an ignorable mechanism that builds relationships among the observed variables. Then, if necessary, and if it is the focus of the research, one can continue

[2] As will be seen in Chapter 3, this procedure is identical to one round in SRMI if all of the incomplete variables are univariate Normal.

investigating any MNAR mechanisms that might be occurring. While Rubin (2003*a*) does not suggest that MNAR MDM research is unimportant, he re-emphasises that at least a MAR MDM should be assumed when there is missing data.

Schafer (2003) agrees, suggesting that as a general principle more time should be spent on building an 'intelligent' model for the data, rather than spending too much time on modelling the MDM, unless the departure from a MAR MDM is expected to be severe. Collins et al. (2001) show that, with continuous data, an incorrect MAR assumption does not have a major impact on estimates or standard errors unless the absolute correlation between the the cause of missingness and the incomplete outcome is above 0.5. However, should a significant departure from a MAR MDM exist, the problem can be dealt with using the methods reviewed later in this subsection, Subsection 2.3.3.

Zhang (2003), too, mentions that often the MDM could have parts that are ignorable and parts that are not. However, as long as the ignorable part is substantial, and the nonignorable part is included in the imputation process, "the bias caused by treating all of the missing data as ignorable would be negligible" (Zhang 2003, p. 590–591).

Indeed, the MI process can incorporate even uncertainty concerning a MNAR MDM. "As long as the imputations reflect the correct amount of uncertainty there is nothing in the theory of MI that prevents appropriate inferences under $P(Y|X, R = 0)$. Multiple imputation will also work for nonignorable response mechanisms" (van Buuren 2007, p. 223).

One alternative method to account for a non-MAR MDM is proposed by Schafer (2003). This method involves creating extra covariates from the random variable R. Besides this possibility, there have been several attempts to correctly model a MNAR MDM. Some of these studies are reviewed below.

Glynn et al., 1993

Glynn et al. (1993) illustrate the application of MI to the estimation of means and regression parameters when nonresponse is nonignorable, and follow-up data is available for a sample of nonrespondents. The authors use a mixture model for MI; a mixture of a population model for respondents and a population model for the follow-up data obtained from nonrespondents. The imputation models used (both Normal and ABB) are derived using information from the nonrespondents that provide data once followed up.

Glynn et al. (1993) show that the use of this mixture model implemented through MI corresponds to a fully Bayesian mixture model or the classical double-sampling estimate (see, for example, Cochran 1977) in the simplest of cases, and yet it performs well in relatively more complex situations as well. The procedure works well, even when the underlying mixture model is incorrect. One of the most pertinent conclusions drawn by the authors is that, using their method, there is no need to specify a model for the probability of nonresponse, even though this model is nonignorable. It is assumed that the follow-up data better represents the missing data, so the models built from the follow-up data skew the imputations towards the correct nonignorable MDM. The authors also show that the fraction of missing information in incomplete data with a nonignorable MDM is larger than when the MDM is ignorable, so care should be taken to at least increase the number of imputations (in the case of Normal data).

Van Buuren et al., 1999

Van Buuren et al.'s (1999) paper applies a form of MI on incomplete blood pressure variables (systolic and diastolic measures) in order to examine the relationship between blood pressure and mortality (measured through a fully observed censoring flag and survival date) in elderly individuals when controlling for age and sex (and health in one model). The two most important aspects of this study are (i) that van Buuren et al. (1999) perform MI on individual variables sequentially, and (ii) that the authors multiply impute under both MAR and MNAR mechanisms. The former technique introduces the SRMI method. The latter exploration sheds light on how *a priori* assumptions can be used to model data under the MNAR mechanism, and how using differing mechanisms can inform the researcher about the underlying MDM. This part of the study will be reviewed in the current subsection. On the same general topic as Meng (1994) introduced in his paper concerning departure from a MAR MDM, van Buuren et al. (1999, p. 682) write:

> "Of particular interest is that MI allows display of the sensitivity of the in-

> ferences to different mechanisms that could have created the non-response. There is no need to assume one 'true' response model and stick to that. Several plausible mechanisms can be tried. If none of these mechanisms changes the relation of interest, then inference is robust against the specified causes of the non-response. On the other hand, if the results do depend on the specific form of the non-response model, then more precise statements can be made regarding the exact conditions under which the obtained results apply."

Van Buuren et al.'s (1999) overlying model is Cox regression of mortality on blood pressure, adjusted for age and sex (and health). Approximately 12.5% of cases are missing one or more blood pressure measurements, and, importantly, it is expected that individuals with lower blood pressure are more likely to have the blood pressure measurements missing. This means that the data might be MNAR.

Assuming that the data is merely MAR, a single variable is imputed as follows (van Buuren et al. 1999):

1. Calculate $W = (X'_{obs}X_{obs})^{-1}$, $\hat{\beta} = WX'_{obs}Y_{obs}$, and $\hat{Y}_{obs} = X_{obs}\hat{\beta}$.
2. Draw a random variable c from the $\chi^2_{n_{obs}-r}$ distribution, where r is the number of predictors for Y.
3. Calculate $\sigma^2_* = (Y_{obs} - \hat{Y}_{obs})'(Y_{obs} - \hat{Y}_{obs})/c$
4. Draw an r-dimensional Normal random vector $D \sim N(0, I_r)$.
5. Calculate $\beta^* = \hat{\beta}+\sigma_* W^{1/2}D$, where $W^{1/2}$ is the triangular square root of W obtained from the Cholesky decomposition.
6. Calculate predicted values $Y^*_{mis} = X_{mis}\beta^*$.
7. For each missing value $i = 1, \ldots, n_{mis}$, find the respondent whose $\hat{Y}_{obs}$ is closest to $Y^*_{mis,i}$, and take Y_{obs} of this respondent as the imputed value for case i.
8. Repeat steps 2–7 m times to create the m (multiply) imputed data sets, $Y^{(1)}_{mis}, Y^{(2)}_{mis}, \ldots, Y^{(m)}_{mis}$.

This method is very similar to that which will be introduced as the sequential regression Normal method, only differing in the penultimate two steps. It incorporates uncertainty due to deviations around the regression lines in steps 2 and 3, and also incorporates variation of the regression line itself due to finite sampling in step 4.

In order to impute multivariate data, the sequential univariate regressions are applied. By the way of Gibbs sampling, sampling from the univariate conditional distributions iteratively, the multivariate joint distribution can be approximated (if it exists). This process is essentially the method used by van Buuren et al. (1999) to impute the systolic and diastolic blood pressure measures.

Each incomplete entry is initialised by filling in a random draw from the marginal distribution of Y_{obs}.

1. Y_1 is imputed by the univariate procedure listed above, conditional on all other data (observed and imputed combined).

2. Y_2 is then imputed, conditional on all other data (using the most recent imputations for Y_1).

3. This process repeats itself until all the incomplete variables are imputed.

4. The first three steps are then repeated on a second pass of the data, using all imputations created during the first pass.

5. Step 4 is then repeated until 20 passes have been made. The data set that results after 20 passes is one imputed data set.

6. Step 5 is repeated m times to complete m imputed data sets.

The authors find that fewer iterations on each data set are required than for regular Markov chain simulation techniques. This type of algorithm converges usually within relatively few iterations, meaning the 20 iterations specified in the procedure above are more than sufficient.

Van Buuren et al. (1999) then add a location term to the imputation model to adjust for a MDM that is not random. The prior belief that missing blood pressure observations are more likely to occur at lower blood pressures suggests that an adjustment should be made to those cases where $R_i = 1$. Incorporating this idea into the regression switching method implies the addition of a location term to the imputation model for systolic blood pressure (the effect of the location parameter is carried over via the regression switching to the diastolic blood pressure imputations):

$$Y_1 = X\beta + (1 - R_1)\,\delta + \epsilon$$

where R_1 is the binary response indicator of systolic blood pressure, and δ is chosen by the imputer to be the mean unit difference between responders and non-responders in excess of that induced by X.

The Cox regression results on the imputed data sets are combined using the standard combining rules given by Rubin (1987), assuming that the sample is large enough to use the Normal confidence intervals for Q. Relative risk estimates and their confidence limits in the Cox proportional hazards model can then be obtained as $\exp(\hat{Q})$ and $\exp(\hat{Q} \pm 1.96\sqrt{T})$, with variables notated as previously. Various values of δ, $(0, -5, -10, -15, -20)$ are chosen for the analysis, and the results are compared. Of course, as delta becomes more negative, the point estimates for imputed values become lower. However, the proportional hazards for each blood pressure grouping remain similar to the complete-case analysis, regardless of the value of δ. The lack of change in the overall proportional hazards model may be due to the small amount of missing data, or because the differences in mortality between responders and nonresponders are simply too small to exert a serious impact on the estimates. The authors caution that although the imputation results are similar to the complete-case analysis results, the complete-case analysis was not necessarily appropriate in the first place.

Carpenter and Kenward, 2007

Carpenter & Kenward (2007) also apply sensitivity analysis on the MDM. The authors impute under the MAR mechanism, obtain parameter estimates for each imputed data set, and then obtain MNAR parameter estimates as a weighted average of the MAR parameter estimates, where the weights depend on the extent of the departure from the MAR mechanism. The weights are calculated using importance-sampling, or an application of the sampling importance resampling (SIR) algorithm, which Meng (1994) used in his study of non-congeniality in MI. Carpenter & Kenward's (2007) approach provides a simple approximate sensitivity analysis for imputers when the MAR and MNAR distributions overlap and the fraction of missing information, γ or $\hat{\gamma}$ from Subsection 2.2.3, is not too great. The method allows researchers to assess the necessity of MNAR modelling, and is detailed below.

Carpenter & Kenward (2007) model the response of patients in a longitudinal clinical trial as follows:

$$\begin{aligned}&\text{logit } \Pr(R_i = 1)\\ &= \beta_0 + \beta_1 \times \mathrm{I}_{0,1}\left[\text{patient } i \text{ on active treatment}\right] + \beta_2 X_i + \beta_3 Y_i \end{aligned} \qquad (2.83)$$

where $\mathrm{I}_{0,1}[\cdot]$ is an indicator variable with argument for $\mathrm{I}_{0,1} = 1$ in $[\cdot]$, β_0 is the log-odds of observing Y_i, β_1 is the adjusted change in the log-odds ration of observing Y_i if the patient is randomised to active treatment, and β_2 is the additional change in the log-odds ratio for a one-unit change in the baseline X_i, which itself need not be fully observed but will remain notated as X_i in the review of this study for the sake of simplicity.

Of course, if $\beta_3 = 0$ then the MDM is MAR. If $\beta_3 \neq 0$, then distributional assumptions will be needed to estimate β_3, since information about β_3 within the model may be scarce. As with the study by van Buuren et al. (1999), the authors therefore estimate the model in expression (2.83) assuming that $\beta_3 = 0$, *i.e.* that the MDM is MAR, and then suggest suitable values for the additional dependence of R_i on Y_i through fixed choices of β_3, testing the sensitivity to the each suggested MNAR mechanism. When $\delta \neq 0$, this process involves jointly fitting a model for the observed responses, Y_i, and model (2.83), which usually takes the form of numerical integration or MCMC modelling.

Let Z be the baseline data (baseline response and treatment allocation, *i.e.*, $\mathrm{I}_{0,1}$ and X together). As usual, Q is the scalar quantity of interest. Carpenter & Kenward (2007) make the standard assumption that without missing data, Q is Normally distributed. Assuming the MAR MDM, we have, as before, m versions of the missing data, and m parameter estimates and their variances, $\hat{Q}_{*l}, U_{*l}, l = 1, \ldots, m$. Also, as before, we have the regular MAR MI estimates as per Equations (2.5)–(2.8). To obtain an estimate of Q when the data are assumed to be MNAR, a suitable β_3 must be chosen. The patients are then re-ordered so that it is patients $i = 1, \ldots, n_*$ that have withdrawn and patients $i = n_* + 1, \ldots, n$ that are complete. Also, let $Y_i^{(j)}$ denote the the jth MAR imputation of patient i's response. Then, for each imputation j, compute:

$$\tilde{w}_j = \exp\left(-\beta_3 \sum_{i=1}^{n_*} Y_i^{(j)}\right), \tag{2.84}$$

and

$$w_j = \frac{\tilde{w}_j}{\sum_{k=1}^{m} \tilde{w}_k}. \tag{2.85}$$

The derivation of these weights given model (2.83) is straight forward. The authors use the sampling importance resampling (SIR) algorithm to determine the weights in a two-imputation case for a single observation, and extend the basic result thereafter. SIR is used when one wishes to estimate the mean of g, a probability distribution which is known up to a normalising constant, but can only readily sample from f, although the support of f and g coincide. Then:

1. Sample $Y^{(1)}, \ldots, Y^{(m)} \sim f$
2. Calculate $w_j = \frac{g\left(Y^{(j)}\right)}{f(Y^j)}$
3. Then we have that:

$$E_g\left[Y\right] \approx \frac{\sum_{j=1}^{m} w_j Y^{(j)}}{\sum_{j=1}^{m} w_j} \tag{2.86}$$

or

$$E_g\left[h\left(Y\right)\right] \approx \frac{\sum_{j=1}^{m} w_j h\left(Y^{(j)}\right)}{\sum_{j=1}^{m} w_j} \tag{2.87}$$

This approximation improves as $m \to \infty$. Now, in this algorithm, f represents the MAR MDM that we know we can draw from, *i.e.* $[Y|Z, R = 1]$, while g is the MNAR MDM that we would like to draw from, *i.e.* $[Y|Z, R = 0]$, and h is the completed data estimator, $\hat{Q}$. Looking at only a single observation (that is not in the treatment group) Y_1 with $R_1 = 0$

and X_1 observed. The weight then follows the ratio:

$$\begin{aligned}\frac{[Y_1|Z_1, R_1 = 0]}{[Y_1|Z_1, R_1 = 1]} &= \frac{[Y_1, Z_1, R_1 = 0]}{[Y_1, Z_1, R_1 = 1]} \frac{[Z_1, R_1 = 1]}{[Z_1, R_1 = 0]} \\ &= \frac{[R_1 = 0|Y_1, Z_1]}{[R_1 = 1|Y_1, Z_1]} \frac{[Z_1, R_1 = 1]}{[Z_1, R_1 = 0]} \end{aligned} \tag{2.88}$$

In model (2.83), for $[R_1 = 0|Y_1, Z_1]$, if the patient is in the non-treatment arm of the clinical trial, this is $\{1 + \exp[\beta_0 + \beta_2 X_1 + \beta_3 Y_1]\}^{-1}$. Thus, $[R_1 = 1|Y_1, Z_1]$ reduces to $\exp[\beta_0 + \beta_2 X_1 + \beta_3 Y_1]\{1 + \exp[\beta_0 + \beta_2 X_1 + \beta_3 Y_1]\}^{-1}$ Therefore, the weights for the two imputations, $Y^{(1)}$ and $Y^{(2)}$ are:

$$\tilde{w}_1 = \exp\left(-\beta_3 Y^{(1)}\right) \left\{ \exp\left[-\left(\beta_0 + \beta_2 X_1\right)\right] \frac{[Z_1, R_1 = 1]}{[Z_1 . R_1 = 0]} \right\}, \tag{2.89}$$

and

$$\tilde{w}_2 = \exp\left(-\beta_3 Y^{(2)}\right) \left\{ \exp\left[-\left(\beta_0 + \beta_2 X_1\right)\right] \frac{[Z_1, R_1 = 1]}{[Z_1 . R_1 = 0]} \right\}. \tag{2.90}$$

Since the terms in braces are identical, the normalised weights $w_1 = \tilde{w}_1/(\tilde{w}_1 + \tilde{w}_2)$ and $w_2 = \tilde{w}_2/(\tilde{w}_1 + \tilde{w}_2)$ are proportional to $\exp(-\beta_3 Y_1^{(1)})$ and $\exp(-\beta_3 Y_1^{(2)})$, respectively. It is easy to see that the indicator variable will also be a part of the common terms, as would any function of the observed data. This simple case is intuitively extended for all observations and all cases, giving us the weighting formula in Equation (2.84). Given the completed data, the probability of observing $Y_i^{(j)}$ is independent of the probability of observing $Y_{i'}^{(j)}, i \neq i'$. It then follows from Equations (2.89) and (2.90) that the weight for imputation m is:

$$w_j \propto \exp\left(-\beta_3 \sum_{i=1}^{n_*} Y_i^{(j)}\right). \tag{2.91}$$

Thus, the log-weight is proportional to a linear combination of the imputed data.

Then, under the MNAR model implied by the choice of β_3, the estimate of Q and its variance, U, are simply re-weighted versions of the regular combining rules given in Equations (2.5)–(2.8):

$$\bar{Q}_{\text{MNAR}} = \sum_{j=1}^{m} w_j \hat{Q}_j; \tag{2.92}$$

with variance:

$$T_{\text{MNAR}} \approx \bar{U}_{\text{MNAR}} + \left(1 + \frac{1}{m}\right) B_{\text{MNAR}}, \tag{2.93}$$

where

$$\bar{U}_{\text{MNAR}} = \sum_{j=1}^{m} w_j U_j, \tag{2.94}$$

and

$$B_{\text{MNAR}} = \sum_{j=1}^{m} w_j \left(\hat{Q}_j - \bar{Q}_{\text{MNAR}}\right)^2. \tag{2.95}$$

Carpenter & Kenward (2007) mention that the accuracy of the approximation justifying Equations (2.92) and (2.93) improves as the number of imputations increases (with problems often requiring $m \geq 50$). However, T_{MNAR} is still likely to underestimate the variance, even with large m, since the effective sample size after reweighting is often less than m. Therefore, the t-distribution degrees of freedom for the inferences based on the MNAR model should be decreased.

In their simulation study, Carpenter & Kenward (2007) show that as m tends to infinity, if the correct β_3 is chosen, the estimates will be unbiased. The bias that the MAR results offers is whittled away by increasing m and using the proposed reweighting scheme. One of the advantages of this method is that parameters other than β_3 need not be specified. So the method is robust to misspecification of the model in Equation (2.83), providing β_3 is correctly specified. Also note that $[Y|X, R = 1]$ is estimated from the observed data. If the fraction of missing information is large, the true distribution may not be summarised adequately by the observed data. As for application of this model to real data, Carpenter & Kenward (2007, p. 269) write,

> “it is important to know whether we have enough imputations for a reasonably

> reliable answer, and also whether the range of parameter estimates from the MAR model is sufficiently wide to give acceptable support to the [MN]AR distribution — the key assumption to this method."

So the MAR distribution's support should complement that of the MNAR mechanism, and the number of imputations should be large. The latter may be a problem when the analyser is not the imputer, since it is not common to publish that many multiply imputed completed data sets. It is also important that the choice of β_3 be guided by prior knowledge or from expert opinion, since there is usually insufficient information in the incomplete data to estimate this. One option given by the Carpenter & Kenward (2007, p. 274) is to "calculate the [β_3] corresponding to one standard error change in the parameter of interest, and then refer this value to experts to assess its plausibility."

In order to perform sensitivity analysis, the authors use two types of graphs. The first is a graph of the parameters versus the normalised weight for each imputation, with a line plotted at the MAR estimate of the parameter (where there are equal weights). This graph will show which imputations are being more highly weighted. The second graph is a running re-weighted estimate versus number of imputations, *i.e.*, a graph of the running re-weighted estimate as more imputed data sets are added to the weighted combining rules. Again a line is plotted for the unweighted MAR parameter estimate. These graphs will show the model's sensitivity to the MAR MDM, and will give close approximations to the MNAR estimates if they are close to the MAR estimates, or will at least point in the direction of the MNAR estimates if the MAR and MNAR distributions do not over

When discussing potential extensions to their study, Carpenter & Kenward (2007) mention that a check on the reliability of resulting MNAR imputations is whether the re-weighted estimates agree with those obtained under a MAR MDM.

Siddique and Belin, 2008

Siddique & Belin (2008) combine ABB and predictive mean matching (PMM) in multiple imputation on nonignorable missingness in incomplete data. In the basic ABB procedure, a new bootstrapped sample from Y_{obs} (size n_{obs}) is considered as being the set of donor cases for the n_{mis} imputation draws. Instead of each observation being drawn into the bootstrapped sample with equal probability, Siddique & Belin (2008) extend the method proposed by Rubin & Schenker (1991) where the n_{obs} observed values $y_i \in Y_{obs}$ have probability of being drawn equal to $\frac{y_i^c}{\sum_{j=1}^{n_{obs}} y_j^c}$. In Rubin & Schenker's (1991) method, for different values of c, the bootstrapped samples which offer the donor cases for imputation draws will tend to over-represent large values (if $c > 0$ and $y_i > 0$ for example) or over-represent small values (if $c < 0$ and $y_i > 0$ for example). In this way, more larger (or smaller) values will be imputed, according to the nonignorable MDM that is hypothesised. Alternative, other 'shapes' of donor cases can be generated, by, for example, sampling based on difference from the median or another quantile.

Siddique & Belin (2008) create the bootstrap sample of donors using Rubin & Schenker's (1991) method, but then the probability of being selected as a donor case also uses weights based on PMM. In PMM, values Y_{obs} in the bootstrapped sample are regressed on a their observed covariates, X. Using the parameters from this regression, predicted values are calculated for all observations, observed and missing, $\hat{Y}$. Missing values of Y are then drawn from the ABB donor or bootstrapped sample with a probability inversely proportional to the distance between the missing values' predictions and the donors' predictions. Without going too much into the technicalities, the distance formula incorporates a parameter k that allows the distancing formula to go from one extreme (nearest-neighbour HD) to the other (simple random HD). Thus, all donor cases have a non-zero probability of being selected as imputed values, but some will have higher probabilities than others, based on their proximity in predicted value to the predicted value of the point being imputed.

In their experiments, Siddique & Belin (2008) multiply impute using their ABB/PMM mixed method. They find that the best results are obtained when the nonignorable MDM is altered for each imputed data set (known as a Mixture ABB in their paper). The MDMs can be altered with a focus on a hypothesised nonignorability (favouring large values, for example), or they can be spread across several alternatives. In this way, the authors do not have to assume they know the nonignorable MDM, but the variation over imputed data sets represents sensitivity to the nonignorable MDM. In essence, a Mixture ABB MI allows imputers to test the assumption of a nonignorable MDM.

2.4 Conclusion

Since Rubin (1978) laid the groundwork for multiple imputation and ignorable MDMs, research into MI has increased substantially. However, the fundamental concepts have remained almost unchanged. The theory behind proper, valid multiple imputation has remained sound over the last four decades, and research has reached into the effects of nonignorable MDMs, MI of particular missingness patterns, and most importantly, new methods reducing the complexity of the joint modelling procedure, most notably SRMI.

However, over recent years some extensions have been attached to the age-old fundamental procedures, such as the use of larger numbers of imputed data sets for post-imputation Bayesian analysis, as discussed by Zhou & Reiter (2010).

The most profound outcome of the path of research and debate within MI has been the illustration, time and again, of its convenience and efficiency. It is clear that MI is now the preferred method for filling in incomplete data sets if those data sets are to be made available for later public use: when the imputation and analysis tasks are separated, MI is the most appealing solution to the missing data problem. Multiple imputation allows the uncertainty inherent in imputation to be carried over from the imputation task to the analysis tasks, including the uncertainty due to not knowing how the data became missing, uncertainty in the imputation model, and sampling uncertainty.

Chapter 3

Sequential Regression Multiple Imputation

3.1 Introduction

A recent MI approach uses sequences of appropriate regression models to multiply impute missing data. Hence the name Sequential Regression Multiple Imputation (SRMI). This approach is also known as the fully conditional specification (FCS) or approach (for reasons that will be explained shortly), or MI through chained equations (MICE), as well as stochastic relaxation, regression switching, variable-by-variable imputation, partially incompatible MCMC, iterated univariate imputation, or the ordered pseudo-Gibbs sampler. This thesis will refer to all of these methods with the common acronym SRMI, since they are all essentially the same procedure.[1]

One of the main problems of MI from a Bayesian context is that a multivariate model needs to be chosen for the observed data. In practice, however, survey data consists of many variables with different distributions, and often displays seemingly unsystematic patterns of missing data. These properties of survey data make joint modelling approaches extremely difficult to implement, since typical multivariate distributions are not flexible

[1] The reason SRMI is chosen above the more common FCS acronym, is due to the more explanatory nature of the terms 'sequential regression', which adequately describe some of the steps within the procedure.

enough to accommodate such varying structure.

In the sequential procedure, each variable can be modelled individually within the imputation process. Imputers can have much more control over imputations from variables with inherent restrictions. This is not easily done when variables are jointly modelled in an imputation procedure.

The SRMI method of MI was proposed by van Buuren et al. (1999), and independently by Raghunathan, Lepkowski, van Hoewyk & Solenberger (2001), although the system had been used even earlier by researchers such as Kennickell (1991).

3.2 The SRMI process

3.2.1 Overview

As explained by Raghunathan et al. (2001) and He & Raghunathan (2009), SRMI works, in essence, in a two-dimensional process as follows. Reviewing our standard notation, let $Y_j\,(j = 1, \ldots, p)$ denote the variables with missing values, X denote the matrix of q fully observed variables, and let $Y_{-j} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_p)$ denote the $p - 1$ variables in Y excluding Y_j. In SRMI, a conditional model $P\,(Y_j|Y_{-j}, X, \theta_j)$ is specified for each Y_j, with θ_j denoting the respective model parameters. The first dimension of the SRMI process is a single iteration, or pass, of the process, essentially 'filling in' the missing data values. The second dimension of the procedure is the repetition of this 'filling in' process, using the previously filled-in values. Thus, in each iteration of the imputation procedure, θ_j is drawn from $P\left(Y_j|Y_j^{obs}, Y_{-j}, X\right)$ using the observed part of the variable Y_j, namely Y_j^{obs}, and the completed part Y_{-j} (from the previous iteration if there was one), and X; the missing part of the variable Y_j, namely Y_j^{mis}, is then imputed. The conditional model process is repeated, cycling through all the Y_j's. Each conditional density is modelled through the appropriate regression model, chosen specifically based on the distribution of each variable.

Note that the first round of imputations, *i.e.* the first iteration, is slightly different, as mentioned above in the text "...from the previous iteration *if there is one*". Raghunathan et al. (2001) break down the first iteration in detail. The joint conditional density of $Y_1, Y_2, \ldots, Y_p$ given X can be factored as:

$$\begin{aligned} f\left(Y_1, Y_2 \ldots, Y_p | X, \theta_1, \theta_2 \ldots, \theta_p\right) = \\ f_1\left(Y_1 | X, \theta_1\right) f_2\left(Y_2 | X, Y_1, \theta_2\right) \ldots f_p\left(Y_p | X, Y_1, Y_2, \ldots, Y_{p-1}, \theta_p\right) \end{aligned} \quad (3.1)$$

where $f_j, j = 1, \ldots, p$ is the conditional density function and θ_j a vector of parameters for the respective conditional distribution. So the first iteration of the SRMI procedure conditions only on the data that has been filled in already in that iteration.

When the missing data have a non-monotone pattern, the target distribution is the joint conditional distribution of Y_{mis} and θ given Y_{obs}, $P(Y_{mis}, \theta | Y_{obs})$. One can simulate from this distribution using the MCMC method, as proposed by Zhang (2003), which proceeds as follows:

1. Replace the missing data Y_{mis} by some assumed values.
2. Simulate θ from the resulting completed data posterior $P(\theta | Y_{obs}, Y_{mis})$. Let $\theta^{(t)}$ be the current simulated value of θ from this complete data posterior distribution.
3. (Imputation or I-step): The next iterative sample of Y_{mis}, namely $Y_{mis}^{(t+1)}$, can be drawn from the conditional predictive distribution of Y_{mis} given Y_{obs} and $\theta^{(t)}$:

$$Y_{mis}^{(t+1)} \sim P\left(Y_{mis} | Y_{obs}, \theta^{(t)}\right) \quad (3.2)$$

4. (Posterior or P-step): Conditioning on $Y_{mis}^{(t+1)}$, the next simulated value of θ can be drawn from its completed data posterior distribution,

$$\theta^{(t+1)} \sim P\left(\theta | Y_{obs}, Y_{mis}^{(t+1)}\right). \quad (3.3)$$

5. Repeating steps 3 and 4 from a staring value of θ, say, $\theta^{(0)}$, yields a Markov chain $\{(\theta^{(t)}, Y_{mis}^{(t)}) : t = 1, 2 \ldots\}$. The stationary distribution is the target distribution, $P(Y_{mis}, \theta | Y_{obs})$.

Consequently, the marginal stationary distributions of the subsequence $\{\theta^{(t)} : t = 1, 2 \ldots\}$

and $\{Y_{mis}^{(t)} : t = 1, 2 \ldots\}$ are the observed data posterior distribution $P(\theta|Y_{obs})$ and the posterior predictive distribution $P(Y_{mis}|Y_{obs})$, respectively. When t is sufficiently large, $\theta^{(t)}$ can be viewed as a single simulation from the observed data posterior distribution $P(\theta|Y_{obs})$, and $Y_{mis}^{(t)}$ can be viewed as a single simulation from the posterior predictive distribution $P(Y_{mis}|Y_{obs})$.

Multiple imputations of Y_{mis} should be independent given Y_{obs}. For this reason, Zhang (2003) suggests that Y_{mis} draws only be taken from widely separated items along a single chain, or from the final draws on each of m chains. A comment on the efficiency of this type of sampling is given in Subsection 3.4.7.

Since the convergence from MCMC is distributional, *i.e.* the draws do not converge, yet the distributions from which they are drawn do, different methods have been proposed in order to monitor this convergence. One of these methods will be reviewed later in this chapter.

In the following section we explain the process behind SRMI in more detail, since this method is the focus of the present chapter, and indeed, this thesis. *Firstly*, the imputation steps can be explained in more detail, and, *secondly*, the Gibbs sampling procedure used in the application of SRMI can be reviewed thoroughly. Both of these explanations are presented in Raghunathan et al. (2001).

3.2.2 Step-by-step SRMI and Gibbs sampling

Given an incomplete dataset, the data set's incomplete variables are sorted from the variable with the least missing entries to the variable with the most missing entries. Let the variable with the most missing entries be the vector Y_1, the variable with the next fewest missing be Y_2, *etc.*, until Y_p. Let X again be the part of the dataset that is complete. Finally, let θ_j once more be the vector of unknown regression and dispersion parameters in the conditional model for Y_j. The sorting of the dataset follows as an extension to the fact that in model-based imputations the joint conditional density of $Y_1, Y_2, \ldots, Y_p$ given X can be factored as in Equation (3.1).

The first round of imputations then begins; the variable with the least amount of missing data entries (apart from the complete variables) is selected. This variable is regressed on the complete data according to a regression model that is assumed to fit the distribution of the variable, as mentioned above. The model first processed is illustrated in Equation (3.1) by f_1. The regression is Bayesian by nature, but utilizes a diffuse or non-informative prior. If $\Theta = (\theta_1, \theta_2, \ldots, \theta_p)$ then the prior for each model is $\pi(\Theta) \propto 1$. A set of regression parameters is then drawn from the regression model and a single draw from the posterior predictive distribution of the model (the predictive distribution of the missing values given the observed values) is made for every missing data entry in that variable. These draws are the imputed values for that variable.

The SRMI process then selects the variable with the next fewest missing values, and the procedures explained the above paragraph are repeated as follows. A new regression model, illustrated by f_2 in Equation (3.1), is chosen according to the assumed distribution of Y_2, the variable now being regressed. This new variable is regressed on the complete data and the newly completed variable from the previous step (*i.e.* the variable with the least missing values, all of which have now been imputed with a single imputation). Again a set of regression parameters is drawn from the new regression model and a single draw from the posterior predictive distribution of this model is made for every missing data entry in the variable. This step is repeated until all of the variables in the dataset are filled in by appropriate regression predictions. By the nature of this process, the term 'sequential regression imputation' is justified.

The reason that the data is sorted according to missingness is given by the fact that the starting distribution in the Gibbs sampling procedure should be as close as possible to the target distribution $P(Y_{mis}|Y_{obs})$, since the Gibbs sampling procedure can be strongly influenced by the initial distribution (Brand 1998, p. 53). By filling in the data set variable by variable, from least missing to most missing, we obtain the best possible starting distribution.

Once an entire dataset has been filled in with imputed values for the original missing entries, this completed dataset is subjected to an update round, round two, starting essentially at the second step above. Thus, the iterative process involved in SRMI is brought to light. The process involved in the updating rounds differs slightly to that of steps two and three above.

The first difference depends on the pattern of the missing data. For a monotone pattern of missing data, if a datum for an observation is missing in variable Y_j, then the data for that observation will be missing in variables $Y_{j+1}, Y_{j+2}, \ldots, Y_p$. When this pattern occurs the imputations in the first round are approximate draws from the predictive distribution of the missing values given the observed values. Draws in subsequent rounds can be improved upon using the SIR (sampling, importance-weighting, resampling) or another rejection algorithm (Raghunathan et al. 2001). When the pattern of missing data is not monotone, a Gibbs sampling algorithm must be developed to improve upon the previous round's estimates. Raghunathan et al. (2001) suggest that the missing values in Y_j at round $(w + 1)$ need to be drawn from the conditional density:

$$f_j^* \left(Y_j | X, \theta_1^{(w+1)}, Y_1^{(w+1)}, \ldots, \theta_{j-1}^{(w+1)}, Y_{j-1}^{(w+1)}, \theta_{j+1}^{(w)}, Y_{j+1}^{(w)}, \ldots, \theta_p^{(w)}, Y_p^{(w)} \right) \tag{3.4}$$

where $Y_i^{(w)}$ is the vector Y_i that was filled in with imputed values in round w. Equation (3.4) is computed based on the joint distribution specified in Equation (3.1). This draw process would be extremely difficult to complete, since the density in Equation (3.4) is difficult to compute in most practical situations without restrictions (Raghunathan et al. 2001, He & Raghunathan 2009). However, Raghunathan et al. (2001) propose that, instead, the draw in round $(w + 1)$ for Y_j is taken from the predictive distribution corresponding to the conditional density:

$$g_j \left(Y_j | X, Y_1^{(w+1)}, Y_2^{(w+1)}, \ldots, Y_{j-1}^{(w+1)}, Y_{j+1}^{(w)}, \ldots, Y_p^{(w)}, \phi \right) \tag{3.5}$$

where ϕ is a vector of regression parameters with diffuse prior.

In other words, in imputation rounds after the first round the values that were originally missing in each variable are now predicted from regression models, regressing those variables on all of the other variables in the dataset. This process implies that the variables with values imputed from the first round are used as regressors in the second round, in addition to the newly updated variables from the current round. The process can be viewed as an approximation to the Gibbs sampling procedure in Equation (3.4). In some particular cases this approximation is equivalent to drawing values from a posterior predictive distribution under a fully parametric model. For example, if all of the variables are continuous and Normally distributed with constant variance, then the algorithm governing Equation (3.5) converges to a joint predictive distribution under a multivariate Normal distribution with an improper prior for the mean and covariance matrix (Raghunathan et al. 2001).

This fourth step is then repeated as many times as the researcher deems fit (usually to a point where the inferences made on the data during subsequent rounds converge).

3.3 Regression Models Used in SRMI

The SRMI modelling procedure is open to using new regression modelling procedures. Numerous ‘regular’ regression models exist that have been used in SRMI modelling; these models will be presented in Subsection 3.3.1. Besides these models, many other possible regressions exist that could be incorporated into an SRMI procedure. One such example is the family of generalised linear mixed models, which will be discussed in Subsection 3.3.3.

3.3.1 Imputing from generalised linear models

From Equation (3.5) it is evident that, in order to obtain predictions for the missing data in a variable, a particular regression model needs to be utilised according to the assumed distribution of the variable in question. Regular regression models considered are the Ordinary Least Squares (OLS) regression model for a variable that is Normally distributed, the logistic Generalised Linear Model (GLM) for a variable that is dichotomous or binary, the Poisson GLM for a variable that displays count data, and a polytomous regression model for variables with three or more categories.

In this section, note that the univariate outcome variable is denoted by Y, while the covariates used in the regression of Y are denoted by X, as opposed to the notation in this thesis that regards Y as an incomplete data matrix, and X as a complete one. Similarly, Y_{mis} and X_{mis} are the outcome and covariate rows, respectively, where the outcome is missing, and Y_{obs} and X_{obs} are the outcome and covariate rows where the outcome is observed.

Normal data

When the variable in question is distributed Normally, *i.e.* $Y \sim N\left(\mu, \sigma^2 I\right)$, then the OLS regression model is applicable, where $E\left[Y\right] = X\beta$. This method is very similar to that reviewed by Zhang (2003) in the predictive model (PM) method for monotone missingness. As noted in Subsection 3.2.2, a random draw from the posterior of the parameters and σ^2 is needed, and from there a random draw can be made from the posterior predictive distribution of the variable.

The parameter estimates from OLS are known to be $\hat{\beta} = \left(X'X\right)^{-1} X'Y$. In order to generate a random draw from the posterior of σ^2 we note that:

$$U = \frac{\left(Y - X\hat{\beta}\right)' \left(Y - X\hat{\beta}\right)}{\sigma^2} \sim \chi^2_{n-k} \tag{3.6}$$

where n is the number of observations in the regression and k is the number of parameters. Generating a random draw, u, from the χ^2_{n-k} distribution, and using the parameter estimates, $\hat{\beta}$, one can generate an estimate for σ^2, namely, σ^2_*, using the following equation:

$$\sigma^2_* = \frac{\left(Y - X\hat{\beta}\right)' \left(Y - X\hat{\beta}\right)}{u} \tag{3.7}$$

Using this estimate one can draw a set of parameters, β^*, from the posterior distribution of the parameters, using:

$$\beta^* = \hat{\beta} + \sigma_* T z_1, \tag{3.8}$$

where T is the symmetric square root of $\left(X'X\right)^{-1}$, the covariance matrix of $\hat{\beta}$, and z_1 is a random draw from the Standard Normal distribution.

Using β^* and σ^2_*, one can impute missing values using the following equation:

$$Y^*_{mis} = X_{mis}\beta^* + \sigma_* z_2, \tag{3.9}$$

where z_2 is another random draw from the Standard Normal distribution.

This procedure is identical to the procedure incorporated into the Normal version of the

PM method for monotone data, discussed in Subsection 2.3.2.

Binary data

When the variable in question is binary, one should implement a special case of the Binomial model, in which $Y \sim Bin\left(n, \pi\right)$. With dichotomous data the elements of n are ones. Parameters can be estimated from the general logistic regression model, with the logit link function, namely:

$$logit\left(\pi\right) = \ln\left(\frac{\pi}{1-\pi}\right) = X\beta \tag{3.10}$$

Maximum likelihood estimates of the parameters β, and therefore also of the vector of probabilities $\pi = \frac{\exp(X_{mis}\beta)}{1+\exp(X_{mis}\beta)}$, are obtained by maximizing the following log-likelihood function:

$$l\left(\pi; Y\right) = \sum_{i=1}^{n}\left[Y_i \ln \pi_i + \left(1 - Y_i\right)\ln\left(1 - \pi_i\right)\right] \tag{3.11}$$

From Equation (3.10) we have that

$$\pi_i = \frac{\exp\left(X_i\beta\right)}{1 + \exp\left(X_i\beta\right)} \tag{3.12}$$

and therefore

$$\ln\left(\pi_i\right) = X_i\beta - \ln\left[1 + \exp\left(X_i\beta\right)\right] \tag{3.13}$$

and

$$\ln\left(1 - \pi_i\right) = -\ln\left[1 + \exp\left(X_i\beta\right)\right] \tag{3.14}$$

Using these results in Equation (3.11) yields the following log-likelihood function to be maximised:

$$\begin{aligned} l\left(\pi; Y\right) &= \sum_{i=1}^{n}\left\{Y_i\left[X_i\beta - \ln\left[1 + \exp\left(X_i\beta\right)\right]\right] - \left(1 - Y_i\right)\ln\left[1 + \exp\left(X_i\beta\right)\right]\right\} \\ &= \sum_{i=1}^{n}\left\{Y_iX_i\beta - \ln\left[1 + \exp\left(X_i\beta\right)\right]\right\} \end{aligned} \tag{3.15}$$

For maximum likelihood estimation, the scores with respect to the $(p+1)$ elements of β are required, $U_0, U_1, \ldots, U_p$, or in other words, the derivatives of the log-likelihood function with respect to the elements of β, as well as the information matrix, F. The estimates are then obtained by solving the iterative equation $F^{(m-1)}\hat{\beta}^{(m)} = F^{(m-1)}\hat{\beta}^{(m-1)} + U^{(m-1)}$, where the superscripts denote the number of the iteration. The initial settings for the elements of $\hat{\beta}$ are zeros. Estimates are taken once convergence has been achieved, and at that stage the covariance matrix is taken as the inverse of the information matrix. For more details on the process, see Dobson (2002).

To impute missing values from this distribution, a random draw, β^*, is drawn from the posterior of the parameters as before in Equation (3.8), although this time a different MLE estimate $\hat{\beta}$ is used. Then a vector of probabilities is generated:

$$\pi_* = \frac{\exp\left(X_{mis}\beta^*\right)}{1+\exp\left(X_{mis}\beta^*\right)} \tag{3.16}$$

Finally, a vector of Uniform random variables is generated that has the same length as π_*, and this vector is compared with π_*. If an element of the vector of Uniforms is less than or equal to the corresponding element of π_* then a '1' is imputed for the missing value associated with that element of π_*. Alternatively, if an element of the vector of Uniforms is greater than the corresponding element of π_* then a '0' is imputed for the missing value associated with that element of π_*. This process details approximate draws from the posterior predictive distribution of the missing values (Raghunathan et al. 2001).

Count data

For count data, where $Y \sim Pois\left(\lambda\right)$, the Poisson regression model is used. The mean of Y is λ, and is modelled as follows:

$$\lambda = \exp\left(X\beta\right) \tag{3.17}$$

The linear predictor is $g\left(\lambda\right) = X\beta$, where $g\left(.\right)$ is the log link function.

Once more a random draw, β^*, is taken from the posterior of the parameters of the regression model, as before in Equation (3.8). A parameter set, λ^*_{mis}, is then generated as follows:

$$\lambda^*_{mis} = \exp\left(X_{mis}\beta^*\right) \tag{3.18}$$

A missing datum is then imputed by drawing a random number from a Poisson distribution with the element of λ^*_{mis} corresponding to that missing datum as the distribution's parameter.

Categorical and ordinal data

For data Y that can take one of k values, $j = 1, 2, \ldots, k$, let $\pi_j = \Pr(Y = j|X)$. A polytomous regression model is fitted, relating Y to X as follows:

$$\log\left(\frac{\pi_j}{\pi_k}\right) = X\beta_j,\ j = 1, 2, \ldots, k-1. \tag{3.19}$$

With the restriction that $\sum_{j=1}^{k}\pi_j = 1$, then $\pi_k = [1 + \sum_{j=1}^{k-1}\exp(X\beta_j)]^{-1}$. Let $\hat{\beta}$ be the MLE estimate for a polytomous regression with regression coefficients $(\beta_1, \beta_2, \ldots, \beta_{k-1})$, and let V be the asymptotic covariance matrix with Cholesky decomposition T.

Again a random draw, β^*, is taken from the posterior of the parameters of the regression model, as before in Equation (3.8). Now let

$$P^*_i = \frac{\exp\left(X_{mis}\beta^*_i\right)}{1 + \sum_{i=1}^{k-1}\exp\left(X_{mis}\beta^*_i\right)}, \tag{3.20}$$

where β^*_i is the appropriate elements of β^*, $i = 1, 2, \ldots, k-1$, and $P^*_k = 1 - \sum_{i=1}^{k-1}P^*_i$.

Then let $C_0 = 0$, $C_j = \sum_{i=1}^{j}P^*_i$ and $C_k = 1$, the cumulative sums of the probabilities. To impute values, generate a random Uniform number u and take j as the imputed category if $C_{j-1} \leq u \leq C_j$. As with count data, the imputations are from approximate predicitve posterior distributions, since the corresponding parameter draws are from asymptotic Normal approximate posterior distributions.

3.3.2 Other sequential procedures

It should be noted that several of the methods detailed in Subsection 2.3.1 can be incorporated into the SRMI procedure, as mentioned by Brand (1998). These methods include a hot-deck (HD) error term variant and a round-off option, both methods similar to methods introduced in He & Raghunathan (2009) and summarised in Subsection 3.4.4.

The HD error-term variant handles skew or heavy-tailed error-term distributions. In this variant, imputations for a missing data entry are made up of the sum of the predicted values of the outcome based on the posterior regression coefficients, plus an error drawn from a suitable subset of error terms. When this set comprises the q error terms of the q observations with predicted outcomes closest to the predicted value of the observation considered, the imputation will be robust against heteroscedastic error terms. A reasonable choice for q may be $0.3 * n_{obs}$, with n_{obs} the number of observed values for y (Brand 1998, p. 59).

Brand (1998) also briefly describes a method in which values outside the domain of y are not imputed, the round off option. In this case, any generated imputation is replaced by the nearest observed value of y.

If even GLMs or the other proposed model-based methods are not adequate for a variable in the incomplete data, a nearest neighbour classification can be used, as illustrated by Brand (1998, pp. 59–60):

> "With nearest neighbour imputation an imputation y_i^* for a missing entry y_i is generated by drawing y_i^* from an estimate $\hat{P}(y_i|X_i^T)$ of the predicitive distribution of y_i given the corresponding row X_i^T of X. In order to reflect the uncertainty about $P(y_i|X_i^T)$, the imputation y_i^* is generated from $\hat{P}(y_i|X_i^T)$ according to the Bayesian boostrap method. The estimate $\hat{P}(y_i|X_i^T)$ is the empirical distribution of the observed values $y_{i_1}, \ldots, y_{i_q}$ of y which are chosen such that the corresponding rows $X_{i_1}^T, \ldots, X_{i_q}^T$ are the $q = [f_{dc} * n_{obs}]$ rows of X_{obs} closest to X_i^T, with f_{dc} the donor class fraction and $[.]$ the entire function. In this context, 'close' is defined by a distance function $d = d(X_i^T, X_j^T)$, with X_i^T and X_j^T the i-th and j-th row of X. A reasonable value of f_{dc} may be 0.1"

Brand (1998, p. 98) further summarises this process into easy-to-follow steps, as given below:

1. Select from X_{obs} the q rows $X_{i_1}^T, \ldots, X_{i_q}^T$ which are closest to X_i^T as measured by a distance function d, such as the Euclidean, Mahalanobis or Jaccard distance (for continuous commensurate, continuous non-commensurate, and categorical data, respectively).

2. Draw $q-1$ Uniform(0,1) and let $a_1, \ldots, a_{q-1}$ be their ordered values. Also, let $a_0 = 0$ and $a_q = 1$

3. Let $p_j = a_j - a_{j-1}$ for $j = 1, \ldots, q$

4. Draw y_i^* from $y_{i_1}, \ldots, y_{i_q}$ with probabilities $p_1, \ldots, p_q$

Steps 2–4 indicate that imputations are drawn according to a Bayesian bootstrap method, while the probabilities p_j in Step 3 are drawn to reflect uncertainty about the predictive distribution given the observed values.

This model-free approach may warrant further investigation, especially in any study aiming, in part, to create a robust sequential procedure that would make SRMI more attractive amongst non-statisticians.

3.3.3 Imputing from generalised linear mixed models

Creation of the multiply imputed datasets can incorporate models using fixed and/or random effects. In other words, the sequential regressions used in the MI process can incorporate both fixed and random effects, to be more accurate and appropriate for certain types of data. However, the process of including random effects into generalised linear models is not simple when prediction is concerned. In order to predict missing values, random effects have to be estimated in each generalised linear mixed model for each group in the data. This is not explicitly possible from model estimation, but rather requires a form of Gibbs sampling to determine the separate random effects. This complicated approach is not considered in this thesis.

3.3.4 Additional considerations for SRMI imputation

Whatever model or process is chosen to predict a particular variable, it is important to recall several key aspects. The first of these is that the imputation model should be more general than the final analysis model (see Subsection 2.2.6), and secondly as much correlated information should be included within the imputation procedures as is possible, in order to possibly shift an MNAR model towards becoming MAR, and any variables possibly related to missingness should be present as well(see p. 62 Brand 1998).

It would seem that this may lead imputers to include a significant number of variables in each imputation step; in order to reduce this number, stepwise procedures can be followed. In this way, only the most important variables for imputation of incomplete variable will be kept. This procedure is followed by Rubin (2003*b*), a study which is discussed in detail later in Subsection 3.4.2.

3.4 Recent Research in SRMI

SRMI has only become popular over the last decade or so. This means that any advances in the SRMI field besides the actual creation of the FCS, MICE, and SRMI procedures can be thought of as being ‘recent’. For this reason, this section is substantial and includes discussions on the studies that have used SRMI, studies that have compared SRMI with other imputation procedures, studies that have attempted forming SRMI diagnostic measures and studies that have used the unique characteristics of SRMI to advance the MI field.

3.4.1 Studies using SRMI

Over the last decade SRMI has proved to be a particularly simple and effective method of multiply imputing for incomplete data sets. For this reason, many studies have used this MI method. Some of these studies will be mentioned in this subsection, in order to obtain a more comprehensive view of the usage of SRMI.

Raghunathan et al.'s 2001 study

Besides laying the groundwork for the common SRMI procedure, Raghunathan et al. (2001) also include two illustrative examples for SRMI as well as a simulation study, all of which should be noted here for very specific reasons.

In the first application, extremely complex incomplete data is imputed using SRMI in a study of the relationship between cigarette smoking and primary cardiac arrest. In particular, the variables measuring the number of years smoked were bounded by different limits in different cases, depending on the respondents' answers to questions such as, 'how long ago did you quit smoking', 'did you smoke in school', *etc.* Such intricate bounds on this incomplete variable make joint modelling of the entire dataset rather complex, and so the SRMI algorithm, which can easily impute values from truncated distributions, becomes an obvious choice.

In the second application, examining the effects of parental psychological disorders on several measures of child development, the SRMI method is shown to be at least as good as a fully Bayesian MI approach, given that the data is incomplete due to a MAR MDM.

In their simulation study, Raghunathan et al. (2001) create a dataset consisting of a Normal variable, U, a Gamma random variable, Y_1, with mean and variance dependent on the Normal variable and another Gamma random variable, Y_2 with mean and variance dependent on the other two variables. The data is made incomplete using a logistic regression approach, informing the approach used in Chapter 4 of this thesis. There were no missing values in U; the missing values in Y_1 depend on U through a logistic function $\text{logit}\left[\Pr(Y_1 \text{ is missing})\right] = 1.5 + U$; and missing values in Y_2 depend on U and Y_1 through the logistic function $\text{logit}\left[\Pr(Y_2 \text{ is missing})\right] = 1.5 - 0.5Y_1 - 0.5U$. This mechanism generated 22% missing data in Y_1 and 29% missing data in Y_2, whereas the complete-case analysis would use only 48% of the data. Their study shows that SRMI correctly provides wider, more conservative confidence intervals on regression parameters than complete-data analysis would do, showing that the SRMI procedure is correctly incorporating the additional uncertainty due to missingness into the overall inferences.

National Health Interview Survey of 2001

Schenker et al. (2006) use SRMI to impute missing income data in the National Health Interview Survey (NHIS). The setting for the imputations in this study is well-suited to SRMI — some of the variables are hierarchical in nature, with family-level and individual-level measures; some variables have structural dependencies, with values depending on other variables; some variables should be imputed within bounds; incomplete variables have different non-Normal distributional forms. Each of these complications could easily be incorporated into the sequential regression models, within the following steps (Schenker et al. 2006, p. 926):

1. Impute missing values of person-level covariates and employment status for adults.
2. Create family-level covariates.
3. Impute missing values of family income and family earnings, as well as any missing values of family-level covariates(due primarily to missing person-level covariates for children).
4. Impute the proportion of family earnings to be allocated to each employed adult with missing personal earnings, and calculate the resulting personal earnings.

The authors' procedure then followed the standard SRMI sequence with four update rounds after the initial round. As usual, in the initial round, step 1 did not include income and employment as covariates, while steps 2 to 5 did (once all the gaps in the data set had been filled in). However, in the update rounds employment status was not re-imputed, to avoid incompatibilities with imputed values of personal earnings. The entire process was completed five times ($m = 5$) for five imputed data sets. According to the authors, other variables were also imputed and created, but were not retained in the final public-use data. Note that Schenker et al. (2006) maintained the hypothesised structural dependencies and truncated imputations, as mentioned before, and they incorporated design variables as imputation covariates. Some variables were transformed via Box-Cox analyses to Normality during imputation and were transformed back afterwards (see Box & Cox 1964).

An interesting part of Schenker et al.'s (2006) study is their adherence to Meng's (1994) recommendation that more general imputation models are required when imputation and analysis tasks are separated (see Subsection 2.2.5 of this thesis). Accordingly, Schenker et al. (2006) use about 60 predictors in the SRMI procedure, including variables related to sample design.

The results obtained are as expected; in the context of assessing MI, the MI standard errors for poverty ratios are lower than those from complete-case analysis (if both procedures produce unbiased estimates), but more than those of single imputation (which, of course, are not incorporating the true uncertainties within the imputation procedure).

Finally, the authors compare the MI results with results from poststratification reweighting, the latter producing results similar to those from complete-case analysis. It is clear that MI uses more additional information than the poststratification adjustment does.

Income poverty and inequality in South Africa from 1996-2001

Ardington et al. (2006) use SRMI to test the sensitivity of poverty and inequality measures in South Africa to the imputations of certain covariates. The authors summarise a vast literature (based on studies of incomplete data) that show how poverty and inequality within racial groups both increased in South Africa over the period between 1996 and 2001. Their use of SRMI to validate these findings shows that the previous literature was indeed correct, although they provide better confidence intervals via the MI procedure than were offered by the default hot-deck single imputation confidence intervals reported with the public release of the data. Ardington et al. (2006) find small increases in poverty for the poorest of the poor, and increases in inequality across the board.

Moreover, the authors analyse the effect of a high proportion of household incomes totalling zero (around 25%). They adjust their imputation procedure to take account of this drastic proportion, rather than ignoring the zeros or arbitrarily assigning small amounts to these households, both practices having previously been thought of as being acceptable. Besides this analysis, the authors check the sensitivity of the results to the assumptions regarding point estimates for income that were generated from the income bands recorded in the surveys. The initial setting is to take the midpoints of bands, the point estimate of zero (which was included in the survey question response options), and the lower bound of the highest unbounded band.

The authors followed the imputation processes and recommendations made by Raghunathan et al. (2001) and estimated mean household income, a poverty head count index, and the Gini coefficient inequality measure from the resulting multiply imputed data sets. The results were combined using the regular combining rules presented in Subsection 2.2.3. The authors use province of residence, urban/rural location, and race as complete predictors, while the incomplete variables, ordered from least missing to most missing include age (a count variable), gender, emploment status (unemployed vs employed), occupation (four categories), years of education (a count variable), and income (an ordered categorical variable of 12 income bands).

Ardington et al. (2006) report on the sensitivity analysis for the 2001 data, and then report the 1996 to 2001 interval's estimates changes based on their best methods. They find that ignoring the missing values downwardly biased estimates of mean income and inequality, and upwardly biased estimates of poverty, results that were expected or are consistent with preliminary analyses. The confidence intervals generated are far wider than those generated from the single hot-deck imputation procedure.

Within a survey year, to test the sensitivity to the large proportion of households having zero income, the authors recode the zeros as missing and allow the imputation procedure to validate those original zeros. These households can then be assigned either positive or zero incomes, depending on the imputation. The authors admit that this method will bias post-imputation estimates (due to the fact that the true zeros are also recoded as missing), but they argue that the procedure has merit as a sensitivity analysis. Systematic anomalous entry types are also recoded to missing. The results are then thought of as being a boundary limit. With no imputations, hot-deck imputations, and SRMI, the proportions of households with no income, the proportions of employees earning no income, and the proportion of children earning anomalous incomes, are all similar in magnitude, at around 24%, 2% and 0.14% respectively. After the recodings, however, these magnitudes drop to 13.48%, 0.47%, and 0.01%. These percentages should be considered as lower bounds, and we should expect, for example, the true number of households with zero income to be between 13% and 23% (the latter being the SRMI value before the recodings). Recoding the implausible values to missing also increases mean per capita income and decreases the

Gini coefficient and the percentage of poor households, although these changes are not substantially different from the SRMI results before recoding.

The final sensitivity test is the analysis of the estimates based on a generated income measure. This measure looks to spread all the observations' incomes over intra-band distributions. The intra-band distributions used are based on empirical evidence from another survey, and observations with income measures within a certain band are randomly spread within that band according to the distribution of incomes within the same limits from the alternative survey. The results indicate a falling mean income per capita, and rising poverty and inequality, although all of these changes are not extreme, which indicates a lack of sensitivity to the type of income measure: band midpoints or a continuous distribution. Since the distributions within bands still follow empirical evidence, however, the authors prefer this continuous measure to the band-midpoint measure.

General Social Survey of 1998

Penn (2007) applies SRMI to 1 747 observations (of individuals over the age of 25) from the U.S. General Social Survey (GSS) of 1998. In this data set, 11.7% of records are missing the income measure, primarily due to individuals refusing to answer that specific survey question. By looking at the distributions of the income-complete and incomplete observations, Penn (2007) shows that complete-case analysis would definitely produce biased results. The purpose of the study is to verify results from the study by McBride (2001), which, after using only complete-case analyses, showed that a person's self-reported happiness is reliant on the person's standard of living compared with that of their parents. Penn (2007) use the following variables in the imputation: happiness, parents' standard of living, educational level, age, marital status, gender, health status, race, family income, working status, occupation (9 categories), and educational level. All but the last three variables are used for the completed-data analysis, so the imputation model is more general than the analysis model. Of the 1 747 observations, income is missing for 205, parents' standard of living has 26 missing values, happiness has 21 missing values, and fewer for health status and age. At least one item is missing for 13.9% of the sample.

Penn (2007) sets $m = 6$ in his analysis, and performs the ordered probit regression of happiness on both the completed-data and the incomplete data, for comparison. Penn (2007) finds that not only do the coefficients of the relative standard of living between parents and children grow (meaning that if they have a better standard of living than their parents did at their age, they tend to be happier), but the standard errors for this categorical variable decrease after imputation, moving the two statistically insignificant categories into significance. The fact that these standard errors decrease may be due to the increase of the analysed sample size, from 1 503 to 1 747, or, more likely, because the imputation model is imputing correct information and superefficiency[2] is occurring (Meng 1994, Rubin 2003*a*).

3.4.2 Nested multiple imputation

Another recent MI technique is that developed by Shen (2000) and used by Rubin (2003*b*), namely nested MI. The two novel ideas incorporated into this MI procedure are distributionally incompatible MCMC (another name for SRMI), and nested MI to enhance efficiency at a fixed cost. These procedures were applied to the National Medical Expenditure Survey (NMES). In essence, the MCMC procedure is a form of the sequential regression MI procedure. Rubin calls the MCMC procedure partially incompatible because a joint distribution of all the data might not exist, even though the MCMC method works as if there was one. This problem is discussed further in Subsection 3.4.7.

Essentially, in nested MI, the imputations are split into two parts, the computationally expensive, and the computationally inexpensive. For the first part of the imputation, m_1 multiply imputed data sets are created, and for each of these data sets m_2 imputed data sets are created for the second part of the imputation. The regular combining rules are then used to combine estimates from the $m_1 m_2$ imputed data sets.

As before, with the complete data we have that $(Q-\hat{Q}) \sim N(0,U)$. The $m_1 m_2$ completed data sets are used to calculate the following values of the statistics $\hat{Q}$ and U: $(\hat{Q}_{i,j}, U_{i,j}), i = 1,\ldots,m_1; j = 1,\ldots,m_2$.

[2] As discussed in Subsection 2.2.5)

Then we have that

$$\bar{Q}_{\text{NEST}} = \frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \hat{Q}_{i,j}, \tag{3.21}$$

and that

$$\bar{Q}_{i,\text{NEST}} = \frac{1}{m_2} \sum_{j=1}^{m_2} \hat{Q}_{i,j}, \tag{3.22}$$

or, $\bar{Q}_{i,\text{NEST}}$ is the average $\hat{Q}$ in the ith nest. Also, as before, let

$$\bar{U}_{\text{NEST}} = \frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} U_{i,j}, \tag{3.23}$$

but, let MSW_{NEST} be the within-nest mean square, and let MSB_{NEST} be the between-nest mean square, where

$$MSW_{\text{NEST}} = \frac{1}{m_1 (m_2 - 1)} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \left(\hat{Q}_{i,j} - \bar{Q}_{i,\text{NEST}} \right)^2, \tag{3.24}$$

and

$$MSB_{\text{NEST}} = \frac{m_2}{(m_1 - 1)} \sum_{i=1}^{m_1} \left(\bar{Q}_{i,\text{NEST}} - \bar{Q}_{\text{NEST}} \right)^2. \tag{3.25}$$

With these quantities, the total variance of $(Q - \bar{Q}_{\text{NEST}})$, T_{NEST} is estimated as follows:

$$\begin{aligned} T_{\text{NEST}} &= \bar{U}_{\text{NEST}} + \frac{1}{m_2}\left(1 + \frac{1}{m_1}\right) MSB_{\text{NEST}} \\ &\quad + \left(1 - \frac{1}{m_2}\right) MSW_{\text{NEST}}. \end{aligned} \tag{3.26}$$

When $m_2 = 1$, this last expression reduces to the familiar Equation (2.6), *i.e.* $T = \bar{U} + (1 + 1/m)B$.

Interval estimates and significance levels for a scalar Q are based on a Student-t reference distribution, as before, with

$$\frac{\left(\hat{Q} - Q\right)}{\sqrt{T}} \sim t_w, \tag{3.27}$$

where the degrees of freedom, w, are calculated as follows:

$$\begin{aligned} w \;=\; & \left\{ \left[\frac{\frac{1}{m_2}\left(1 + \frac{1}{m_1}\right) MSB_{\text{NEST}}}{T_{\text{NEST}}} \right]^2 \frac{1}{m_1 - 1} \right. \\ & \left. + \left[\frac{\left(1 - \frac{1}{m_2}\right) MSW_{\text{NEST}}}{T_{\text{NEST}}} \right]^2 \frac{1}{m_1\left(m_2 - 1\right)} \right\}^{-1} \end{aligned} \tag{3.28}$$

Once again, when $m_2 = 1$, Equation (3.28) reduces to the familiar expression for the degrees of freedom, Equation (2.10), namely,

$$w = v = (m-1)\left[1 + \left(\frac{m}{m+1}\right)\frac{\bar{U}}{B}\right]^2,$$

One of the interesting aspects of this study is the use of slightly different regression models for the conditional modelling, in particular, the use of stepwise regression and a sampling importance resampling (SIR) algorithm.

- Missing binary variables are predicted using a stepwise logistic regression, with parameters drawn from their asymptotic Normal Bayesian approximation improved by SIR.

- Missing categorical variables are predicted in the same way as the binary variables, but sequentially, first predicting the most populated category versus the rest, *etc.*, using the appropriate group of subjects.

- OLS-round-to-observed variables are predicted using a stepwise Normal OLS regression with parameters drawn from the Bayesian posterior distribution, and are then rounded to the closest observed value in the data set.

- Semi-continuous variables (with a positive probability of being zero, but if they are positive then they are continuous) are imputed in two stages: the logistic for “0” vs

"+", and then, if "+", an OLS-round-to-observed routine is followed.

- Missing or previously-imputed medical expenditure data, in the form of two 18x5 tables for each respondent (number of events by code and category within code and expenditure by code and category within code), is imputed using a method similar to the predictive mean matching method discussed in Subsection 3.4.5. The number of events with both known expenditures and missing expenditures is known. For those with missing expenditures, these missing values were originally imputed in the publicised version of the data. These imputed values are used to create a donor pool by looking for observed values of that same cell across a gender group within a certain distance from that imputed value. The imputed values are not simply drawn at random from this pool, but through probabilities attached to each value in each person's donor pool.

Before multiply imputing through the MCMC procedure, Rubin (2003*b*) sorts the rows and columns in order to approximate a monotone pattern of missing data, which can be dealt with non-iteratively, using combinations of standard regressions. Unlike in MCMC, when the pattern of missing data is monotone, the univariate models are automatically compatible. The imputation procedure then follows the steps given below:

1. In order to make this approximate monotone pattern exactly monotone, the missing data (from the left-most variable) that destroy the monotone pattern are imputed using one of the listed methods above. Then the next variable to the right is 'filled in' up to a point where the pattern is monotone again, and then the next variable, *etc.*, until one dataset is obtained with a monotone pattern of missing data. When the number of regressors becomes substantial, stepwise regression procedures are used. Also, when parameter draws are made from posteriors where asymptotic approximations are made, SIR is used to improve the draws.

2. The non-monotone missing values are imputed using one repetition of the SRMI procedure as we know it, based on the models listed above. In other words, the monotone-patterned incomplete data set is filled in using the first round of the procedure explained in Subsection 3.2.2, using the models listed above.

3. The next round of imputations begins by re-imputing the missing values that orig-

inally destroyed the non-monotone pattern. However, in this round the predictors include all imputed values from the previous round (as if they were observed). After this round, the result is another monotone-patterned data set.

4. Use step 2 to fill in the non-monotone part again. This is the iterative step. The process switches between steps 2 and 3 until the whole process is iterated five times, which Rubin (2003*b*) mentions took a substantial amount of time (2 days per iteration) in their research process. The small number of iterations, however, is less concerning than it could be, since the missing data is made to be approximately monotone (if it was perfectly monotone to start with, convergence would have been immediate). The end result of these four steps is a single completed data set.

Moving to the imputation of the health expenditure variable, one row of the health expenditure data indicated entries with missing health event codes. The expenditures in this row have to be reallocated to one of the other codes. These reallocation codes are multiply imputed within each completed data set arising from the four steps given above. First the number of events with missing codes are allocated across the other codes, and then the expenditures on these events with missing codes are reallocated.

Firstly, a prior distribution is created for each individual based on all other similar individuals. Then a procedure known as Iterative Proportional Fitting (see Bishop, Feinberg & Holland 1975) is used to create a table with margins of the individual's table, but interactions formed from the similar people (in terms of covariates or coarse categories of covariates). This is formed into a table of proportions that sums to one in each column across the 17 real rows. Then twice the values in this table are added to the actual table for that person. So each person has two prior observations added to each column. Then a Dirichlet draw is taken from the table, column by column, and the missing events are allocated with these probabilities.

Secondly, dollar amounts for the missing code events are drawn from a truncated posterior Normal distribution with mean and variance of the cell to which the event is allocated, but truncated to lie between zero and the maximum observed value for that type of event.[3]

[3] In other words the population cell mean is drawn, centered at the cell sample mean with variance equal to the within-cell variance divided by the number of events in that cell on which that sample mean is based and, for each event allocated to that cell, a value centered at the drawn population mean is drawn with variance equal to the within-cell variance.

Once all the dollar amounts are allocated in this way, they are renormed to add up to the original total for the missing codes for that individual. The table values are then totalled up with the imputed numbers of events and expenses for each of the reallocated events for each individual.

Finally, the nesting comes into play. Since the imputations for the non-health expenditure variables are computationally expensive to run when compared with the imputations for the health expenditures, Rubin (2003*b*) creates ten imputed health expenditure parts of the data set for each imputed data set created from the previous four listed steps.

3.4.3 Synthetic data

One of the more recent uses of SRMI has been in disclosure limitation in public use data sets, spearheaded in part by Reiter (see, for example, Reiter 2005, Reiter & Raghunathan 2007, Reiter 2009, Reiter 2012, Wang & Reiter 2012, Paiva, Chakraborty, Reiter & Gelfand 2014). Disclosure limitation aims to ensure the respondents of surveys remain anonymous once the data set that they are a part of is published. Even stripping out identifying information (such as age, race, marital status, gender, *etc.*) might not be enough to limit disclosure of respondents (Reiter & Raghunathan 2007). Thus, data disseminators sometimes revert to methods that alter the observed value of the data. However, the more alteration made to the data, the lower the accuracy of inferences made from the data, even though the additional alterations help to anonymise the data (Reiter & Raghunathan 2007). However, using MI, and SRMI in particular (since survey data variables are difficult to model jointly), one is able to multiply impute some or all of one or more variables of a public use dataset in order to both limit disclosure and keep resultant inferences accurate.

The science behind disclosure limitation using multiple imputation boils down to two types of synthesised data: partially synthetic data and fully synthetic data. Partially synthetic data is a collection of data sets in which some variables for the surveyed units have been fully or partially multiply imputed, whereas fully synthetic data consists of units that have been entirely multiply imputed. To create partially synthetic data, sensitive information within the current sample is multiply imputed and the actual values are replaced by these multiple imputations. In order to create a fully synthetic sample, the missing population units are multiply imputed (and possibly even the original sampled units as well), and from this completed population synthetic samples are drawn.[4] The process is repeated to create multiple synthetic samples. Different combining rules have been constructed for both cases. For more information on these rules (and further references), see Reiter & Raghunathan (2007, p. 1476).

The topic of multiple imputation for disclosure limitation is a large area of research. Thus, it would not be wise to delve into the this field in this thesis. However, it is worth noting that the novel area of research within this thesis can easily be applied within SRMI for disclosure limitation; a research area that could be expanded on in the future.

3.4.4 Evaluation of SRMI

Since SRMI is a relatively recent development in the MI field, there have been several studies recently that compare this method with other imputation procedures. These studies are reviewed in this subsection.

SRMI vs available case

Van Buuren et al. (2006) evaluate SRMI in three simulation studies by looking at univariate and multivariate missingness and three types of models. The univariate studies use Irish wind speed data (for linear and logistic imputation methods), and data from women on knowledge of, and attitude and behaviour towards mammography, *i.e.* mammographic experience (for a polytomous imputation method).

For the univariate analysis the outcome variable is replaced in each study by predicted values of the outcome given the other variables, in a sample taken from the original data. This is done 1 000 times for each data set. Then 50% missingness is induced in these simulated outcomes, according to an MCAR mechanism, and MAR mechanisms creating more missing data in larger values (MARRIGHT), more missing in tail values (MARTAIL) and more missing in the centre of the distribution (MARMID).

[4] Only the synthetically sampled, unobserved units actually need to be multiply imputed, not the entire unobserved part of the population (Reiter & Raghunathan 2007).

In the multivariate missingness study on continuous data, the wind speed data is used and two samples of 400 observations are taken; one to approximate the mean and covariance of the original data (the simulated set) and one random sample. Missing values were then created in the data according to a specific non-monotone structure, as generated by Brand (1998, p. 110–113), and summarised in van Buuren, Brand, Groothuis-Oudshoorn & Rubin (2006, Appendix B). This process will be detailed here in order to be replicated later in this study.

Assume a sample size of n, and that $Y_1, \ldots, Y_J$ are initially completely known. Additional complete covariates $X_1, \ldots, X_L$ can be present for which no missing entries are sought. The following must be specified:

- The proportion of incomplete cases $(0 < \alpha < 1)$.
- The allowed patterns of missing data $(R_1, \ldots, R_P$, where $R_P = \{r_{p1}, \ldots, R_{pJ}\}$ is a 0-1 response indicator of length J, with $r_{pj} = 0$ if variable Y_j is missing and $r_{pj} = 1$ otherwise). All response patterns except fully observed and fully incomplete can occur.
- The relative frequency of each pattern, *i.e.* $f = (f_1, \ldots, f_P)$, the relative frequencies for patterns $R_1, \ldots, R_P$, with $\sum_p f_p = 1$.
- The way in which the observed information can influence the response probability of each pattern.

So each case is randomly assigned to one of P candidate blocks with probability f_p. Within each block, a subgroup of $\alpha n f_p$ cases is made incomplete according to a pattern R_p using a probability model specified as follows:

1. Calculate a linear score $s_i = \sum_{j=1}^{J} a_{pj} r_{pj} Y_{ij} + \sum_{l=1}^{L} b_{pl} X_{il}$ for each case in the block, where a_{pj} and b_{pl} are user weights specific to pattern p (for example, regression weights from regressing Y_j on $\{Y_{-j}, X_1, \ldots, X_L\}$ as computed from the initially complete data).
2. Divide the nf_p cases within the candidate block p into k_p subgroups using their value s_i. The user can control the composition of each candidate subgroup by specifying $k_p - 1$ break points q_{pk} for $k = 1, \ldots, k_p - 1$ in the form of quantiles.

3. Specify for each subgroup $h_k(2 < k = k_p)$ the odds w_{pk} of having response pattern R_P relative to that of the reference subgroup h_1. Together with α, these odds determine the probability on response pattern R_p for each case in the candidate block.

4. For each case, draw randomly from a Uniform distribution, and if this draw does not exceed the probability on response pattern R_p the data for that case are set missing according to response pattern R_p.

5. This procedure is then repeated for each candidate block.

For their simulation, van Buuren et al. (2006) generate missingness in their data set $\{Y_1, \ldots, Y_4, X_1, X_2\}$ using the above procedure, with $P = 4, R = \{010111, 001111, 110011, 101011\}, f_1 = f_2 = f_3 = f_4 = 0.25, \alpha = 0.625, k_1 = k_2 = k_3 = k_4 = 2, q_{pl} = 0.5$, and $w_{pk} = 4$ with $p = 1, \ldots, P$ and $k = 1, \ldots, k_p - 1$.

The authors find that the linear regression SRMI procedure restores correlations and eliminates biases in the data set, correlation losses and biases that appear in the available case analyses.

For their third and final simulation study, in each of 500 replication, 1000 draws are made from a bivariate Normal distribution with means equal to 5, variance equal to 1, and correlation equal to 0.6. All values are positive. Missing values are generated in one of three ways:

- MARRIGHT: $\text{logit}(\Pr(Y_1 = \text{missing})) = -1 + Y_2/5$, while $\text{logit}(\Pr(Y_2 = \text{missing})) = -1 + Y_1/5$.
- MARTAIL: $\text{logit}(\Pr(Y_1 = \text{missing})) = -1 + 0.4|Y_2|$, while $\text{logit}(\Pr(Y_2 = \text{missing})) = -1 + 0.4|Y_1|$.
- MARMID: $1 - \Pr(\text{MARTAIL})$.

For MARRIGHT there are about 50% missing entries, 75% incomplete cases, and about 25% completely missing (Y_1, Y_2) pairs, with proportionally more missing data for the higher values of Y_1 and Y_2. The average missing information generated is around 0.63, which is rather extreme. Van Buuren et al. (2006, p. 1059) mention that, “The multivariate missing data were not entirely MAR because the cases where Y_1 and Y_2 (or both) is (are) missing were more frequent for the higher values. The regression lines are, however, not affected because the nonresponse is generated symmetrically around the regression lines." Compatibility was ensured in the Gibbs sampler for the bivariate draws by chaining the multiple imputations Y_1^* and Y_2^* from the conditional models $Y_2^*|Y_1 \sim N(\mu_1^* + \beta_1^* Y_1, \sigma_1^{2*})$ and $Y_1^*|Y_2 \sim N(\mu_2^* + \beta_2^* Y_2, \sigma_2^{2*})$, where $\mu_1^*, \beta_1^*, \sigma_1^{2*}, \mu_2^*, \beta_2^*, \sigma_2^{2*}$ are draws from the appropriate posterior distributions. Incompatibility was generated by replacing the imputation step for Y_2 by $Y_2^*|Y_1 \sim N(\mu_1^* + \beta_1^* Y_1^2, \sigma_1^{2*})$, and, separately, $Y_2^*|Y_1 \sim N(\mu_1^* + \beta_1^* \log(Y_1), \sigma_1^{2*})$. The authors generate $m = 5$ completed data sets for each model, while in each data set, the Gibbs sampler is only iterated 5 times as well. The complete data model is the linear model $Y_1 = \alpha + \beta Y_2 + \varepsilon$, with analysis interest focussed on β.

The authors find that for the incompatible models, serious bias and undercoverage of the true estimate is eliminated using SRMI, meaning that incompatibility is a relatively minor issue in their SRMI applications; *i.e.* the SRMI procedure is robust against incompatibility.

SRMI vs poststratification reweighting

Note that the study by Schenker et al. (2006), reviewed in Subsection 3.4.1, showed that postratification reweighting produced results similar to those of the complete-case analyses, which implies that postratification reweighting uses less information than SRMI, and thus is an inferior method in the case of their analyses.

SRMI vs MN

Several studies compare SRMI and the regular MN method. These will be summarised briefly below.

Faris et al. (2002) compare MN, SRMI, CSR, and data enhancement (through merging of additional data) in a study of 6 065 cardiac care patients in Alberta, Canada, in 1995.

The outcome of interest is the binary variable measuring one-year mortality. Categorical variables in the predictor set are split into binary indicators. The original data set consists of 6 276 individuals, but is reduced when 6 026 individuals are matched with hospital discharge administrative data in the 'enhanced' data sets — the combined clinical and administrative data sets. Several of the variables overlap in these two data sets. For the two categorical variables in the study that did not overlap (left ejection fraction and Duke Index of coronary artery disease severity), missing values were recoded into an additional observed category (but only in the enhanced data set). When they compared the individuals with and without administrative data, the authors note that one may assume that the individuals with administrative data are simply a random sample of the original individuals. Sources discussing the merits and drawbacks of this method are given by Faris et al. (2002, p. 186).

The authors then compare logistic regression results from mortality regressed on the clinical data after MN MI on the original data, SRMI on the original data, complete-case analysis of the enhanced data set, as well as enhancement and SRMI combined, and finally, enhancement and SRMI with the administrative variables included in the imputation model. The imputations from the SRMI procedure are taken after only five rounds in this study, as the Gibbs sampler seems to converge at that point. Ten imputed data sets are used in MI analyses.

In order to assess the procedures, *firstly* the logistic model fit was assessed using (1) the C statistic, which is the area under the receiver operating characteristic or ROC curves[5] with bootstrapped confidence intervals, and (2) the changes in deviance from the null model, *i.e.* $-2 \times \log L$ of the model versus $-2 \times \log L$ of the null. *Secondly*, the ability of the coefficients to predict the outcomes for the complete cases in a 1996 follow up survey was assessed.

The authors find the following: for the 1995 data methods, the C statistic is greatest for SRMI, followed by CSR, MN, and lastly, enhancement. Validating using the 1996 data shows the greatest C statistic for SRMI again, followed by CSR and enhancement, followed by MN. In both cases, the SRMI C statistic ranks first in more than 90% of the bootstrapped estimates. The SRMI model also has the largest decrease in deviance from the null model for 1995, followed by CSR, then enhancement, then MN; SRMI is best in

[5] ROC curves will have a maximum value of one if when those dying all have larger fitted values than those surviving.

more than 99% of the bootstrapped values. Validating using the 1996 data shows SRMI above CSR, followed by MN, followed by enhancement; SRMI is again first in 98% of the bootstrapped estimates. The enhanced SRMI combination performs better than the original enhanced fit, while the administrative SRMI model shows no improvement over the original SRMI model. All of these results show the superiority of the SRMI method in this case, although the performance of all methods is relatively satisfactory.

Van Buuren (2007) compares joint modelling through the MN specification to an SRMI procedure on the Fourth Dutch Growth Study data set of measures of pubertal development in 3801 Dutch girls (these were the observations with complete age, height and weight measures). About 34% of the so called Tanner stage development data is missing. The data includes menarche stage (two categories), breast development (five categories) and pubic hair (six categories).

The imputations made for these incomplete variables under the MN scheme are rounded to fit into the categorical nature of the data. The two-category menarche variable is imputed using a logistic regression model, while the two incomplete categorical variables, breast development and pubic hair, are imputed using polytomous logistic regression, which in itself can raise modelling issues (see van Buuren 2007, p. 233). For each the MI procedures, five imputed data sets were created.

The author uses correspondence analysis on the two categorical variables to determine whether the inherent structure of the data is preserved between complete-case analysis and MI via the MN and SRMI methods, and finds that the SRMI based correspondence analysis preserves the canonical correlations of the complete-case analysis better than the MN method does.

Van Buuren (2007) then regresses log weight on the incomplete and completed data, and finds all three procedures (complete-case analysis, MN, and SRMI) produce the same significant model fit. Standard errors from the imputed data are much narrower due to the increased sample size (from 2200 in the complete-case analysis to 3 801 in the completed data sets). Suspicious of these similar results, van Buuren (2007) creates reference curves for the complete-case method versus the imputation methods. For each stage transition of breast development, a reference curve was fitted conditional on age by a series of four logistic additive models. From these curves it is clear that the MN method imputes data for breast development that does not fit the complete-case distribution across age, while the SRMI approach does succeed in doing so. For this reason, van Buuren (2007) recommends that the MN approach is not chosen above the SRMI approach when the incomplete variables are categorical in nature. It is possible that the rounding of the MN imputations adds to biases seen in the results.

Lee & Carlin (2010) compare the SRMI method with the standard MN method in their study of estimation of regression coefficients from simulated data after MI. Their analysis is similar to that of van Buuren (2007), except the rounding of the MN method's imputations is adapted to the true distribution of the categorical variable. The simulated data sets are obtained by sampling from a synthetic population of 971 327 girls, grades 7–10, created to resemble the sample from the US National Longitudinal Study of Adolescent Health.

The variables are synthesised sequentially, starting by drawing 1 million observations from the 3 × 5 race–grade table, and then adding one variable at a time using predictive simulation from regression models based on the original data. At each step, the model conditions on the previously generated values, incorporating them into complex regressions that included nonlinear relations and numerous interactions, to create sufficient population complexity. Since the outcome variable, emotional distress at wave II, is a continuous measure between 0 and 3 that is strongly positively skewed, the 0 score observations are dropped so as to not complicate a logarithmic transform ($ldistW2$). Data sets for the study each draw 1 000 individuals from this synthetic population. The analysis of the regression of the log of distress on other covariates (diet, log of distress at wave I, Black race indicator, Hispanic race indicator, grade, health and physical fitness) is primarily concerned with the main coefficient, that of diet. The so-called true values of the coefficients are obtained from the same OLS model applied to the full synthetic population. The authors use the original non-significant diet effect and an artificially inflated significant diet effect as comparisons (but both produce a similar set of results, given below).

The data are set missing according to one of three models set out below, each model using a logistic regression of the following form (where $ldistW2$ is the outcome variable):

$$\operatorname{logit} \Pr(\text{missing}) = \alpha + \beta_1 diet + \beta_2 Black + \beta_3 Hisp + \beta_4 grade + \beta_5 ldistW2$$

1. Missing data on emotional distress at wave I

2. Model 1, plus independent missing data pairs on health and physical fitness

3. Model 2, plus independent missing data on diet.

For Models 1 and 2, the coefficients are fixed to create a substantial association between variables and missingness, as follows: $\alpha = 3, \beta_1 = \beta_2 = \beta_3 = 1, \beta_4 = 0.2, \beta_5 = 0.3$. The MDM for Models 1 and 2 is automatically MAR, but to make Model 3 MAR, β_1 is set to 0.

Imputations from the SRMI or MN methods are rounded to fit into the given scales. Adaptive rounding is additionally used as an option for the binary diet variable (where rounding is based on a Normal approximation to the Binomial, making use of the marginal distribution in the observed data). For the SRMI approach, diet is imputed using a logistic regression, while health and physical fitness used ordinal (proportional odds) logistic regressions. The distress variable is either transformed to a Normal distribution (via log transformation and log transformation with an offset to make 0 skewness in the observed values) or is left as is. Imputations of 0 are then replaced with the smallest value in the sample, while observations above 3 are truncated at 3. For the SRMI approach, predictive matching is also used as an option for the Normal variables, imputing the observed value with the predictive mean closest to that of the imputation for a missing value. The regular combining rules are used on 20 imputed data sets for each method.

The authors find that the best results (under all missingness models) for the diet coefficient are obtained using the MN method that uses the zero-log-skewness adjustment or the prediction matching method in SRMI. All methods alleviate the biases and poor coverages of the complete-case analysis. For the other coefficients (not associated with the MDM), all methods provided adequate results, and it is shown that precision is improved by imputing rather than using complete-case analysis. The zero-log-skewness adjusted MN method performs even better than the SRMI approach with predictive matching in the context of coverage, when the adaptive rounding is used.

These studies results are important to note, since Lee & Carlin (2010) show that the MN method can be adjusted to impute properly even for a binary variable. It may seem that the added complexity inherent in an SRMI model may not always be justified. However, sensitivity to non-Normality may make the SRMI approach seem to be the more robust option. Of course, the inherent ease of dealing with ordinal and categorical variables in the SRMI model may be enough to sway a researcher towards that option.

SRMI vs SI

The study by van der Heijden, Donders, Stijnen & Moons (2006) compares SRMI with SI, complete-case analysis, and the missing-indicator approach; a form of the latter is used by Faris et al. (2002) in their non-imputation data enhancement technique that was compared with the MN method and SRMI, as discussed earlier in this subsection. The data set used by van der Heijden et al. (2006) consists of 398 consecutive patients 18 years or older who were referred to a Dutch hospital because acute pulmonary embolism (PE) was suspected. Numerous tests were completed on these patients, which found that 43% of the patients did have PE. The predictors chosen for the analysis of PE are based on those recommended by previous studies.

None of the outcome values are missing, although the covariate data is incomplete. In 38% of the patients, one or more predictors are missing, while the data is certainly not MCAR, since, based on the results of certain tests, doctors may have skipped subsequent tests, considering them to be uninformative given the prior test result. Van der Heijden et al. (2006) use complete-case analysis, the missing-indicator method[6], SI (conditional and unconditional mean imputation), and SRMI, with convergence within the imputed data sets occurring after five rounds, and create ten imputed data sets.

The authors then use a backward selection process for the overall model on each of these five methods, and compare regression coefficients, standard errors of the coefficients, and

[6] In this method, a missing value in a variable is recoded into a separate indicator variable attached to the incomplete variable itself, while the missing values in the incomplete variable are recoded into zeros, for example; complete case analysis will, then, not drop the observations for which these variables are 'missing'.

areas under the ROC curves — the so-called C statistic of Faris et al. (2002).

Van der Heijden et al. (2006) find that the model selected from complete-case analysis is different from those selected after the imputation procedures. This is natural, since the data is not MCAR. The standard errors are smallest for the conditional mean imputation (as expected). The indicator variables for the missing-indicator approach all achieve significance, while the coefficients in this method are larger than for the other models (since there are simply more coefficients). All areas under the ROC curves are above 0.75, but since there are more significant (but clinically meaningless) predictors in the missing-indicator approach, this method produces the highest area, albeit surely overestimated, while conditional mean imputation and SRMI produce the lowest values. The most significant outcome of this study is that the complete-case analysis is different to that of the post-imputation analysis, warning researchers of the hazards of simply relying on complete-case analysis methods. Interestingly, the missing-indicator, SI and MI methods' results do not differ greatly, in post-imputation analysis coefficient direction, magnitude and precision. This may be due to the relatively small amount of missing data. However, the authors do warn against the use of the missing-indicator approach, for several valid reasons (see van der Heijden et al. 2006, p. 1108).

SRMI vs S-HD

Barnes, Gutierrez-Romero & Noble (2006) use SRMI on data taken from the South African census in 2001. In this census, over 50% of individuals above the age of 18 report zero income, and additionally, 16% of the income values are missing (Barnes et al. 2006). The authors of this study compare the SRMI method with the imputation method used by Statistics South Africa (StatsSA) when they published the data, namely single hot-deck imputation S-HD. The S-HD process is the same as the HD process, except that only one value is imputed for every missing value.

As mentioned by Ardington et al. (2006), income is recorded in income bands, so the variable is essentially categorical. Other variables used in the study include age, gender, population group, employment status, occupation, education, income, province and location, the latter two being the only complete variables. As was done by bArdington et al.

(2006), implausible values of income are recoded as follows:

- If household income is zero, income is set missing for household members aged 15 and older, and to zero for those younger than 15.
- For those younger than 15 with recorded income greater than R6 400 per month, income is set missing.
- For those recorded as being employed but with zero income, income is set missing.

The results reported by this paper are minimal. The only statistics that are compared are the proportions lying in each of the income bands before and after imputation. Barnes et al.'s (2006) results are similar to those obtained by Ardington et al. (2006), as well as those obtained StatsSA's S-HD imputation procedure, although the latter seems to favour slightly higher income band imputations for all but the lowest income band (the zero band). The results of this study and that of Ardington et al. (2006) are similar although age and education are modelled as Poisson variables in Ardington et al.'s (2006) paper, while they are modelled as Normal variables by Barnes et al. (2006) — since the latter authors prefer not to assume a constant failure rate for the inter-arrival times within the variable — and although a simple logit regression is used for the income bands, instead of the ordered logit used by Ardington et al. (2006). The most noteworthy conclusion is that no major outlier problem exists, since such a problem would make the S-HD method's results differ substantially from those of SRMI.

A word of caution should be given after reviewing the paper by Barnes et al. (2006). It is important to remember that it is not just the point estimates that are of importance in assessing imputations, but also the confidence intervals generated by the procedure. As has been mentioned before, no single imputation procedure will produce valid confidence intervals, since the uncertainties associated with choosing an imputation model are not incorporated into the analysis estimates as they are in MI.

SRMI vs SI and MI HD/ABB

Ambler et al. (2007) compare the SRMI procedure with several others in a study of risk modelling. Complete case analysis, single imputation procedures and MI procedures are compared to SRMI in an analysis model with a binary outcome, *i.e.* a logistic analysis model. The single imputation procedures include mean, mean/mode and conditional mean imputation, all explained in Subsection 1.2.3, while the MI procedures include hot-decking via HD, HD_{CP}, HD_{obs} and HD_{Y} (outcome only) procedures combined with the ABB procedure, all explained in Subsection 2.3.1. The data used were the medical characteristics of 20 738 aortic and/or mitral valve surgery patients in Great Britain and Ireland between 1995 and 2003, with in-hospital mortality as the outcome variable.

The authors find the most prevalent missingness patterns (among observations), impute the data using SRMI, create binary responses from the fitted logistic regression coefficients based on the observed data (hereafter known as the true coefficients), and then find the (rescaled) fitted probabilities of each observation belonging to each of the most prevalent patterns, also using logistic regressions. Simulated data sets are created by sampling without replacement from the completed data. For each of these data sets binary responses were created from the true coefficients, and data was made missing according to both MCAR and MAR rules separately. For MCAR, the observations were randomly assigned to a missing data pattern according to that pattern's prevalence. For MAR, the same rule is applied, although in this case the covariates for an individual are set missing with probability equal to that of the individual having been in the particular pattern originally (the fitted probabilities calculated before). The different MI techniques are then applied to five imputed versions of each of these simulated data sets and the overall logistic regression is again run; this process allows a comparison between the true coefficients and the multiply imputed coefficients. The measures used to assess the imputations over 1000 simulated datasets are as follows:

- **Measure of agreement**: The proportion of observations correctly classified into the correct risk group.

- **Rank correlation**: A Spearman rank correlation between the true ranking of disease severity and the ranking after the imputation procedures on the simulated data sets.

- **Root mean squared error (RMSE)**: The RMSE between the fitted and true probabilities for patients.

- **Regression based calibration measure**: To assess the calibration of the fitted model, the fitted log-odds are regressed against the true log-odds for each simulation data set using a standard linear regression. The coefficients of this regression provide information about the calibration of the fitted model. Slope coefficients close to one are wanted.

- **Regression coefficients and confidence interval coverage**: The biases between the true regression coefficients and those obtained from the completed simulated data sets is assessed, averaging over all simulated data sets for a particular method.

In essence the authors show that mean/mode imputation provides an improvement over complete-case analysis. Mean imputation goes one step further, being less biased. Conditional mean imputation, however, outperforms both of these methods in this study. However, all methods suffer from the deficiency of not accounting for imputation model uncertainty, which is provided for in MI.

Of the MI measure, $\mathrm{HD}_{\mathrm{obs}}$ performs the worst, even worse than some of the single imputation procedures, but the authors admit that this may have been due to the setup of the simulated data sets (*i.e.* high proportions of missing data). SRMI, HD, $\mathrm{HD}_{\mathrm{CP}}$, HD_{Y} generally perform well and very similarly with respect to agreement, rank correlation and RMSE. Additionally, SRMI and HD_{Y} methods exhibit good calibration and provide better classification of low and high risk patients. This result may be due to the fact that these methods include the outcome in the imputation procedure. The SRMI procedure produces the lowest biases in the regression coefficients and the confidence intervals have coverage values close to the nominal level.

3.4.5 Non-Normal errors in the imputation regressions

He & Raghunathan (2009) contribute to this research area by assessing several methods of Normality-based SRMI when the underlying conditional distributions of the variables are non-Normal. In a simulation study, they assess the following sequential imputation methods when these methods are (incorrectly) applied to data that is non-Normal, with missing values that are MCAR:

- **Sequential Normal linear regressions**. This method is equivalent to imputation under the multivariate Normal model using Gibbs sampling (Schafer 1997, van Buuren 2007),[7] and is the method for Normal data specified in Subsection 3.3.1. In summary, assuming a Jeffrey's prior for β and σ^2, $P(\beta, \sigma^2) \sim 1/\sigma^2$, an ordinary least squares regression model is fitted to the n_{obs} complete cases, then a value σ_* is drawn with $\sigma_*^2 \sim SSE/\chi^2_{n_{pbs}-p}$, where SSE is the residual sum of squares from the OLS regression fit. Then a value β^* is drawn with $\beta^* \sim N(\hat{\beta}, \sigma_*^2(X'_{obs}X_{obs})^{-1})$, where $\hat{\beta}$ is the OLS estimate of β, and X_{obs} denotes the part of the covariate matrix which have corresponding Y's observed. Finally, for an incomplete case i, a value Y_i^* is imputed from $N(X_i\beta^*, \sigma_*^2)$.

- **Predictive mean matching (PMM)**. This method described by Schenker & Taylor (1996) has its origin in hot-deck imputation. Missing Y values are imputed from nearby complete cases. The predictive mean of an observation is given as $\hat{Y}_i = X_i\beta^*$ where β^* is drawn as in the Normal method given above. For each incomplete case this method draws an observation randomly from from a set of "possible donors" which are observations with predictive mean close to that of the incomplete case. The value of Y for the chosen case is then donated to the incomplete case.

- **Local residual draw (LRD)**. This method, also described by Schenker & Taylor (1996), imputes the value $Y_i^* = X_i\beta^* + r^*$, where r^* is drawn at random and with replacement from the set of complete "donor cases" as defined above. The method can adjust for the lack of fit of the Normal regression model by fitting local residuals, rather than by drawing local observed values, as in the previous method.

  Schenker & Taylor (1996) discuss the possible bias inherent in having too many donor cases for the previous two methods, and the possibly overstated correlation inherent in having too few donor cases. For this reason Schenker & Taylor (1996) develop an adaptive technique for choosing the number of donor cases. The authors find, however, that there is little difference in results between their adaptive

[7] Another special case relating joint modelling to the SRMI approach is when three variables are modelled using logistic regressions. The joint model for this is effectively a multivariate log-linear model with no three-way interaction term (see van Buuren 2007).

technique and a non-adaptive fixed number of donor cases. Additionally, He & Raghunathan (2009) also show that there is not much change in the overall analysis if different (reasonable) fixed numbers are used for the donor cases.

- **Adjustment of Normal regression by sampling from observed residuals (or expanded residual draw, or ERD)**. This method, proposed by Rubin (1987), is a modification of the sequential Normal method. This method first obtains the standardised residuals:

$$\frac{Y_i - X_i\beta^*}{\sqrt{\frac{SSE}{n_{obs}-p}}}$$

From these residuals, n_{mis} values are sampled with replacement (*i.e.* as many as there are missing observations), multiplied by σ_*, and, finally, added to $X_i\beta^*$. These standardised residuals then have the correct conditional moments, but a distribution whose shape is adjusted to reflect that of the actual error terms. In fact, this method's "donor set" for residual has just been expanded to include all complete cases, and the residuals donated are standardised, so the overall adjustment is partially parametric.

- **Adjustment by fitting Tukey's g-and-h distribution to errors**. The final method these authors analyse, is the Normal method adjusted to fit Tukey's (1977) g-and-h distribution to the error terms. Tukey (1977) proposed the gh family based on a transformation of the standard Normal Z,

$$T_{gh}(Z) = \mu + \tau\frac{e^{gZ}-1}{g}e^{hZ^2/2}, \tag{3.29}$$

where μ is the location parameter, $\tau(> 0)$ is the scale parameter, and g and h are scalars that govern the skewness and kurtosis or elongation of the data, respectively. He & Raghunathan (2009) use a linear regression model for their method, with their error terms modelled using the centered gh distribution,

$$\begin{aligned} Y_i &= X_i\beta + \epsilon_i, \qquad (3.30)\\ \epsilon_i &= \tau\left(\frac{e^{gZ}-1}{g}e^{hZ^2/2} - E_{gh}\right), \end{aligned}$$

where $E_{gh} = \frac{1}{g\sqrt{1-h}}\left(e^{g^2/[2(1-h)]} - 1\right)$ is the mean of the standardised gh distribution with $\mu = 0$ and $\tau = 1$ in Equation 3.29. First, β^* is drawn as if the error distribution is Normal, and then parameters τ, g, and h are estimated from a bootstrap sample of the observed residuals $Y_i - X_i\beta^*$ using a quantile-based method (He & Raghunathan 2009, Appendix). Then, for a missing case i, independent standard Normal Z_i's are simulated and the missing value of Y_i is estimated as $Y_i = X_i\beta^* + \tau(\frac{e^{gZ}-1}{g}e^{hZ^2/2} - E_{gh})$.

Each of these imputation methods is applied sequentially and multiply (with each dataset imputed five times) on the following simulated data, with 20% missing values generated completely at random:

$$\begin{aligned} Y_1 &\sim U(0,2), \\ Y_2 &= 1 + Y_1 + \epsilon_2 \\ Y_3 &= 1 + Y_1 + Y_2 + \epsilon_3 \end{aligned}$$

The authors then consider two sets (one with less variation and one with more variation) of each of the following distributions for each of ϵ_2 and ϵ_3: Lognormal, centred Student's t, and Uniform. More details of the processes used in their study are given in Chapter 4.3.

Their simulation study consists of 1000 replicates, and each replicate includes 1000 cases. On each replicate, once missing values have been generated, the given SRMI procedures are applied, and in each case 5 imputed data sets are created (within each data set the SRMI procedure is iterated 5 times). If "donor cases" are chosen in any method, their number is restricted to 20. The quantities of interest in their study are the marginal mean of Y_3, the proportions of Y_3 that are less than its different population quantiles (5%, 25%, 50%, 75%, and 95%), and the coefficients of regressing Y_3 on Y_1 and Y_2. Inferences made before deleting missing values are taken as a benchmark, while the results from applying complete-case analysis on the incomplete data are used. The performance of these methods is evaluated using relative bias, $RBIAS$, and the root of the relative mean squared error, $RRMSE$.

$$RBIAS = \left|\frac{Bias}{True}\right| \times 100\% \tag{3.31}$$

$$RRMSE = \sqrt{\frac{MSE(Method)}{MSE(Before\ deletion)}} \tag{3.32}$$

Additionally, coverage rates of the 95% confidence intervals across the 1000 replicates are examined, which should be close to nominal if the imputation method is working well. A reasonable upper limit for $RBIAS$ is 5%. $RRMSE$ measures the increase of $RMSE$ relative to that of the analysis before deletion of the missing observations. In general, it is expected to be more than 1, reflecting loss of efficiency when analysing incomplete data. However, it can be less than 1, too, implying "superefficiency" of the MI models, as discussed in Rubin (1996). To recap, superefficiency happens when the analysis uses more (correct) data than would be used in the complete-case analysis of the incomplete data, *i.e.* the imputation methods are imputing correct values.

In essence, the study shows that the methods are reasonable for both estimating means and proportions (although the sequential Normal method is the worst for the proportions), and coverage rates are adequate. However, when estimating a regression coefficient for a regression on the completed data, all methods are left wanting when the ϵ's follow the distributions with the wider variances. The key conclusion from this study is that it is extremely important for a researcher to analyse the incomplete data thoroughly before applying an imputation method, since it is shown that simply applying a regular Normal method (even one adjusting from non-Normal errors) might not be adequate for a particular estimation procedure in the presence of errors with non-Normal distributions and large variances.

Following on the study by He & Raghunathan (2009), de Jong, van Buuren & Spiess (2014) investigate a possible MI solution for missingness in non-Normal variables. The authors focus on univariate missingness, conditional on other covariates (*i.e.* MAR missingness for a single variable). The application is simple enough to extend to SRMI, but de Jong et al. (2014) choose not to do this so that their new approach can be assessed without the distracting interactions that might arise when solving a multivariate missingness problem.[8] They compare multiple imputation from linear models (LMs), GLMs, linear and nonlinear PMM, and their novel approach, generalised additive models for location, scale and shape

[8] This is the approach followed in Chapter 6 in this thesis.

(GAMLSS). In GAMLSS, a target distribution is chosen for the incomplete variable, say Normal or generalised Beta. If the Normal distribution is chosen, fitting and smoothing algorithms are used to model the mean and variance of the distribution conditional on other known covariates. If the Beta distribution is chosen, then fitting and smoothing algorithms conditionally model the skewness and kurtosis of the incomplete variable in addition to the mean and the variance.

In simulation experiments, de Jong et al. (2014) find that when the incomplete variable is originally Normal, Normal-based GAMLSS performs adequately and comparable to the LM MI, which is expected to perform best (since there is no model misspecification). There is slight loss in efficiency for the GAMLSS model, due to its flexibility. When the incomplete variable is Uniform, then the authors show that a generalised Beta GAMLSS is robust and performs well. When the incomplete variable is distributed as a Uniform squared, then GAMLSS performs well if it is based on the Normal distribution. If the incomplete variable is distributed as a Student's t with three degrees of freedom, however, then although GAMLSS seems to provide the best results of the given methods, none of the considered models provide adequate imputations. This seems to suggest that even GAMLSS is not suitable for imputation on variables with heavy-tailed distributions (de Jong et al. 2014, p. 21). Finally, Normal-based GAMLSS seems to perform well for a Poisson-type incomplete variable, even when the Poisson GLM and polytomous regression model underperform. In conclusion, the authors suggest that GAMLSS imputation should be used if there is uncertainty concerning the the implementation of a parametric imputation model, providing an alternative to PMM MI. However, they concede that additional models will need to be considered for heavy-tailed imputation procedures.

3.4.6 Additional SRMI diagnostics

Kolmogorov-Smirnov, scatterplots and residuals

Although the diagnostic methods presented in this section can be applied to any multivariate imputation method, Abayomi, Gelman & Levy (2008) apply these diagnostic measures in the SRMI context. Abayomi et al. (2008, p. 273) mention that "the development of diagnostic techniques for MI... has been retarded by the belief that the assumptions of the procedure are untestable from observed data.” In particular, researchers believe the MAR assumption is untestable, since imputed values are merely guesses for unobserved, unknown data. The response of Abayomi et al. (2008) is twofold:

1. In the context of the problem being studied, differences between imputed and observed values can be examined, while the distribution of the completed data as a whole (as well as the imputations alone) can also be checked to be sensible (when compared with the observed data). This is an external diagnostic — a comparison with outside knowledge.

2. The fit of the regression models that are used in the imputation process can be assessed. This is an internal diagnostic — specific to observations and modelling.

From these responses, it is clear that there is no internal diagnostic for the MAR mechanism, or any hypothesised MNAR mechanism.

When an incomplete variable contains gross outliers, predictive regression models will impute values biased towards the outliers. For this reason, diagnostic measures for this type of case are distributional plots before and after imputation, and of the imputations themselves.

Using SRMI, imputers can check the distributional plots from the specific predictive regression models against those of the observed data — deviations suggesting a possible model error or possible deviation from the assumed MDM. However, it is not certain that if the distributions differ there is indeed a problem, since some deviation is expected from a MAR MDM.

Abayomi et al. (2008) *firstly* compare the empirical distributions of the observed and imputed data using the Kolmogorov-Smirnov (KS) test for each variable, as well as visually. The p-value from the KS test is approximate, since the two sets of values are not independent; the MAR assumption means that the imputed values depend on the observed. Any dramatic differences between the distributions deserve attention. Since the p-values are approximate, the authors suggest examining the 10% of the variables whose KS tests are the most deviant. *Secondly*, the authors examine bivariate scatterplots of the variables against both internal and external measures. *Thirdly*, Abayomi et al. (2008) examine the residuals from the conditional regression models in the SRMI procedure. By fitting a LOWESS curve (see Cleveland 1979), the authors are then able to improve the conditional regressions using residual refinement, but only if the assumption of random missingness is true.

The authors found these methods to be quite helpful in diagnosing problems with the MI model used in an SRMI procedure on the Environmental Sustainability Index (ESI) of 2001. While these methods might not diagnose the problems themselves, at least they help identify the symptoms of the possible problems.

Calibrated posterior predictive p-value

The calibrated posterior predictive p-value (or *cppp*) is introduced by Hjort, Dahl & Steinbakk (2006) and reviewed by Cabras, Castellanos & Quirós (2011) as a goodness of fit (GOF) measure for the (parametric) Bayesian regressions chained within SRMI.

The *cppp* is based on the posterior predictive p-value, or *ppp*. Let Y be the variable that is being imputed following the sampling model $f\left(y|\beta\right)$ where $\beta \in \Theta$ is distributed according to prior $\pi\left(\beta\right)$, then the *ppp* is defined as

$$ppp\left(y\right) \quad = \quad \Pr\left(D\left(Y^{rep}, \beta\right) \geq D\left(y, \beta\right)|y\right), \tag{3.33}$$

where D is a discrepancy measure[9], y represents the observed data, and the distribution of Y^{rep} is the posterior predictive distribution, $p\left(Y|y\right) = \int_{\Theta} f\left(y|\beta\right)\pi\left(\beta|y\right)d\beta$. Without loss of generality, larger values of Equation 3.33 indicate incompatibility of the regression model. The *ppp* is usually approximated by a Monte Carlo sum where β and Y^{rep} are drawn from the posterior distributions $\pi\left(\beta|y\right)$ and $f\left(Y|\beta\right)$ respectively (Cabras et al. 2011, p. 430).

Due to the fact that the *ppp* cannot be interpreted under the Uniform$(0, 1)$ distribution, as noted by Hjort et al. (2006) and Cabras et al. (2011) amongst others, it is suggested that the *cppp* is used to overcome this disadvantage. This *cppp* is a post-processed *ppp*

[9] It should also be noted that the choice of the discrepancy measure also influences the GOF, so some thought has to go into the choice of D.

measure. Mathematically,

$$cppp\,(y) \;=\; \Pr\left(ppp\,(Y) \leq ppp\,(y)\right) \tag{3.34}$$

where Y comes from the prior predictive distribution, $p\,(Y) = \int_{\Theta} f\,(y|\beta)\,\pi\,(\beta)\,d\beta$. The *cppp* is then Uniform$(0,1)$ under the null model, as opposed to the *ppp*, which is not. However, proper priors are required for β, in an area in which default and improper priors are usually used. Cabras et al. (2011) provide a possible solution to this problem by using minimal training samples to turn improper priors into proper ones.

Upon preliminary review, it seems as though this *cppp* statistic could be valuable in assessing the need for a continuous-data SRMI model that deviates from the Normal distribution. The *cppp* is part of a statistical test to determine objectively whether data fits the chosen model. For this reason, the *cppp* method, some discrepancy measures, and the practical details concerning the Monte Carlo simulation within the *cppp* measure will be discussed in Chapter 5.

3.4.7 Problems in SRMI

The most obvious problem associated with SRMI is the need to specify models for each incomplete variable. This need complicates the modelling process, even though the more complicated process may be more accurate in certain contexts. Herein lies the second obvious problem — the effectiveness and accuracy of SRMI has not been studied in enough detail, given that the joint distribution of the specified conditionals might not exist (van Buuren et al. 2006). This problem is discussed in more detail below. Finally, an issue which seems to be more of a debate on correct procedure than an insurmountable obstacle is the issue of efficiency in the Gibbs sampling algorithm. This issue is discussed in the next paragraph.

Gibbs sampling and efficiency

A criticism offered by Nielson (2003) is that, when running the Gibbs sampler, a better estimator might be obtained by combining the estimates from the last k iterations in each of the m chains, rather than just the final estimate from each of the m chains. One can effectively even reduce the value of m. This process will reduce simulation noise. In fact, simulation noise could be reduced by running the iteration chains for longer, and increasing k. Meng & Romero (2003) note, however, that if this is done, the km imputations are no longer conditionally independent, given the observed data — which makes subsequent analyses much more complicated. One would have to derive more general combining rules for specific dependent structures, or a general set of combining rules for arbitrarily independent and dependent multiply-imputed data sets. It may be more practically efficient (although less theoretically efficient) if km multiple datasets are produced in the standard way — at least then each will be independent from all the others.

Gibbs sampling and posterior approximations

Van Buuren et al. (2006), van Buuren (2007) and He & Raghunathan (2009) mention that an uncertain issue in SRMI is that of the possible incompatibility of the conditional models specified, or, rather, that a joint distribution may not exist given the conditional distributions of the assumed forms. As mentioned by van Buuren et al. (2006), the implicit joint distribution will not exist if the space spanned by the parameters of k incomplete variables has more dimensions than is appropriate — this occurs when these parameters are not independent of each other, which is commonly the case. Little & Rubin (2002) also warn that if the sampling algorithm does not converge, then there an approximate posterior does not exist. In fact, van Buuren et al. (1999) mentioned this incompatibility when sequential conditional models were first being considered. In their paper, imputations on two variables are made sequentially, and iterated, as is regularly done in SRMI, although, as already discussed, these authors call the process "regression switching". They mention that under quite general conditions the draws from this sequential process converge to the appropriate multivariate posterior density, $P(Y_{mis}|Y_{obs}, X, R)$, as in a Gibbs sampling framework. However, they note that the specification of the two conditional distributions for the incomplete variables $P(Y_1|Y_2)$ and $P(Y_2|Y_1)$ could be incompatible, in that no joint distribution $P(Y_1, Y_2)$ exists. If this is the case, the algorithm will alternate between the isolated conditional distributions, although the results may still be adequate when evaluated by classic frequentist criteria, as shown by Brand (1998). A problem with having incompatible conditional distributions is highlighted by van Buuren et al. (2006), when they mention that there may be a possible effect of this incompatibility on imputation inferences. Gelman (2004), however, points out that having a joint distribution may be less important than incorporating other information about the specific variables that joint modelling may not be able to do, such as zero/nonzero features bounds, skip patterns, nonlinearity and interactions.

Gelman & Rubin (1992) develop a monitoring statistic in their work that looks at the statistics based on iterative procedure like the Gibbs sampler. This monitoring statistic continually assesses the variation between and within simulated sequences, until within variation roughly equals between variation. This is the point at which the statistics based on these sequences converge, since, "[o]nly when the distribution of each simulated sequence is close to the distribution of all the sequences mixed together can they all be approximating the target distribution" (Little & Rubin 2002, p. 206).

Gelman & Rubin (1992) use this monitoring statistic in a three-stage target distribution simulation procedure. *Firstly* they generate a starting distribution with the same number of modes as the target distribution; *secondly* they overlay this starting distribution with an over-dispersed approximation; *thirdly*, they downweight draws from the overlaid approximation that have relatively low density under the target distribution (*i.e.* here the authors use importance re-sampling). The monitoring statistic is used to determine how much closer draws from the approximating distribution will be to the target distribution if iterative simulations can continue indefinitely.

The process to calculate the monitoring statistic is as follows. For each scalar estimand ψ, label the draws from D parallel sequences[10], each of length L, as $\psi_{d,l}(d = 1, \ldots, D, l =$

[10] Note that D in these expressions is equivalent to m in all other parts of this thesis.

$1, \ldots, L$), and compute $B^{(seq.)}$ and $\bar{V}^{(seq.)}$, the between and within sequence variances:

$$B^{(seq.)} = \frac{L}{D-1} \sum_{d=1}^{D} \left(\bar{\psi}_{d\cdot} - \bar{\psi}_{\cdot\cdot} \right), \tag{3.35}$$

where

$$\bar{\psi}_{d\cdot} = \frac{1}{L} \sum_{l=1}^{L} \psi_{d,l} \; ; \quad \bar{\psi}_{\cdot\cdot} = \frac{1}{D} \sum_{d=1}^{D} \psi_{d\cdot} \; ,$$

and

$$\bar{V}^{(seq.)} = \frac{1}{D} \sum_{d=1}^{D} s_d^2 \; , \tag{3.36}$$

where

$$s_d^2 = \frac{1}{L-1} \sum_{l=1}^{L} \left(\psi_{l,t} - \bar{\psi}_{d\cdot} \right)^2 .$$

The marginal posterior variance of the estimand, $V(\psi|Y_{obs})$ can be estimated by a weighted average of $\bar{V}^{(seq.)}$ and $B^{(seq.)}$,

$$\widehat{V}^{+} \left(\psi | Y_{obs} \right) = \frac{L-1}{L} \bar{V}^{(seq.)} + \frac{1}{L} B^{(seq.)}, \tag{3.37}$$

which overestimates the marginal posterior variance assuming the starting distribution is appropriately overdispersed, but is unbiased if the starting distribution equals the target distribution (*i.e.* under stationarity). This estimate is analogous to the classical variance estimate for cluster sampling. As $T \to \infty$, the expectation of $\bar{V}^{(seq.)}$ approaches $V(\psi|Y_{obs})$. For this reason, one can monitor the potential scale reduction, $\widehat{SR}$, that the current distribution of ψ could be reduced by, should $T \to \infty$, namely:

$$\sqrt{\widehat{SR}} = \sqrt{\frac{\widehat{V}^{+} \left(\psi | Y_{obs} \right)}{\bar{V}^{(seq.)}}} \; . \tag{3.38}$$

Clearly, $\sqrt{\widehat{SR}} \to 1$ as $T \to \infty$. So, if this scale reduction is not below a certain threshold, say, 1.2, then the sampling iterations should continue. When the threshold is reached, subsequent draws from the D sequences can be treated as draws from the target distribution.

Even though these monitoring measures exist, it is important to note once more the results obtained by van Buuren et al. (2006), reviewed earlier in this chapter. These authors find that incompatible conditionals are potentially less of a problem in the SRMI context than they were originally thought to be. The use of the monitoring statistics however, should provide backing for SRMI for even the most sceptical of SRMI researchers.

Further research areas identified

Van Buuren et al. (2006) mention that one area needing work in SRMI studies is that of choosing overdispersed starting values for multivariate missing data problems. Another research area identified by these authors is the maintenance of higher order interactions within the SRMI paradigm. The latter is clearly related to the need to make a MNAR MDM more MAR, by including additional explanatory variables into the SRMI models, although neither of these problems falls within the context of this thesis' goal.

He & Raghunathan (2009) also believe that there is a general lack of research on robustness of the SRMI algorithms when the conditional models are misspecified. Moreover, most SRMI algorithms use linear regressions. This thesis will, therefore, in part attempt to add to the literature on the study of robustness of SRMI under misspecification of the assumed conditional regression models, although non-linearity will not be addressed.

3.5 Conclusion

Complex k-dimensional (incomplete) data structures can often more easily and more accurately be modelled by k one-dimensional models than by a full k-dimensional model (Gelman & Rubin 1992, van Buuren et al. 2006). This is, of course, due to the fact that models for individual variables in a data set can easily incorporate limits or bounds, skip patterns, interactions, and other distributional complications. These separate univariate distributional complications are not easy to build into a joint model of a data set.

For this reason, SRMI has become increasingly popular over recent years, in vastly different fields, such as economics and medicine. This popularity has resulted in breakthroughs not only in the SRMI field, but also in MI in general as well.

Breakthroughs within the field of SRMI include increased understanding of the advantages and limitations of the method and incorporated procedures, advances in diagnostic measures, and comparisons with other MI procedures.

Advances in the MI field include research into MDMs, research into imputing under non-Normal errors in data, and advancing the field of synthetic data utilisation. The advantages of both joint modelling MI and SRMI have been combined in nested modelling, making use of the best of both MI and SRMI worlds.

While these advances have been significant, much remains to be uncovered in the field. The need for a robust SRMI model is the primary focus of the new work introduced in the chapters to follow.

Chapter 4

A New Robust Sequential Regression Model

4.1 Introduction

Investigation of the literature on MI and SRMI suggests that there is a need for a robust model within SRMI that can handle heavy-tailed and possibly skew data.[1] Such a model could be chosen as a default within an SRMI routine instead of the Normal regression model, because this default model would be able to handle non-Normal errors (including heavy tails and skewness).

One model which could fill the role of a robust sequential regression model is the Student's t-distribution. With heavier tails than the Normal distribution, and the possibility of incorporating a skewness parameter, the t-distribution model could serve as a robust counterpart to the Normal OLS regression model (even with the PMM, LRD and ERD adaptations introduced in Subsection 3.4.5). If the errors are indeed Normal, then this robust model will be able to reduce to the Normal case by increasing the degrees of freedom of the t-distribution. In this thesis, the t model will be built using a Bayesian paradigm.

The objective of this chapter is to show that the skew t-distribution in SRMI can reproduce

[1] Robust in the context of yielding satisfactory imputation results whether the underlying data model is Normal, non-Normal, or possibly even initially misspecified.

the error distribution under a variety of Normal and non-Normal symmetric and skew specifications. Additionally, beyond simply replicating the original distributions, we would like to show that the imputations made from the skew t-distribution have good coverage of the original data points that are made missing.

4.2 The Student t-Distribution

We follow the setup presented in Sahu, Dey & Branco (2003) and Fonseca, Ferreira & Migon (2008, p. 326). Consider a linear regression model in which an observation vector $y = (y_1, \dots, y_n)'$ satisfies

$$y = X\beta + \epsilon \tag{4.1}$$

where $\beta = (\beta_0, \beta_1, \dots, \beta_p)$, $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$ is the error vector and $\epsilon_1, \dots, \epsilon_n$ are i.i.d. according to the Student-t distribution with location zero, scale parameter σ and ν degrees of freedom. Here $X = [x_1, \dots, x_n]'$ is the $n \times p$ matrix of explanatory variables, taken to be of full rank p. We denote the model parameters by $\theta = (\beta, \sigma, \nu) \in \mathbb{R}^p \times (0, \infty)^2$. Thus, the likelihood function is given by:

$$L\left(\beta, \sigma, \nu | y, X\right) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)^n \nu^{n\nu/2}}{\Gamma\left(\frac{\nu}{2}\right)^n \pi^{n/2} \sigma^n} \prod_{i=1}^{n} \left[\nu + \left(\frac{y_i - x_i'\beta}{\sigma}\right)^2\right]^{-(\nu+1)/2}. \tag{4.2}$$

The likelihood for the t-distribution given in Equation (4.2) can be restructured as follows:

$$L \propto \prod_{i=1}^{n} \left(\frac{\lambda_i \tau}{2\pi}\right)^{\frac{1}{2}} \exp\left[-\frac{\tau}{2}\left(y_i - \beta x_i\right)^2\right] \times \prod_{i=1}^{n} \left[\frac{(\nu/2)^{\nu/2}}{\Gamma\left(\nu/2\right)} \lambda_i^{\nu/2-1} \exp\left(\frac{-\nu\lambda_i}{2}\right)\right] \tag{4.3}$$

where x_i is the row of the covariate matrix X that corresponds to observation y_i, $\tau = \sigma^{-2}$ and the λ_i are weights indicating the influence of each observation on ν. Integrating out the λ_i in Equation (4.3) yields Equation (4.2).

4.2.1 Fitting the t-distribution

The conditional distributions of the parameters

When the t-distribution is used for errors on the posterior predictive distribution, generating the imputations is simply a matter of applying the posterior-drawn regression parameters to the covariates and adding an appropriate t error. The difficulty is in finding the degrees of freedom for this error. This thesis uses a Gibbs sampling process for the parameters σ^2, $\lambda_i, i = 1, \ldots, n$, β, and ν, while ν itself is drawn via a Metropolis-Hastings algorithm in each step of the Gibbs sampler. The Gibbs sampler requires the formulation of the conditional posterior distributions for each of the parameters of the model. These conditional distributions are presented within this subsection.

Given Equation (4.3), and ignoring uninformative priors, it can be shown that the conditional distribution of the β vector is multivariate Normal, namely,

$$\beta|y, \sigma^2, \Lambda, X \sim N\left\{\left(X'\Lambda X\right)^{-1} X'\Lambda y, \sigma^2 \left(X'\Lambda X\right)^{-1}\right\}, \tag{4.4}$$

where the matrix Λ is the diagonal matrix with diagonal elements $\lambda_1, \lambda_2, \ldots, \lambda_n$.

The conditional posterior of σ^2 is derived to be an Inverse Gamma distribution, such that

$$\frac{\left(y - X\beta\right)' \Lambda \left(y - X\beta\right)}{\sigma^2}|y, \beta, X \sim \chi^2_n. \tag{4.5}$$

The posterior for ν, conditional on Λ, and its priors, is given in the following equations.

$$p\left(\nu|y, \Lambda\right) \propto \frac{\nu^{\frac{1}{2}\nu n}}{2^{\frac{1}{2}\nu n}\left[\Gamma\left(\frac{\nu}{2}\right)\right]^n} \left|\Lambda\right|^{\frac{1}{2}\nu - 1} \exp\left[-\frac{1}{2}\nu \sum_{i=1}^{n} \lambda_i\right] p\left(\nu\right), \tag{4.6}$$

with the prior on ν taking one of four forms, namely the truncated Exponential[2],

$$p\left(\nu\right) \propto e^{-\nu\xi}, \nu > 2, \xi = 0, 1, \tag{4.7}$$

[2] This distribution is truncated so that the mean and the variance exist (Sahu et al. 2003)

the Independence Jeffrey's prior,

$$p_{IJ}(\nu) \propto \left(\frac{\nu}{\nu+3}\right)^{\frac{1}{2}} \left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+3)}{\nu(\nu+1)^2}\right]^{\frac{1}{2}}, \tag{4.8}$$

the probability-matching prior or reference priors for the orders (ν,μ,σ^2), (ν,σ^2,μ), and (μ,ν,σ^2),

$$p_{PM,R1}(\nu) \propto \left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+3)}{\nu(\nu+1)^2}\right]^{\frac{1}{2}}, \tag{4.9}$$

and the reference priors for the orders (μ,σ^2,ν), (σ^2,μ,ν), and (σ^2,ν,μ),

$$p_{R2}(\nu) \propto \left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+5)}{\nu(\nu+1)(\nu+3)}\right]^{\frac{1}{2}}. \tag{4.10}$$

Note that $\psi'(\cdot)$ is the trigamma function. Derivations for these priors are given in Appendix A. Working with the natural log posterior and log priors, which is easier, we have,

$$\begin{aligned}\log(p(\nu|y,\lambda)) \propto{}& \frac{1}{2}\nu n\log(\nu) - \frac{1}{2}\nu n\log(2) - n\log\left(\Gamma\left(\frac{\nu}{2}\right)\right) \\ & - \left(\frac{1}{2}\nu-1\right)\sum_{i=1}^{n}\log(\lambda_i) - \left(\frac{1}{2}\nu-1\right)\sum_{i=1}^{n}\lambda_i - \log(p(\nu)),\end{aligned} \tag{4.11}$$

$$\begin{aligned}\log(p_{IJ}(\nu)) \propto{}& \frac{1}{2}(\log(\nu) - \log(\nu+3)) \\ & + \frac{1}{2}\log\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+3)}{\nu(\nu+1)^2}\right],\end{aligned} \tag{4.12}$$

$$\log(p_{PM,R1}(\nu)) \propto \frac{1}{2}\log\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+3)}{\nu(\nu+1)^2}\right], \tag{4.13}$$

$$\log(p_{R2}(\nu)) \propto \frac{1}{2}\log\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+5)}{\nu(\nu+1)(\nu+3)}\right]. \tag{4.14}$$

For $\lambda_i, i = 1, \ldots, n$, we can show that

$$\lambda_i | y, \beta, \tau, \nu \sim \Gamma \left\{ \frac{1}{2} (\nu + 1), \left[\frac{1}{2} \lambda_i \left(\frac{1}{\sigma^2} (y_i - x_i \beta)^2 + \nu \right) \right]^{-1} \right\}. \qquad (4.15)$$

The algorithm for the Gibbs sampler (and Metropolis sampler for ν) utilises these conditional distributions. We initialise the parameters, and then update multiple times via these conditional distributions.

Programmable algorithm for fitting the t-distribution

1. Initialise the following parameters:

$$\begin{aligned} \sigma^2_{(1)} &= var\,(y) \\ \Lambda_{(1)} &= 0.5 I_{n \times n} \\ \beta_{(1)} &= \sigma^2 \left(X' \Lambda X \right)^{-\frac{1}{2}} \times \epsilon + \left(X' \Lambda X \right)^{-1} \left(X' \Lambda y \right) \end{aligned}$$

   where ϵ is a $((p+1) \times 1)$ vector of standard Normal draws, X is the observed covariate matrix, y is the vector of observed responses, and $I_{n \times n}$ is the $n \times n$ identity matrix. The 0.5 coefficient of $\Lambda_{(1)}$ was chosen through trial and error.

2. Initialise ν by drawing from the Metropolis sampler for ν: preseed a ν value as $10 \times u + 4$, where u is drawn from a $U(0, 1)$ distribution, and set the log of the posterior of ν as $-\infty$; jump away from the preseeded value by a value drawn from a $N(0, 0.25)$ distribution (making sure the result is greater than 2); compare the log posterior of ν from the jumped-to value of ν to the log posterior of ν from the jumped-from value of ν. If the difference is greater than a random draw from a $U(0, 1)$ distribution, then the new, or jumped-to value of ν is retained. Repeat this process to obtain at least 100 ν values (the burn-in period), and then take the next draw of ν as the draw from the conditional posterior of ν in the Gibbs sampler.

3. Now the parameters σ, Λ, β, and ν have initial values $\sigma_{(1)}, \Lambda_{(1)}, \beta_{(1)}$, and $\nu_{(1)}$. We then start the updating process within the Gibbs sampler: for $s = 2, 3, \ldots, m$, where m is the number of update rounds including burn-in and the number of draws required

from the conditional posterior distributions:

$$\sigma^2_{(s)} = \frac{\left(y - X\beta_{(s-1)}\right)' \Lambda_{(s-1)} \left(y - X\beta_{(s-1)}\right)}{u}, \tag{4.16}$$

where u is a random draw from a χ^2_n distribution;

$$\beta_{(s)} = \left[\sigma^2_{(s)} \left(X'\Lambda_{(s)}X\right)^{-1}\right]^{\frac{1}{2}} \times \phi + \left(X'\Lambda_{(s)}X\right)^{-1} X'\Lambda_{(s)}y, \tag{4.17}$$

where ϕ is a $p \times 1$ vector of random draws from a $N\left(0,1\right)$ distribution; and $\Lambda_{(s)}$ is a diagonal matrix with diagonal elements $\lambda_{i,(s)}$,

$$\lambda_{i,(s)} = \frac{w}{\nu_{(s-1)} + \frac{1}{\sigma^2_{(s)}} \left(y_i - x_i'\beta_{(s)}\right)^2} \tag{4.18}$$

where $i = 1, 2, \ldots, n$, x_i' is the i^{th} row of the covariate matrix (corresponding to the i^{th} observation), y_i is the response for the i^{th} observation, and w is a random draw from a $\chi^2\left(\nu_{(s-1)} + 1\right)$ distribution. A new $\nu_{(s)}$ is then drawn from the same procedure as in Step 2, except this time it is conditional on y and $\Lambda_{(s)}$.

This simulation procedure becomes difficult to process when the number of covariates in X becomes large. In this case, inverting $X'\Lambda_{(s)}X$ can be computationally time-consuming, and for this reason the simpler non-matrix alternative procedure for estimating the parameters for the t-distribution is considered as well. Furthermore, the non-matrix representation is easily extended to incorporate skewness, which is critical to this research.

The (alternative) conditional distributions of the parameters

For ease of notation within the following equations, we calculate $\tilde{y}_{iq}$ for each observation $i = 1, \ldots, n$ and covariate $q = 0, 1, \ldots, p$, where $\tilde{y}_{iq} = y_i - \beta_{-q}X_{-q}$, where $-q$ represents all variables in X besides variable q. In other words, for $q = 0$:

$$\tilde{y}_{i0} = y_i - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_p x_{ip} \tag{4.19}$$

For $q = 1$:

$$\tilde{y}_{i1} = y_i - \beta_0 - \beta_2 x_{i2} - \beta_3 x_{i3} - \ldots - \beta_p x_{ip} \tag{4.20}$$

For $q = 2, \ldots, p$:

$$\tilde{y}_{iq} = y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_{q-1} x_{i(q-1)} - \beta_{q+1} x_{i(q+1)} - \ldots - \beta_p x_{ip} \tag{4.21}$$

Finally, for $q = p$:

$$\tilde{y}_{ip} = y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_{p-1} x_{i(p-1)} \tag{4.22}$$

It should be noted that instead of the parameter σ^2, the parametrisation $\tau = \sigma^{-2}$ is used in the following equations.

It can be shown that the following conditional distributions exist for the β_q:

$$\begin{aligned}&\beta_q | y, \beta_{-q}, \tau, \Lambda \sim \\ &N\left\{\left(\tau \sum_{i=1}^{n} \lambda_i x_{iq}^2 + \frac{1}{\sigma_{\beta_q}^2}\right)^{-1} \left(\tau \sum_{i=1}^{n} \lambda_i x_{iq} \tilde{y}_{iq} + \frac{\mu_{\beta_q}}{\sigma_{\beta_q}^2}\right), \left(\tau \sum_{i=1}^{n} \lambda_i x_{iq}^2 + \frac{1}{\sigma_{\beta_q}^2}\right)^{-1}\right\}\end{aligned} \tag{4.23}$$

where x_{iq} is element (i, q) of the data matrix X (and when $q = 0$, $x_{i0} = 1$ for all i), and μ_{β_q} and $\sigma_{\beta_q}^2$ are the conjugate Normal prior mean and variance for β_q respectively.

For τ, it can be shown that:

$$\tau | y, \beta, \Lambda \sim \Gamma\left\{\frac{n}{2} + a_\tau, \left(\frac{1}{2}\sum_{i=1}^{n} \lambda_i \hat{y}_i^2 + 2b_\tau\right)^{-1}\right\} \tag{4.24}$$

where a_τ and b_τ are the conjugate Gamma prior parameters for τ, and $\hat{y}_i = y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip}$.

The conditional posterior for ν is identical to that given in the above list of conditional posteriors in the matrix-formation of these posteriors.

As for the λ_i, it can be shown that

$$\lambda_i | y, \beta, \tau, \nu \sim \Gamma \left\{ \frac{1}{2} (\nu + 1), \left[\frac{1}{2} \left(\tau \hat{y}_i^2 + \nu \right) \right]^{-1} \right\}. \tag{4.25}$$

This is the same distribution as given above, except with σ^2 replaced by $1/\tau$.

Programmable (alternative) algorithm for fitting the t-distribution

1. Initialise the following parameters:

$$\begin{aligned} \tau_{(1)} &= \frac{1}{var\,(y)} \\ \Lambda_{(1)} &= 0.5 I_{n \times n} \\ \beta_{(1)} &= \sigma^2 \left(X' \Lambda X \right)^{-\frac{1}{2}} \times \epsilon + \left(X' \Lambda X \right)^{-1} \left(X' \Lambda y \right) \end{aligned}$$

where ϵ is a $((p+1) \times 1)$ vector of standard Normal draws, X is the observed covariate matrix, y is the vector of observed responses, and $I_{n \times n}$ is the $n \times n$ identity matrix.

2. Draw an initial $\nu_{(1)}$ in exactly the same manner as before, using the same Metropolis sampler described above.

3. Now the parameters τ, Λ, β, and ν have initial values $\tau_{(1)}, \Lambda_{(1)}, \beta_{(1)}$, and $\nu_{(1)}$. Start the updating process within the Gibbs sampler, for $s = 2, 3, \ldots, m$, where m is the number of update rounds including burn-in and the number of draws required from the conditional posterior distributions. First, in each step, recalculate the $\tilde{y}_{iq,(s)}$ using $\beta_{-q,(s-1)}$. Then, for each $\beta_q, q = 0, 1, \ldots, p$,

$$\begin{aligned} \beta_{q,(s)} = & \left(\tau_{(s)} \sum_{i=1}^{n} \lambda_{i,(s-1)} x_{iq}^2 + \frac{1}{\sigma_{\beta_q}^2} \right)^{-\frac{1}{2}} \times \epsilon \\ & + \left(\tau_{(s)} \sum_{i=1}^{n} \lambda_{i,(s-1)} x_{iq}^2 + \frac{1}{\sigma_{\beta_q}^2} \right)^{-1} \left(\tau_{(s)} \sum_{i=1}^{n} \lambda_{i,(s-1)} x_{iq} \tilde{y}_q + \frac{\mu_{\beta_q}}{\sigma_{\beta_q}^2} \right) \end{aligned} \tag{4.26}$$

where ϵ is a random standard Normal value, and μ_{β_q} and $\sigma_{\beta_q}^2$ are the conjugate Normal prior mean and variance for β_q respectively, as before. For simplicity and diffusion, $\mu_{\beta_q} = 0$ and $\sigma^2_{\beta_q} = 10000$.

Then calculate $\hat{y}_{i,(s)}$ and draw the $\tau_{(s)}$ from the conditional Gamma distribution given above in Equation (4.24), namely,

$$\Gamma\left\{\frac{n}{2} + a_\tau, \left(\frac{1}{2}\sum_{i=1}^{n} \lambda_{i,(s-1)}\hat{y}^2_{i,(s)} + 2b_\tau\right)^{-1}\right\},$$

where the priors are chosen as $a_\tau = b_{tau} = 0.1$.

Similarly, each λ_i is drawn from the Gamma distribution given above in Equation (4.25), namely:

$$\Gamma\left\{\frac{1}{2}\left(\nu_{(s-1)} + 1\right), \left(\frac{1}{2}\lambda_{i,(s)}\tau_{(s)}\hat{y}^2_{i,(s)} + \nu_{(s-1)}\right)^{-1}\right\}.$$

Finally, $\nu_{(s)}$ drawn from the Metropolis algorithm presented above, conditional on the new $\lambda_{i,(s)}, i = 1, \ldots, n$.

In order to fit the t-distribution regression model to a set of data, this research allows the Gibbs sampler described above to generate several values of each parameter after a particular burn-in period. The mean of the draws for a parameter is then used as the parameter estimate. This method is followed when the fitting procedure is tested later in this section.

Simulating from the predictive posterior distribution

The Gibbs sampler described above allows one to draw a single set of parameters that is used to generate a response prediction based on a new set of observed covariate values. Making a single draw of each parameter in the model for the data, and then drawing with error, one effectively draws from the predictive posterior of the data.

This is the procedure that is followed within SRMI: the t-distribution regression model is fitted to the observed data, the Gibbs sampler (eventually, after burn-in) provides a single draw of each of the parameters from the approximate joint posterior of the parameters, and then the parameter set is used to generate a prediction (with error) for the responses that are missing (but whose covariates are complete).

4.2.2 Fitting the skew *t*-distribution

In order to fit the skew t-distribution, this thesis follows the fitting procedure outlined by Sahu et al. (2003).

The conditional distributions of the parameters

Although the t-distribution constitutes a step towards a robust imputation procedure, of even greater importance is the fact that skewness can easily be built into the t-distribution. The Gibbs sampling procedure that is required to estimate skewness within the data (given the data is t-distributed) is only marginally more complicated than the sampler for the symmetric t.

The likelihood for the symmetric distribution given in Equation (4.3) is altered slightly to incorporate a skewness parameter, δ, as follows:

$$L \propto \prod_{i=1}^{n} \left(\frac{\lambda_i \tau}{2\pi}\right)^{\frac{1}{2}} \exp\left[-\frac{\tau}{2}\left(y_i - \beta x_i - \delta z_i\right)^2\right] \times \prod_{i=1}^{n} \left[\frac{(\nu/2)^{\nu/2}}{\Gamma\left(\nu/2\right)} \lambda_i^{\nu/2-1} \exp\left(\frac{-\nu\lambda_i}{2}\right)\right] \tag{4.27}$$

where the other parameters are as before, except that we now have z_i as zero-truncated positive Normal values that show the influence of the skewness, δ, on each observation.

Skewness confounds several of the previous equations, so for the sake of completeness, all of the conditional posterior distributions used in the Gibbs sampler will be given below. Now, for each observation $i, i = 1, \ldots, n$, and for each covariate $q, q = 0, 1 \ldots, p$, $\tilde{y}_{iq} = y_i - \beta_{-q} X_{-q} - \delta z_i$, where $-q$ represents all variables in X besides variable q. In other words, for $q = 0$:

$$\tilde{y}_{i0} = y_i - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_p x_{ip} - \delta z_i \tag{4.28}$$

For $q = 1$:

$$\tilde{y}_{i1} = y_i - \beta_0 - \beta_2 x_{i2} - \beta_3 x_{i3} - \ldots - \beta_p x_{ip} - \delta z_i \tag{4.29}$$

For $q = 2, \ldots, p$:

$$\tilde{y}_{iq} = y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_{q-1} x_{i(q-1)} - \beta_{q+1} x_{i(q+1)} - \ldots - \beta_p x_{ip} - \delta z_i \tag{4.30}$$

Finally, for $q = p$:

$$\tilde{y}_{ip} = y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_{p-1} x_{i(p-1)} - \delta z_i \tag{4.31}$$

We also define $\tilde{\tilde{y}}_i = y_i - \beta x_i - \delta z_i$ separate from $\hat{y}_i = y_i - \beta x_i$.

With skewness incorporated into the $\tilde{y}_{iq}$, the same conditional distributions exist for the β_q:

$$\beta_q | y, \beta_{-q}, \tau, \Lambda \sim$$
$$N \left\{ \left(\tau \sum_{i=1}^{n} \lambda_i x_{iq}^2 + \frac{1}{\sigma_{\beta_q}^2} \right)^{-1} \left(\tau \sum_{i=1}^{n} \lambda_i x_{iq} \tilde{y}_{iq} + \frac{\mu_{\beta_q}}{\sigma_{\beta_q}^2} \right), \left(\tau \sum_{i=1}^{n} \lambda_i x_{iq}^2 + \frac{1}{\sigma_{\beta_q}^2} \right)^{-1} \right\}$$

where μ_{β_q} and $\sigma_{\beta_q}^2$ are the conjugate Normal prior mean and variance for β_q respectively. Once again, $\mu_{\beta_q} = 0$ and $\sigma_{\beta_q}^2 = 10000$.

For τ, we have that:

$$\tau | y, \beta, \Lambda \sim \Gamma \left\{ \frac{n}{2} + a_\tau, \left(\frac{1}{2} \sum_{i=1}^{n} \lambda_i \tilde{\tilde{y}}_i^2 + 2b_\tau \right)^{-1} \right\} \tag{4.32}$$

where a_τ and b_τ are the conjugate Gamma prior parameters for τ.

The conditional posterior for the $z_i, i = 1, \ldots, n$ is derived to be:

$$z_i | y, \beta, \tau, \delta, \Lambda \sim N \left\{ \left(\tau \lambda_i \delta^2 + 1 \right)^{-1} \tau \lambda_i \delta \hat{y}_i, \left(\tau \lambda_i \delta^2 + 1 \right)^{-1} \right\} I \left(Z_i > 0 \right), \tag{4.33}$$

where $I\left(Z_i > 0\right)$ is an indicator function to ensure that only positive z_i exist (in order to make sense of the sign of the skewness parameter δ).

The conditional posterior distribution of the skewness parameter, δ, can be shown to be:

$$\begin{aligned}&\delta|y,\beta,\tau,\Lambda,z_1,\ldots,z_n\sim\\&N\left\{\left(\tau\sum_{i=1}^{n}\lambda_i z_i^2+\frac{1}{\sigma_\delta^2}\right)^{-1}\left(\tau\sum_{i=1}^{n}\lambda_i z_i\hat{y}_i+\frac{\mu_\delta}{\sigma_\delta^2}\right),\left(\tau\sum_{i=1}^{n}\lambda_i z_i^2+\frac{1}{\sigma_\delta^2}\right)^{-1}\right\},\end{aligned}\tag{4.34}$$

where μ_δ and σ_δ^2 are the conjugate Normal prior parameters for δ.

For the λ_i, it can be shown that

$$\lambda_i|y,\beta,\tau,\nu,\delta,z_1,\ldots,z_n\sim\Gamma\left\{\frac{1}{2}(\nu+1),\left[\frac{1}{2}\left(\tau\tilde{\hat{y}}_i^2+\nu\right)\right]^{-1}\right\},\tag{4.35}$$

with the skewness incorporated in the distribution by replacing $\hat{y}_i$ with $\tilde{\hat{y}}_i$.

The posterior for ν, conditional on Λ, and its priors, are exactly the same as in Equation (4.6).

The algorithm for the Gibbs sampler (and Metropolis sampler for ν) when we wish to incorporate skewness into the imputation model, is based on the conditional distributions listed above.

Programmable algorithm for fitting the skew t-distribution

The algorithm explained in the previous subsection lays the groundwork for simulation from a skew t-distribution. As for the symmetric t-distribution, we initialise the parameters, and then update multiple times via the conditional distributions.

1. Initialise the following parameters:

$$\begin{aligned}
\tau_{(1)} &= \frac{1}{var\,(y)} \\
\Lambda_{(1)} &= 0.5 I_{n\times n} \\
\beta_{(1)} &= \sigma^2 \left(X'\Lambda X\right)^{-\frac{1}{2}} \times \epsilon + \left(X'\Lambda X\right)^{-1}\left(X'\Lambda y\right) \\
z_{i,(1)} &= -1, i = 1, \ldots, n \\
\delta_{(1)} &= 0
\end{aligned}$$

where ϵ is a $((p+1)\times 1)$ vector of standard Normal draws, X is the observed covariate matrix, y is the vector of observed responses, and $I_{n\times n}$ is the $n \times n$ identity matrix.

2. Draw an initial $\nu_{(1)}$ in exactly the same manner as before, using the Metropolis sampler described above.

3. Now the parameters $\tau, \Lambda, \beta, z_i, \delta$ and ν have initial values $\tau_{(1)}, \Lambda_{(1)}, \beta_{(1)}, z_{i,(1)}, \delta_{(1)}$, and $\nu_{(1)}$. Start the updating process within the Gibbs sampler, for $s = 2, 3, \ldots, m$, where m is the number of update rounds including burn-in and the number of draws required from the conditional posterior distributions. First in each step, recalculate the $\tilde{y}_{iq,(s)}$ using $\beta_{-q,(s-1)}$. As in the symmetric case, for each $\beta_q, q = 0, 1, \ldots, p$,

$$\begin{aligned}
\beta_{q,(s)} = {} & \left(\tau_{(s)} \sum_{i=1}^{n} \lambda_{i,(s-1)} x_{iq}^2 + \frac{1}{\sigma_{\beta_q}^2}\right)^{-\frac{1}{2}} \times \epsilon \\
& + \left(\tau_{(s)} \sum_{i=1}^{n} \lambda_{i,(s-1)} x_{iq}^2 + \frac{1}{\sigma_{\beta_q}^2}\right)^{-1} \left(\tau_{(s)} \sum_{i=1}^{n} \lambda_{i,(s-1)} x_{iq} \tilde{y}_q + \frac{\mu_{\beta_q}}{\sigma_{\beta_q}^2}\right) \qquad (4.36)
\end{aligned}$$

where ϵ is a random standard Normal value, and μ_{β_q} and $\sigma_{\beta_q}^2$ are as before, the conjugate Normal prior mean and variance for β_q respectively, again chosen to be 0 and 10000, respectively. The skewness is incorporated into this process through the skewness-adjusted $\tilde{y}_{iq,(s)}$.

Then calculate $\hat{y}_{i,(s)}$ and draw the $\tau_{(s)}$ from the conditional gamma distribution given above in Equation 4.24, namely,

$$\Gamma\left\{\frac{n}{2}+a_{\tau},\left(\frac{1}{2}\sum_{i=1}^{n}\lambda_{i,(s-1)}\hat{y}_{i,(s)}^{2}+2b_{\tau}\right)^{-1}\right\},$$

where the priors are chosen again as $a_{\tau}=b_{\tau}=0.1$.

For the Normal z_i values we simply use the following simulation for each i:

$$\begin{aligned} z_{i,(s)} &= \left(\tau_{(s)}\lambda_{i,(s-1)}\delta_{(s-1)}^{2}+1\right)^{-\frac{1}{2}}\times\epsilon \\ &+\left(\tau_{(s)}\lambda_{i,(s-1)}\delta_{(s-1)}^{2}+1\right)^{-1}\tau_{(s)}\lambda_{i,(s-1)}\delta_{(s-1)}\hat{y}_{i,(s)}, \end{aligned} \tag{4.37}$$

where ϵ is a standard Normal draw.

For δ we have another Normal simulation, as follows:

$$\begin{aligned} \delta_{(s)} &= \left(\tau_{(s)}\sum_{i=1}^{n}\lambda_{i,(s-1)}z_{i,(s)}^{2}+\frac{1}{\sigma_{\delta}^{2}}\right)^{-\frac{1}{2}}\times\epsilon \\ &+\left(\tau_{(s)}\sum_{i=1}^{n}\lambda_{i,(s-1)}z_{i,(s)}^{2}+\frac{1}{\sigma_{\delta}^{2}}\right)^{-1}\left(\tau_{(s)}\sum_{i=1}^{n}\lambda_{i,(s)}z_{i,(s)}\hat{y}_{i,(s)}+\frac{\mu_{\delta}}{\sigma_{\delta}^{2}}\right), \end{aligned} \tag{4.38}$$

where ϵ is a standard Normal draw and the prior mean and variance are chosen such that $\mu_{\delta}=0$ and $\sigma_{\delta}=1000$.

Each λ_i is then drawn from the gamma distribution given above in Equation 4.25, with skewness incorporated into $\tilde{\tilde{y}}_i$, namely:

$$\Gamma\left\{\frac{1}{2}\left(\nu_{(s-1)}+1\right),\left(\frac{1}{2}\lambda_{i,(s)}\tau_{(s)}\tilde{\tilde{y}}_{i,(s)}^{2}+\nu_{(s-1)}\right)^{-1}\right\}.$$

Finally, $\nu_{(s)}$ drawn from the Metropolis algorithm explained above, conditional on the new $\lambda_{i,(s)}, i=1,\ldots,n$.

.

4.3 Methodology of the Simulation Study

The simulation study presented in this chapter is an analysis of the robustness of a mis-specified sequential imputation method based on the t-distribution (and its skew specification), as a continuation of the work presented by He & Raghunathan (2009). The rationale behind this study is to find a robust Normal-family imputation method that remains strong in the presence of non-Normal data. In this way, imputers with less expertise will have a robust SRMI model to use under circumstances when the data may not be Normally distributed, thereby figuratively extending the reach of SRMI to those who might not have thought of using this process for their MI procedures.

The simulation study will evaluate the situations where predictive mean matching (PMM), local residual draw (LRD), and expanded residual draw (ERD) can reduce bias in SRMI procedures. While these adaptations have already been tested for Normality-based SRMI by He & Raghunathan (2009), it is of interest to see if they are useful when the symmetric t-distribution is used in SRMI, and to see if these adaptations can compare in effectiveness to the actual skew specification of the t-distribution.

4.3.1 Assessment Methods

For an imputation method to be robust, the model should replicate the original data and predict plausible imputations. The purpose of this chapter is to use the robust SRMI model to replicate the original simulated data after it has been made incomplete. The overall analysis of multiple completed data sets is unnecessary, so this chapter will refrain from running these post-imputation analyses and computing $RBIAS$ and $RRMSE$ as was done by He & Raghunathan (2009). The only results that need to be assessed are the fit of the completed data to the original data and the fit of the imputation draws to the values that were made missing.[3] This assessment requires the construction of two quantile-quantile (QQ) plots, and a statistic to measure the deviance of these plots from the optimal solution.

[3] For an imputation method to be robust, the model should replicate the original data and predict plausible imputations.

1. *Firstly*, for one data scenario (with $n = 200$) and one MDM, a plot of the quantiles of the completed data (for each variable with missingness) is drawn against the quantiles of the original data (for each corresponding original complete variable). Since MI creates multiple completed data sets and the overall analyses after MI are averaged over these multiple completed data sets to obtain a final estimate, a 'pooling' procedure is followed when calculating the quantiles of the completed data — the five MI completed data sets for a particular MI method are pooled before the quantiles are calculated.

   For each variable with missingness, the mean squared error (MSE) of the deviation of the quantiles of the completed data from the quantiles of the original data is then computed. Additionally, the MSE of the QQ plot for the incomplete (INC) data and complete-case (CC) or case-deleted data is calculated for comparison. Across multiple simulations within a data scenario and MDM, a distribution of QQ plot MSE calculations is then obtained. The average of these MSE calculations for an imputation method is reported for each data scenario and MDM combination.

   This assessment allows one to compare post-imputation distributions with the original data distributions, as well as with the distributions under the incomplete data and the data set where incomplete observations are deleted.

2. *Secondly*, for each data scenario (with $n = 200$) and MDM combination, 200 multiply imputed data sets are created under each SRMI model. For each variable with missingness, the 1%, 2%, . . . , 99% equal-tail coverage intervals of the imputed values are calculated. The proportions of the original data points that fall inside their 1%, 2%, . . . , 99% imputed intervals is then determined. For an imputation method that perfectly replicates the original data, one should find that, for one variable with missingness, p% of the original data points that were made missing should fall within the p% imputation intervals for that data point. The MSE of the QQ plot of these coverage intervals from the 45° line is reported.

   This assessment allows one to make sure that the imputation model is predicting individual data points within expected intervals.

4.3.2 Simulated data

Complete data is generated under four different data scenarios. The data is then made incomplete using alternating MCAR and MAR mechanisms, and re-filled using various SRMI models, namely the Normal and t, with their PMM, LRD and ERD adaptations for skewness, as well as the skew t model.

Data Scenarios

In this study, simulated data consists of four variables, Y_1, Y_2, Y_3, and Y_4, where: $Y_1 = \epsilon_1$, $Y_2 = 1 + Y_1 + \epsilon_2$, $Y_3 = 1 + Y_1 + Y_2 + \epsilon_3$, $Y_4 = 1 + Y_1 + Y_2 + Y_3 + \epsilon_4$.

The complete-data models take one of four forms:

1. **Normality (and symmetry)**:

$$\epsilon_j \sim N\,(0,1) \text{ for } j = 1,2,3,4.$$

2. **Moderate tails, with skewness**:

$$\begin{aligned}
&\epsilon_1 \sim N\,(0,1)\\
&\epsilon_2 \sim t_6\\
&\epsilon_3 = \alpha_3 - \omega_3 \text{ where } \alpha_3 \sim t_6 \text{ and } \omega_3 \sim N\,(0,1)\\
&\epsilon_4 = \alpha_4 - 2\omega_4 \text{ where } \alpha_4 \sim t_6 \text{ and } \omega_4 \sim N\,(0,1)
\end{aligned}$$

3. **Heavy tails, with skewness**:

$$\begin{aligned}
&\epsilon_1 \sim N\,(0,1)\\
&\epsilon_2 \sim t_3\\
&\epsilon_3 = \alpha_3 - \omega_3 \text{ where } \alpha_3 \sim t_3 \text{ and } \omega_3 \sim N\,(0,1)\\
&\epsilon_4 = \alpha_4 - 2\omega_4 \text{ where } \alpha_4 \sim t_3 \text{ and } \omega_4 \sim N\,(0,1)
\end{aligned}$$

4. **Mixed *gh* distributions**: Again $\epsilon_1 \sim N\,(0,1)$. For the remaining error distribu-

tions of ϵ_2, ϵ_3, and ϵ_4, various options of Tukey's gh distribution are chosen, as was done by He & Raghunathan (2006) and reviewed in Subsection 3.4.5. For more information on the forms of the gh distribution that were chosen, see He & Raghunathan (2006).

For all errors in data scenario 4, $\mu = 0$ and $\sigma = 1$. However, the g and h parameters are varied as follows:

- For ϵ_2, $g = 1$ and $h = -0.25$. This creates a downward-sloping monotonic exponential-type distribution.
- For ϵ_3, $g = 0.75$ and $h = 0.25$. This generates a right-skewed distribution.
- For ϵ_4, $g = 1$ and $h = 0$. This is the well-known Lognormal distribution.

5. **Extreme deviation from Normality**: In this data scenario, the method of He & Raghunathan (2009) is followed roughly, but only one of the 36 scenarios generated by these authors is explored — one with extreme deviation from Normality, and large error variances. Additionally, another variable is added, with a skew t_3 error. The algorithm followed is built according to the method of He & Raghunathan (2009).

   Let the vector of errors for Y_j, $j = 1, \ldots, 4$ be $\xi_j = [\epsilon_{1j}\ \epsilon_{2j}\ \epsilon_{3j}\ \ldots\ \epsilon_{nj}]'$. Also, let $U_j = [u_{1j}\ u_{2j}\ u_{3j}\ \ldots\ u_{nj}]'$. For each observation i, $i = 1, \ldots, n$, the error ϵ_{ij} for this observation on each variable is constructed as follows:

   - $\xi_1 \sim N(0,1)$
   - $\epsilon_{i2} = \frac{u_{i2}-E(U_2)}{Var(U_2)} \times \sqrt{3Var(Y_1)}$, $u_2 = 1 + \exp(1+Z)$, $Z \sim N(0,1)$. So ξ_2 is a vector of centred and widely scaled Lognormal errors.
   - $\epsilon_{i3} = \frac{u_{i3}-E(U_3)}{Var(U_3)} \times \sqrt{2[Var(Y_1) + Var(Y_2)]}$, $u_3 = W$, $W \sim t_3$. So ξ_3 is a vector centred and widely scaled t_3 errors.
   - $\epsilon_{i4} = \frac{u_{i4}-E(U_4)}{Var(U_4)} \times \sqrt{Var(Y_1) + Var(Y_2) + +Var(Y_2)}$, $u_4 = W - 2Z$, $W \sim t_3$, $Z \sim N(0,1)$. So ξ_4 is a vector centred and widely scaled right-skewed t_3 errors.

   This arrangement of error distributions is arguably the most extremely deviated from Normality of the five data scenarios.

Missing Data Mechanisms

In this study, two MDMs are simulated, namely one MCAR mechanism one MAR mechanism. For the MCAR mechanism, every data point has a 20% chance of being deleted in one simulation. This does not guarantee 20% missingness, but, since the MAR mechanism does not either, this point is moot.

For the MAR mechanism, a logistic regression is set up, variable by variable, to generate a probability for each observation in the current variable to be missing.

- Y_1 is complete.
- The probability that observation i is missing in Y_2 is:

  $$p_{i,2} = 0.4 \left[1 + \exp\left(-0.3 - 0.3y_{i,1}\right)\right]^{-1}$$

  Once these probabilities are calculated, each observation with probability less than an independent draw from a $U(0,1)$ distribution is made missing.

- The probability that observation i is missing in Y_3 is:

  $$p_{i,3} = 0.4 \left[1 + \exp\left(-0.3 - 0.3y_{i,1} - 0.3y_{i,2}\right)\right]^{-1}$$

  If $y_{i,2}$ is already missing, this term is ignored, ensuring the MAR MDM does not become an MNAR MDM. Data points are made missing in the same way as for Y_2.

- Finally, the probability that observation i is missing in Y_4 is:

  $$p_{i,3} = 0.4 \left[1 + \exp\left(-0.3 - 0.3y_{i,1} - 0.3y_{i,2} - 0.3y_{i,3}\right)\right]^{-1}$$

  If $y_{i,2}$ or $y_{i,3}$ or both are already missing, the missing terms are ignored, ensuring once more that the MAR MDM does not become an MNAR MDM. Data points are made missing in the same way as for Y_2 and Y_3.

The 0.4 coefficient in the above equations ensures that missingness is comparable to the MCAR MDM. Across 100 simulations of these MDMs we have average missingness for the two MDMs as given in Table 4.1.[4]

Table 4.1: Missingness from the MCAR and MAR MDMs

MDM	**Variable/CC**	**Data scenario**				
		1	2	3	4	5
MCAR	Y2	20.1%	20.0%	20.4%	20.2%	19.6%
	Y3	19.7%	19.4%	20.3%	20.0%	20.0%
	Y4	20.1%	19.6%	20.4%	19.5%	19.7%
	CC	48.7%	48.1%	49.6%	48.7%	48.5%
MAR	Y2	16.9%	17.3%	17.0%	17.3%	17.5%
	Y3	16.9%	16.8%	17.0%	17.1%	17.4%
	Y4	17.4%	17.3%	17.6%	16.9%	17.3%
	CC	42.7%	43.1%	42.9%	42.8%	43.4%

To ensure that the MDMs are indeed MCAR and MAR, we examine the difference between the mean of the original complete data and the mean of the incomplete data. The results are given in Table 4.2.[5]

Table 4.2: Difference after data is made missing

Data scenario 1					**Data scenario 2**				
MDM	**Y1**	**Y2**	**Y3**	**Y4**	**MDM**	**Y1**	**Y2**	**Y3**	**Y4**
MCAR		0.00	0.00	0.01	MCAR		-0.01	-0.01	0.01
MCAR CC	0.00	0.00	0.00	-0.01	MCAR CC	-0.01	0.00	-0.01	-0.04
MAR		-0.03	-0.11	-0.33	MAR		-0.04	-0.15	-0.39
MAR CC	-0.16	-0.20	-0.36	-0.73	MAR CC	-0.16	-0.26	-0.46	-0.91
Data scenario 3					**Data scenario 4**				
MDM	**Y1**	**Y2**	**Y3**	**Y4**	**MDM**	**Y1**	**Y2**	**Y3**	**Y4**
MCAR		0.00	-0.01	-0.02	MCAR		-0.01	0.00	-0.03
MCAR CC	0.00	0.01	0.00	0.00	MCAR CC	-0.02	-0.01	-0.03	-0.06
MAR		-0.04	-0.14	-0.40	MAR		-0.03	-0.13	-0.34
MAR CC	-0.16	-0.25	-0.47	-0.91	MAR CC	-0.19	-0.21	-0.40	-0.80
Data scenario 5									
MDM	**Y1**	**Y2**	**Y3**	**Y4**					
MCAR		0.01	-0.01	0.10					
MCAR CC	0.00	0.01	0.01	0.10					
MAR		-0.05	-0.16	-0.13					
MAR CC	-0.15	-0.25	-0.48	-0.49					

From Table 4.2 we can see that the MCAR MDM is making no difference to the mean of the data, while the MAR MDM results show that the mean of the incomplete data is higher than that of the original complete data. This suggests that the MAR MDM is successfully weeding out smaller values in the data set.

[4] If $n = 1000$ observations are simulated, then there is no discernible difference in the missingness figures.

[5] If $n = 1000$ observations are simulated the results are very similar; no marked differences are notable.

4.4 Simulation Study Analysis

4.4.1 Distributional coverage

Tables 4.3 to 4.7 present the MSE of the QQ plots comparing the quantiles of the incomplete and completed data with the quantiles of the original data. Lower numbers are more desirable. The best result for each variable (per data scenario and MDM) is highlighted in bold, while methods with MSEs within 5% of the best method are italicised.

Table 4.3: QQ MSE for Data Scenario 1

Var.:	Y_2				Y_3				Y_4			
MDM:	**MCAR**		**MAR**		**MCAR**		**MAR**		**MCAR**		**MAR**	
Sample:	200	1000	200	1000	200	1000	200	1000	200	1000	200	1000
INC	0.006	0.002	0.007	0.002	0.027	0.005	0.037	0.023	0.085	0.017	0.190	0.153
CC	0.022	0.006	0.061	0.057	0.079	0.018	0.203	0.206	0.259	0.064	0.771	0.746
N	*0.003*	**0.001**	*0.003*	*0.001*	**0.009**	*0.002*	0.009	*0.002*	*0.024*	**0.004**	**0.019**	*0.004*
N_{PMM}	0.004	0.002	0.004	0.001	0.018	0.004	0.018	0.003	0.046	0.018	0.046	0.021
N_{LRD}	0.003	*0.001*	0.003	*0.001*	0.010	*0.002*	0.010	*0.002*	0.025	0.005	0.022	*0.004*
N_{ERD}	**0.003**	*0.001*	**0.003**	**0.001**	*0.009*	*0.002*	**0.008**	*0.002*	*0.024*	*0.004*	*0.019*	*0.004*
t	*0.003*	*0.001*	0.003	*0.001*	*0.009*	0.002	0.009	*0.002*	*0.025*	*0.004*	0.020	*0.004*
t_{PMM}	0.004	0.002	0.004	0.001	0.018	0.004	0.018	0.003	0.045	0.018	0.049	0.021
t_{LRD}	0.004	*0.001*	0.003	*0.001*	0.010	**0.002**	0.009	**0.002**	0.026	0.005	0.022	0.004
t_{ERD}	0.006	0.003	0.005	0.002	0.010	0.003	0.010	0.003	**0.024**	0.005	0.020	0.004
t_{skew}	*0.003*	*0.001*	*0.003*	*0.001*	*0.009*	*0.002*	*0.009*	*0.002*	*0.024*	*0.004*	0.021	**0.004**

Table 4.4: QQ MSE for Data Scenario 2

Var.:	Y_2				Y_3				Y_4			
MDM:	**MCAR**		**MAR**		**MCAR**		**MAR**		**MCAR**		**MAR**	
Sample:	200	1000	200	1000	200	1000	200	1000	200	1000	200	1000
INC	0.015	0.002	0.016	0.003	0.029	0.007	0.047	0.026	0.133	0.025	0.283	0.154
CC	0.074	0.007	0.146	0.061	0.113	0.023	0.330	0.206	0.474	0.083	1.284	0.756
N	0.013	**0.001**	0.010	*0.001*	*0.016*	*0.004*	0.014	0.004	*0.066*	**0.012**	*0.065*	0.011
N_{PMM}	0.013	0.002	0.013	0.002	0.022	0.005	0.017	0.005	0.097	0.022	0.105	0.019
N_{LRD}	0.011	0.002	0.010	0.001	0.017	*0.004*	0.015	0.004	0.076	0.013	0.072	0.012
N_{ERD}	0.012	*0.001*	0.011	*0.001*	**0.015**	0.004	**0.013**	0.003	**0.065**	0.013	**0.063**	0.011
t	**0.010**	0.002	**0.009**	*0.001*	0.017	*0.004*	*0.014*	0.004	0.070	*0.012*	*0.064*	0.011
t_{PMM}	0.013	0.002	0.012	0.002	0.021	0.005	0.016	0.005	0.096	0.021	0.104	0.019
t_{LRD}	0.011	0.002	0.010	0.001	0.017	0.004	0.015	0.003	0.075	0.014	0.075	0.011
t_{ERD}	0.011	0.002	0.011	0.001	*0.016*	**0.004**	*0.014*	**0.003**	0.073	0.021	0.071	0.017
t_{skew}	*0.010*	0.002	*0.009*	**0.001**	0.017	0.004	0.014	0.004	*0.067*	*0.012*	*0.065*	**0.010**

It is clear from the Tables that under both MCAR and MAR MDMs, complete case data have distributions that deviate the most from the original data, although this difference is somewhat muted under the MCAR mechanism (as expected).

Under the assumption of Normal errors in Data Scenario 1, Normal, Normal with ERD, and the skew t imputation models perform the best under both the MCAR and MAR MDMs.

Table 4.5: QQ MSE for Data Scenario 3

Var.:	Y_2				Y_3				Y_4			
MDM:	**MCAR**		**MAR**		**MCAR**		**MAR**		**MCAR**		**MAR**	
Sample:	200	1000	200	1000	200	1000	200	1000	200	1000	200	1000
INC	0.015	0.004	0.014	0.004	0.037	0.008	0.060	0.024	0.154	0.030	0.307	0.170
CC	0.053	0.017	0.106	0.074	0.143	0.032	0.373	0.210	0.464	0.113	1.242	0.780
N	*0.010*	0.013	*0.009*	0.011	*0.028*	0.006	*0.021*	0.005	*0.077*	0.015	*0.063*	*0.012*
N_{PMM}	0.013	0.005	0.012	0.004	0.036	0.009	0.029	0.009	0.116	0.037	0.110	0.044
N_{LRD}	0.011	*0.003*	*0.010*	**0.002**	0.028	0.007	0.022	0.005	0.082	0.015	0.068	0.013
N_{ERD}	**0.010**	0.011	**0.009**	0.009	**0.027**	0.006	**0.020**	*0.004*	0.079	**0.014**	**0.060**	0.012
t	0.011	0.003	*0.009*	*0.002*	*0.027*	*0.005*	0.022	**0.004**	**0.075**	0.015	*0.061*	*0.012*
t_{PMM}	0.013	0.005	0.011	0.003	0.037	0.009	0.030	0.009	0.118	0.037	0.110	0.043
t_{LRD}	0.012	**0.003**	0.010	0.002	0.029	0.007	0.023	0.005	0.082	0.016	0.069	0.013
t_{ERD}	*0.011*	0.004	*0.009*	0.003	*0.028*	0.007	0.023	0.006	0.080	0.023	0.065	0.023
t_{skew}	0.011	*0.003*	*0.010*	*0.002*	*0.028*	**0.005**	0.022	*0.004*	0.083	*0.015*	*0.062*	**0.012**

Table 4.6: QQ MSE for Data Scenario 4

Var.:	Y_2				Y_3				Y_4			
MDM:	**MCAR**		**MAR**		**MCAR**		**MAR**		**MCAR**		**MAR**	
Sample:	200	1000	200	1000	200	1000	200	1000	200	1000	200	1000
INC	0.005	0.001	0.005	0.002	0.017	0.003	0.034	0.018	0.064	0.012	0.185	0.118
CC	0.016	0.003	0.058	0.040	0.057	0.011	0.214	0.154	0.221	0.042	0.848	0.615
N	*0.001*	**0.000**	**0.001**	*0.000*	**0.003**	*0.001*	*0.003*	*0.001*	*0.010*	*0.002*	0.009	*0.002*
N_{PMM}	0.003	0.001	0.002	0.001	0.008	0.003	0.006	0.002	0.030	0.011	0.025	0.009
N_{LRD}	0.001	0.000	0.001	0.000	0.003	0.001	0.003	0.001	0.010	*0.002*	0.009	0.002
N_{ERD}	**0.001**	*0.000*	*0.001*	*0.000*	*0.003*	**0.001**	*0.003*	*0.001*	**0.010**	*0.002*	**0.008**	*0.002*
t	0.001	0.000	0.001	**0.000**	0.003	*0.001*	*0.003*	*0.001*	*0.010*	**0.002**	*0.009*	**0.001**
t_{PMM}	0.002	0.001	0.002	0.001	0.008	0.003	0.006	0.002	0.030	0.011	0.024	0.009
t_{LRD}	0.001	0.000	0.001	0.000	0.003	0.001	0.003	0.001	0.010	0.002	0.009	0.002
t_{ERD}	0.004	0.003	0.004	0.002	0.006	0.002	0.005	0.002	0.014	0.003	0.012	0.002
t_{skew}	0.001	0.000	0.001	*0.000*	*0.003*	*0.001*	**0.003**	**0.001**	*0.010*	*0.002*	0.009	*0.002*

In Data Scenario 2, with moderate t errors, the Normal model with ERD performs well, but only when $n = 200$. The Normal, t and skew t models perform well under both MDMs, both choices of N, and for all incomplete variables.

Under Data Scenario 3, the Normal model and Normal model with ERD generally perform well when $n = 200$. The t and skew t imputation models consistently perform adequately.

Looking at Data Scenario 4, several imputation models perform well, including the Normal and the Normal with ERD, the t, and the skew t. The LRD adaptations of both the Normal and the t models are also adequate. These results all hold for both sample sizes.

Data Scenario 5 holds mixed results. Strangely, no model is able to replicate the distribution of the original data's Y_2 better than the incomplete data. The error on this variable was Lognormal, so further investigation into this scenario might be warranted in future research.

Amongst the imputation models, the best performers for replicating the original distribution of Y_2 are the Normal and t models with the PMM and LRD adaptations. For Y_3, the

Table 4.7: QQ MSE for Data Scenario 5

Var.:	Y_2				Y_3				Y_4			
MDM:	**MCAR**		**MAR**		**MCAR**		**MAR**		**MCAR**		**MAR**	
Sample:	200	1000	200	1000	200	1000	200	1000	200	1000	200	1000
INC	**0.028**	**0.010**	**0.020**	**0.009**	0.080	*0.016*	0.114	0.029	**1.476**	0.423	**1.232**	0.632
CC	0.087	0.034	0.141	0.123	0.357	0.064	0.513	0.274	13.450	1.346	8.227	1.809
N	0.042	0.034	0.030	0.025	0.093	0.023	0.095	0.015	1.989	0.632	1.630	0.569
N_{PMM}	0.032	0.013	*0.020*	0.011	0.094	0.021	0.109	0.014	2.766	0.495	1.739	0.454
N_{LRD}	0.030	0.010	*0.020*	0.009	0.086	**0.015**	0.092	*0.012*	2.026	0.489	2.419	0.466
N_{ERD}	0.032	0.014	0.021	0.011	0.078	0.017	0.091	0.013	1.592	**0.422**	1.446	**0.431**
t	0.045	0.035	0.029	0.024	**0.066**	0.021	*0.083*	0.013	1.739	0.453	1.415	0.453
t_{PMM}	0.036	*0.010*	0.021	*0.009*	0.093	0.021	0.112	0.013	3.005	0.511	1.764	0.473
t_{LRD}	0.036	0.011	*0.021*	*0.009*	0.095	0.018	0.107	**0.012**	2.094	0.481	2.479	0.489
t_{ERD}	0.033	0.027	0.021	0.018	0.121	0.050	0.157	0.040	6.667	6.489	5.470	5.188
t_{skew}	0.045	0.035	0.027	0.023	0.075	0.020	**0.082**	0.013	1.666	*0.426*	1.502	*0.450*

t and skew t models perform well, together with the Normal and t models incorporating the LRD adjustment. For Y_4, once more the incomplete data seems close in distribution to the original data, while the Normal model with ERD, the t model, and the skew t model perform best amongst the imputation models.

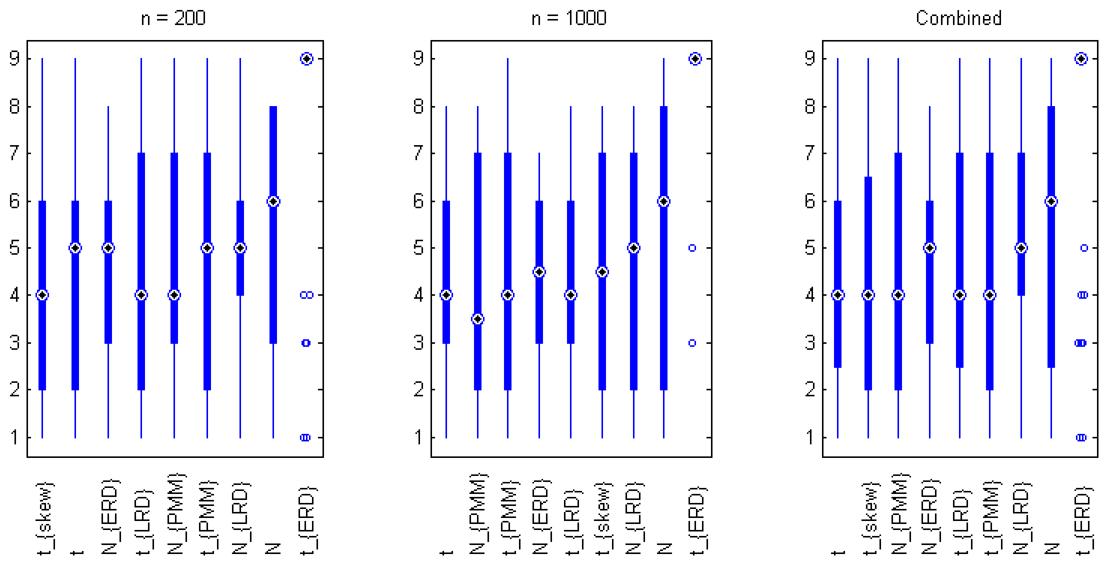
Across all models, it is clear that the skew t model is the most robust. This model shows fewer weaknesses than the other models, while always remaining relatively close to the performance of the best imputation model if it is not the best model itself. It can also be noted that when the distributional assumptions of the errors are less pronounced, *i.e.* when $n = 200$, it is more difficult to choose a better imputation model. However, with $n = 1000$, we are better able to gauge the effectiveness of some of the imputation models, for example, the skew t model.

Before continuing with the second analysis, it is important to keep in mind that the errors across the three variables are uncorrelated, and that a certain amount of 'averaging' of errors across an observation may or may not have allowed less robust models to appear better than they are. However, this property should not concern us too much, since it is clear that even with this advantage, the traditionally less robust models appear are still shown to be less robust that the skew t model.

4.4.2 Imputation coverage

Tables 4.8 to 4.12 provide the MSEs of the QQ plots that compare the coverage of the actual distribution of the imputations over the original values of the data that was made

Figure 4.1: Boxplots of MSE ranks of imputations, across variables, MDMs and data scenarios



missing. Once again, lower numbers are more desirable, and the best result for each variable (per data scenario and MDM) is highlighted in bold, while methods with MSEs within 5% of the best method are italicised. Take note that the maximum MSE for a method across the three variables is also given as a measure of the worst error the method made within the given data scenario and sample size.

The ranks of these MSEs are summarised in a boxplot, Figure 4.1, sorted by the mean of the ranks for each method.

While little information can be gained from this crude ranking analysis, one can at least note that the symmetric and skew t models are performing far better than the Normal model in general, and that the t model with ERD has poor imputation coverage intervals.

There is no systematic evidence of one imputation method providing more accurate coverage of the original data points before these were made missing. Clearly, however, the t model with the ERD adjustment is inadmissable as an imputation method. The errors that the algorithm has to 'donate' to the imputations are simply too wide, too often.

By examining averages across scenarios and MDMs, one finds that the Normal modelwith either PMM, LRD or ERD adjustments, along with the t models with PMM or LRD adjustments, all perform better than the unadjusted Normal, unadjusted t, and the skew

Table 4.8: QQ MSE for imputations under Data Scenario 1

	MDM:	**MCAR**			**MAR**			
N	**Variable:**	**Y2**	**Y3**	**Y4**	**Y2**	**Y3**	**Y4**	**MAX**
200	N	**7.3**	87.3	23.7	93.4	30.4	**13.5**	93.4
	N_{PMM}	67.6	43.4	37.7	**19.1**	21.0	112.5	112.5
	N_{LRD}	41.4	77.7	72.5	47.8	29.6	32.4	77.7
	N_{ERD}	9.7	77.9	67.9	73.3	27.5	17.4	77.9
	t	8.1	65.1	46.1	75.3	20.3	26.8	75.3
	t_{PMM}	66.6	**13.2**	48.8	44.5	29.8	157.8	157.8
	t_{LRD}	35.2	88.4	69.8	56.0	**17.7**	24.2	88.4
	t_{ERD}	349.3	58.2	**22.0**	484.0	123.8	20.7	484.0
	t_{skew}	10.5	66.8	41.8	71.1	23.8	19.1	**71.1**
1000	N	5.4	3.4	6.1	4.7	*5.9*	*11.8*	11.8
	N_{PMM}	5.0	2.7	8.5	8.2	8.2	24.9	24.9
	N_{LRD}	5.8	4.5	7.8	2.8	12.8	15.7	15.7
	N_{ERD}	*4.6*	3.1	6.1	2.0	**5.8**	19.3	19.3
	t	9.3	2.8	**2.1**	2.6	11.0	**11.5**	**11.5**
	t_{PMM}	7.0	**2.1**	10.2	4.2	7.5	27.7	27.7
	t_{LRD}	5.9	6.7	5.8	**1.6**	13.7	12.2	13.7
	t_{ERD}	295.5	95.0	4.4	262.7	59.5	39.6	295.5
	t_{skew}	**4.5**	2.7	2.5	3.5	8.2	13.7	13.7

Table 4.9: QQ MSE for imputations under Data Scenario 2

	MDM:	**MCAR**			**MAR**			
N	**Variable:**	**Y2**	**Y3**	**Y4**	**Y2**	**Y3**	**Y4**	**MAX**
200	N	21.2	70.5	8.6	25.8	**16.6**	90.1	90.1
	N_{PMM}	16.4	33.0	22.7	36.6	97.4	**30.5**	97.4
	N_{LRD}	12.1	51.3	11.2	35.5	50.4	85.8	**85.8**
	N_{ERD}	10.6	52.5	8.2	39.3	36.8	116.0	116.0
	t	12.0	74.4	13.1	53.2	35.1	102.7	102.7
	t_{PMM}	21.2	61.0	17.8	**19.7**	100.5	51.0	100.5
	t_{LRD}	14.1	93.8	**5.0**	48.5	33.5	77.7	93.8
	t_{ERD}	56.5	**19.3**	515.4	32.1	100.9	164.7	515.4
	t_{skew}	**9.7**	75.0	8.1	46.5	29.4	116.5	116.5
1000	N	34.6	8.4	**13.5**	17.8	4.5	8.4	34.6
	N_{PMM}	30.7	5.5	25.8	10.7	4.6	7.1	30.7
	N_{LRD}	20.2	7.7	21.2	4.7	2.7	8.7	21.2
	N_{ERD}	27.0	7.6	18.6	14.1	4.4	6.1	27.0
	t	22.4	**3.2**	23.6	**3.9**	3.4	**3.8**	23.6
	t_{PMM}	21.2	11.6	15.8	7.3	5.8	5.4	**21.2**
	t_{LRD}	**18.3**	5.8	14.6	5.4	**2.0**	9.9	18.3
	t_{ERD}	90.5	69.1	758.3	31.2	105.5	612.8	758.3
	t_{skew}	25.9	5.0	14.6	9.2	6.4	4.1	25.9

t models when $n = 1000$. However, when $n = 200$, the skew t model provides the best coverages on average, followed by the t model. This is also the case if we average the MSEs across sample size. Certainly, the unadjusted t and the skew t models generally perform better than the unadjusted Normal model.

According to maximums across scenarios, sample sizes, and MDMs, the Normal with PMM, LRD, or ERD, and the t with PMM or LRD seem to provide the best coverages.

The excellent performance of the models with modest adjustments for skewness, namely the PMM, LRD, and ERD on Normal errors, may suggest that in the context of imputation it suffices restrict error draws within the realm of observed errors in the data set. If,

Table 4.10: QQ MSE for imputations under Data Scenario 3

	MDM:	**MCAR**			**MAR**			
N	**Variable:**	**Y2**	**Y3**	**Y4**	**Y2**	**Y3**	**Y4**	**MAX**
200	N	21.7	55.3	**23.1**	66.7	76.7	32.8	**76.7**
	N_{PMM}	**15.7**	201.6	27.4	279.4	57.2	55.8	279.4
	N_{LRD}	29.4	183.7	83.9	79.6	35.4	20.7	183.7
	N_{ERD}	23.5	86.7	51.3	67.6	67.5	14.3	86.7
	t	28.0	**51.4**	62.4	94.8	63.4	**10.9**	94.8
	t_{PMM}	*16.4*	178.2	*23.5*	263.9	33.5	30.2	263.9
	t_{LRD}	29.8	234.5	67.6	154.8	**22.3**	20.2	234.5
	t_{ERD}	43.8	392.1	583.6	**53.9**	44.4	558.8	583.6
	t_{skew}	23.1	69.3	63.3	62.1	92.7	17.4	92.7
1000	N	37.9	17.4	**4.6**	51.2	**3.1**	**2.4**	51.2
	N_{PMM}	**2.1**	**5.0**	12.1	**5.1**	11.2	6.9	12.1
	N_{LRD}	8.7	8.1	19.7	5.4	9.4	9.3	19.7
	N_{ERD}	5.9	7.8	14.8	13.0	8.1	5.1	14.8
	t	7.5	7.8	13.3	9.2	7.5	4.7	13.3
	t_{PMM}	2.7	5.4	8.5	10.4	11.3	5.2	**11.3**
	t_{LRD}	8.5	5.6	16.6	7.7	14.1	5.3	16.6
	t_{ERD}	7.6	200.9	835.6	6.4	221.8	711.8	835.6
	t_{skew}	10.5	7.9	14.0	13.5	5.8	3.5	14.0

Table 4.11: QQ MSE for imputations under Data Scenario 4

	MDM:	**MCAR**			**MAR**			
N	**Variable:**	**Y2**	**Y3**	**Y4**	**Y2**	**Y3**	**Y4**	**MAX**
200	N	85.2	21.2	24.0	59.4	82.3	50.1	85.2
	N_{PMM}	23.3	18.5	153.2	47.2	**30.5**	19.6	153.2
	N_{LRD}	24.1	18.5	27.5	55.9	68.2	**10.9**	68.2
	N_{ERD}	85.8	**15.3**	**13.2**	61.8	45.2	24.7	85.8
	t	26.6	17.6	23.3	28.4	63.3	59.6	63.3
	t_{PMM}	17.7	33.3	106.0	47.1	34.5	24.5	106.0
	t_{LRD}	12.8	18.0	19.3	30.2	34.1	16.1	**34.1**
	t_{ERD}	717.5	711.0	381.2	803.0	559.4	444.8	803.0
	t_{skew}	**11.5**	24.0	31.0	**16.3**	41.5	93.2	93.2
	N	10.9	25.6	65.5	53.3	11.2	92.5	92.5
1000	N_{PMM}	5.8	**6.2**	*24.7*	**9.9**	16.9	20.4	24.7
	N_{LRD}	6.3	28.7	31.5	49.7	5.7	49.7	49.7
	N_{ERD}	5.5	13.9	28.1	43.6	5.9	36.3	43.6
	t	14.2	21.9	57.9	52.1	**4.7**	74.6	74.6
	t_{PMM}	6.3	8.4	**23.8**	11.1	16.3	**17.4**	**23.8**
	t_{LRD}	**4.9**	27.6	27.7	38.3	8.0	35.3	38.3
	t_{ERD}	599.4	532.5	208.1	646.4	607.8	233.6	646.4
	t_{skew}	15.8	22.8	67.6	55.7	6.8	74.1	74.1

however, it is important to allow for errors in the imputations that are wider than the existing observed errors (for example, if extreme proportions of the data sets are missing), then one could assume that these adaptations to the symmetric models will not suffice.

In conclusion, the second analysis shows that the most accurate of the pure distribution-based imputation methods is the skew t model, followed by the symmetric t model. If adaptations to incorporate observed skewness are deemed suitable for the data set, the Normal model with any adaptation (PMM, LRD, or ERD) will suffice, and will significantly reduce computation time.

Table 4.12: QQ MSE for imputations under Data Scenario 5

	MDM:	**MCAR**			**MAR**			
N	**Variable:**	**Y2**	**Y3**	**Y4**	**Y2**	**Y3**	**Y4**	**MAX**
200	N	302.7	166.0	47.3	177.8	43.5	49.6	302.7
	N_{PMM}	**7.8**	57.1	25.3	42.6	21.6	54.7	**57.1**
	N_{LRD}	17.3	43.6	41.7	20.7	31.6	71.2	71.2
	N_{ERD}	119.2	50.8	19.6	22.6	**10.7**	55.7	119.2
	t	53.7	20.9	**16.7**	15.4	20.5	57.5	57.5
	t_{PMM}	125.2	17.8	32.2	38.0	32.0	**35.8**	125.2
	t_{LRD}	31.9	**13.7**	40.3	58.1	25.4	47.0	58.1
	t_{ERD}	379.0	824.3	2721.6	367.2	1272.0	2691.9	2721.6
	t_{skew}	31.3	*13.8*	21.5	**7.9**	20.5	60.2	60.2
1000	N	103.5	189.6	**2.5**	271.4	191.0	**2.7**	271.4
	N_{PMM}	14.1	**7.8**	13.9	10.1	3.5	11.7	**14.1**
	N_{LRD}	25.4	8.4	10.0	**7.4**	7.5	8.3	25.4
	N_{ERD}	**8.5**	19.5	12.8	10.5	4.9	11.6	19.5
	t	24.1	16.9	12.7	12.7	**1.4**	9.8	24.1
	t_{PMM}	20.7	11.3	12.6	9.4	2.2	17.4	20.7
	t_{LRD}	25.4	12.5	7.1	10.5	6.4	10.1	25.4
	t_{ERD}	532.4	750.4	2719.8	428.4	974.4	2506.6	2719.8
	t_{skew}	21.3	10.8	7.3	14.0	1.9	14.7	21.3

4.5 Conclusion

Clearly, at the very least Normal SRMI should be used instead of incomplete data analysis or CC data analysis. The Normal model has proved to be relatively robust to misspecification within the SRMI approach for the data variations presented in this chapter when compared with complete case or incomplete data analysis.

However, an imputer can choose a better SRMI model based on the results of this chapter. It seems that in many cases the Normal model with a PMM, LRD or ERD adaptation on the imputed errors will suffice in order to accommodate skewness. The advantage of this choice is the quicker computation time, since the t and skew t models are much more complex in their implementation. It also seems as though the t models with PMM or LRD adaptations are rather robust. The t model with ERD is not recommended due to its generation of errors that lead to poor coverage of imputations over the original data points before they were made missing. Unfortunately, the t model with adaptation does not share the computational simplicity of the Normal model, with or without adaptations.

If one prefers to allow for errors in imputations that are outside of the limits of the errors that are actually observed, a more robust adaptation-free approach should be considered. If this is the case, the obvious choice is the skew t approach. In this chapter the skew t approach has been shown to have favourable properties under many of the simulation scenarios. Moreover, it would seem that the skew t model, while not always being the best choice of imputation model, has shown no serious weaknesses in the context of this simulation study even when compared to the adapted Normal models. The skew t model is, therefore, an acceptable choice of imputation model should the error distributions of the data not be known. The disadvantage of increased computation time (compared with the Normal model) is more than offset by the model's flexibility.

Chapter 5

SRMI Evaluation

5.1 Introduction

The review of previous literature shows that a single measure for the evaluation of imputation has not yet been agreed upon. However, there are several plausible methods that should be considered when evaluating new SRMI methods. These methods include the relative bias ($RBIAS$) and the root of the relative mean squared error ($RRMSE$), as used by He & Raghunathan (2009), the QQ-plots and KS statistic used by Abayomi et al. (2008), and the calibrated posterior predictive p-value, as used by Cabras et al. (2011).

In general, this thesis is concerned with the evaluation of the model used in an SRMI procedure. In particular, an imputer would like to know whether or not the model chosen for a particular incomplete variable is appropriate, given the observed data for that variable. For some models, of course, this check is unnecessary, when a single model is known to be appropriate (for example, a Poisson model for count data). However, where several models might fit a particular variable, we either need to find the optimal model, or at least rule out any ill-fitting choice.

We desire a statistic to determine whether or not the skew t-distribution would suffice when it used in SRMI, knowing only the incomplete data.[1] However, the evaluation problem can be simplified somewhat when we consider that the skew t-distribution is in

[1] Imputation methods can be tested using coverage, bias, and confidence interval width only if the original complete data is known.

fact a robust alternative to the standard practice of using a Normal model for continuous data (that may or may not have been transformed to better represent a Normal bell-shaped density). For this reason, we need only assess the fit of a Normal distribution to a variable's observed data, and if the fit statistic rejects Normality, we can switch the model for the data to the more robust skew t-distribution. Whether or not the skew t-distribution is the best fit becomes a moot point when we consider that no alternative exists for the skew t-distribution at this point. Of course, there is merit in completing the same research for the skew t-distribution as we will be doing for the Normal case, if one considers that non-Normal, non-skew-t data might need transformation to allow the skew t-distribution to fit the data properly. However, for the sake of simplicity, this thesis will be concerned with the null hypothesis of Normality (with the alternative being the skew t-distribution), laying the groundwork for further research into model assessment for a null hypothesis that the data is t- or skew t-distributed.

To start the search for an appropriate assessment method, let us first review the three main methods glanced at in Chapter 3.

5.1.1 Relative Bias and Root Relative Mean Squared Error

The first set of assessment measures briefly considered are those used by He & Raghunathan (2009). As explained in Subsection 3.4.5, $RBIAS$ and $RRMSE$ are calculated as follows:

$$RBIAS = \left|\frac{Bias}{True}\right| \times 100\% \tag{5.1}$$

$$RRMSE = \sqrt{\frac{MSE(Method)}{MSE(Before\ deletion)}} \tag{5.2}$$

To review, a reasonable upper limit for $RBIAS$ is 5%. $RRMSE$ measures the increase of $RMSE$ relative to that of the analysis before deletion of the missing observations. In general, it is expected to be more than 1, reflecting loss of efficiency when analysing incomplete data. However, it can be less than 1 as well, implying "superefficiency" of the

MI models, as discussed by Rubin (1996).

However, assessing the correctness of an imputation regression model's coefficients may not always be the right process for one important reason, namely that a data set created with non-Normal errors might be better fitted by a regression model with intercept and slope coefficients that are different to those used to actually create the data. This is particularly evident in the case of the skew t-distribution, where the skewness parameter directly influences the size of the intercept parameter. If we were to use the $RBIAS$ and $RRMSE$ measures to assess the correctness of the imputed intercept, the results would show poor estimation of the coefficient, although the mean squared error from overall model fit, and possibly from predictions as well, could be vastly superior to a model showing lower coefficient $RBIAS$ and $RRMSE$. For this reason, this thesis is more concerned with a method that can directly compare the distributions of imputation-completed data sets with the actual complete data, and any measure that can possibly provide a distribution-based statistic of SRMI model GOF.

5.1.2 Quantile-quantile plots and the Kolmogorov-Smirnov test

In order to assess the extent to which the overall distribution of the imputed data matches the distribution of the original data, one can plot the quantiles of the completed data against the quantiles of the complete data before deletion. This will allow us to view on a graph the exact stages of the distribution in which the imputation model differs from the original data's model, namely, for example, in the tails, around the mean, to the left or right of the distribution, *etc.*

Along with this visual representation, the Kolmogorov-Smirnov (KS) statistic can be derived to test the null hypothesis that the completed data conforms to the known distribution of the complete data. This test, however, is often quite lenient when it comes to deviations from the original distribution (as we will see in the next chapter), so we may in fact have to rely on the QQ plots more often than on the KS statistic.

Because of the graphical nature of this evaluation method, it involves a fair amount of subjectivity, which leaves little room for research opportunities for this thesis.

5.1.3 Calibrated Posterior Predictive p-Value

The calibrated posterior predictive p-value, or *cppp*, has been used to assess the GOF of imputation regressions, as explained in Section 3.4.6. To review, if $ppp\left(y\right) = \Pr\left(D\left(Y^{rep}, \beta\right) \geq D\left(y, \beta\right)|y\right)$, then the *cppp* is $\Pr\left(ppp\left(Y\right) \leq ppp\left(y\right)\right)$. We have that D is a discrepancy measure, y represents the observed data, the distribution of Y^{rep} is the posterior predictive distribution, and Y is a replicate sample from the prior predictive distribution.

The *cppp* statistic has evolved from the mathematics behind the *ppp* measure, discussed at length in Gelman, Meng & Stern (1996). This measure has a substantial history, as well as the choice of the discrepancy measure (and the actual calibration of the *ppp*).

Hjort et al. (2006) have done some work on *ppp* statistic calibration in the case of Normality which suggests that the statistic could have potential in identifying situations where the traditional Normal model used in SRMI should be replaced by the more robust skew t SRMI model introduced in Chapter 4. Therefore, this chapter will focus on some of the mathematics behind the *ppp* and *cppp* statistics, will attempt to derive an appropriate discrepancy measure in order to meet the objective, and will then check on the distributional properties of the *ppp* and *cppp* statistics in order to assess this method's use within the required context. Simulation checks of the mathematical results follow, to verify the mathematical derivations.

This research will focus on model evaluation for the complete-data case, since the extension to incomplete data is rather trivial because, in essence, the *cppp* statistic is designed to test if data fits a particular broad model; whether the data is complete, incomplete or completed makes no difference. In SRMI, the *cppp* statistic could be generated to test whether the observed part of a variable fits the model attached to that variable in the SRMI process, or it could be generated to test whether the completed variable still fits the model attached to that variable in the SRMI process, or both. In both cases, the *cppp* is generated on what is considered to be complete data. Hence it suffices to develop the theory on the statistic under the complete-data assumption.

5.2 Calibrated Posterior Predictive p-Values

In this section we review two methods of deriving the *cppp* statistics, namely the mathematical formulation, as introduced by Hjort et al. (2006), and the MCMC approximation of the *cppp*, as explained by Cabras et al. (2011).

5.2.1 Mathematical derivation — existing discrepancy

In this subsection, we review the mathematical derivation of the *cppp* for the discrepancy measure given by Hjort et al. (2006). We use the following notation: Y is a random variable with elements $y_i, i = 1, 2, \ldots, n$, and Y_{obs} represents the observed data, while Y^{rep} represents a posterior predictive replicate of the observed data.

Regression context

We focus on the general regression model for data (x_i, y_i) where $y_i = X_i'\beta + \epsilon_i$, for $i = 1, \ldots, n$. The vector X_i is a p-dimensional covariate vector for observation i, and the ϵ_i's are independent with mean zero and standard deviation σ. In matrix notation we have the standard OLS result, $\hat{\beta} = (X'X)^{-1} X'Y = \Omega_n^{-1}X'Y$.

Suppose that $D(y, \theta) = \frac{1}{\sigma^2}\left(\hat{\beta} - \beta\right)' \Omega_n \left(\hat{\beta} - \beta\right)$. We then have that $D(Y^{rep}, \theta) = \frac{1}{\sigma^2}\left(\hat{\beta}^{rep} - \beta\right)' \Omega_n \left(\hat{\beta}^{rep} - \beta\right)$. This is the discrepancy measure used by Hjort et al. (2006, p. 1166).

Consider σ^2 known. We know the following:

$$\begin{aligned}
\hat{\beta}^{rep}|\beta, \sigma^2 &\sim N\left(\beta, \sigma^2 (X'X)^{-1}\right) = N\left(\beta, \sigma^2\Omega_n^{-1}\right) \\
\therefore \Omega_n^{\frac{1}{2}}\hat{\beta}^{rep}|\beta, \sigma^2 &\sim N\left(\Omega_n^{\frac{1}{2}}\beta, \sigma^2\Omega_n^{\frac{1}{2}}\Omega_n^{-1}\Omega_n^{\frac{1}{2}}\right) = N\left(\Omega_n^{\frac{1}{2}}\beta, \sigma^2 I_p\right) \\
\therefore \Omega_n^{\frac{1}{2}}\hat{\beta}^{rep} - \Omega_n^{\frac{1}{2}}\hat{\beta}|\beta, \sigma^2 &\sim N\left(0, \sigma^2 I_p\right) \\
\therefore Z|\beta, \sigma^2 = \frac{1}{\sigma}\Omega_n^{\frac{1}{2}}\left(\hat{\beta}^{rep} - \hat{\beta}\right) &\sim N\left(0, I_p\right) \\
\therefore Z'Z|\beta, \sigma^2 = \frac{1}{\sigma^2}\left(\hat{\beta}^{rep} - \beta\right)' \Omega_n \left(\hat{\beta}^{rep} - \beta\right) &\sim \chi_p^2
\end{aligned}$$

Now suppose that our data is given by Y_{obs}. Also assume that β has a prior distribution of the form $N_p\left(\beta_o, \sigma^2\left(c_0\Omega_0\right)^{-1}\right)$. Hjort et al. (2006) mentioned that the parametrisation is somewhat redundant, since c_0 can be incorporated into the Ω_0, but it is useful to start with the covariance structure and then explore different levels of sharpness (c_0 large) or vagueness (c_0 small).

We can obtain the posterior of β as follows:

$$\begin{aligned} p\left(\beta|\sigma^2, y^{obs}\right) &\propto L\left(\beta|\sigma^2, y^{obs}\right) p\left(\beta\right) \\ &\propto \frac{\left|\Omega_n\right|^{\frac{1}{2}}}{\left(2\pi\right)^{\frac{1}{2}p}\left(\sigma^2\right)^{\frac{1}{2}p}} \exp\left[-\frac{1}{2\sigma^2}\left(\hat{\beta}^{obs}-\beta\right)'\Omega_n\left(\hat{\beta}^{obs}-\beta\right)\right] \\ &\times \frac{\left|c_0\Omega_0\right|^{\frac{1}{2}}}{\left(2\pi\right)^{\frac{1}{2}p}\left(\sigma^2\right)^{\frac{1}{2}p}} \exp\left[-\frac{c_0}{2\sigma^2}\left(\beta-\beta_0\right)'\Omega_0\left(\beta-\beta_0\right)\right] \end{aligned}$$

We now complete the square in the exponent with respect to β:

$$\left(\hat{\beta}^{obs}-\beta\right)'\Omega_n\left(\hat{\beta}^{obs}-\beta\right)+c_0\left(\beta-\beta_0\right)'\Omega_0\left(\beta-\beta_0\right)$$

More specifically, consider only the terms involving β:

$$\beta'\Omega_n\beta-2\beta'\Omega_n\hat{\beta}^{obs}+c_0\beta'\Omega_0\beta-2c_0\beta'\Omega_0\beta_0$$

If we let $\tilde{\beta}=\left(c_0\Omega_0+\Omega_n\right)^{-1}\left(c_0\Omega_0\beta_0+\Omega_n\hat{\beta}^{obs}\right)$, we have:

$$\beta'\Omega_n\beta-2\beta'\Omega_n\hat{\beta}^{obs}+c_0\beta'\Omega_0\beta-2c_0\beta'\Omega_0\beta_0 \propto \left(\beta-\tilde{\beta}\right)'\left(c_0\Omega_0+\Omega_n\right)\left(\beta-\tilde{\beta}\right)$$

As is also mentioned by Hjort et al. (2006), we thus find that:

$$\beta|\sigma^2, y^{obs} \sim N\left(\tilde{\beta}, \sigma^2\left(c_0\Omega_0+\Omega_n\right)^{-1}\right)$$

We also have the following:

$$\begin{aligned}
\left(c_0\Omega_0 + \Omega_n\right)^{-1} \Omega_n\hat{\beta}^{obs} - \hat{\beta}^{obs} &= \left(c_0\Omega_0 + \Omega_n\right)^{-1} \left[\Omega_n - \left(c_0\Omega_0 + \Omega_n\right)\right] \hat{\beta}^{obs} \\
&= - \left(c_0\Omega_0 + \Omega_n\right)^{-1} c_0\Omega_0\hat{\beta}^{obs} \\
\therefore \tilde{\beta} - \hat{\beta}^{obs} &= \left(c_0\Omega_0 + \Omega_n\right)^{-1} c_0\Omega_0 \left(\beta_0 - \hat{\beta}^{obs}\right) \\
\therefore \left(\beta - \hat{\beta}^{obs}\right) |y^{obs}, \sigma^2 &\sim N\left(\left(c_0\Omega_0 + \Omega_n\right)^{-1} c_0\Omega_0 \left(\beta_0 - \hat{\beta}^{obs}\right), \sigma^2 \left(c_0\Omega_0 + \Omega_n\right)^{-1}\right) \\
&\sim N\left(\tilde{f}, \sigma^2 \left(c_0\Omega_0 + \Omega_n\right)^{-1}\right) \\
\therefore \frac{1}{\sigma} \left(\beta - \hat{\beta}^{obs}\right) |y^{obs}, \sigma^2 &\sim N\left(\frac{1}{\sigma}\tilde{f}, \left(c_0\Omega_0 + \Omega_n\right)^{-1}\right) = N\left(f, \left(c_0\Omega_0 + \Omega_n\right)^{-1}\right)
\end{aligned}$$

Now the discrepancy measure for the observed data, as well as the *ppp* for the observed data, can be found as follows:

$$\therefore D\left(y^{obs}, \theta\right) = \left(U + f\right)' \Omega_n \left(U + f\right),$$

where $U \sim N\left(0, \left(c_0\Omega_0 + \Omega_n\right)^{-1}\right)$ and $f = \frac{1}{\sigma}\left(c_0\Omega_0 + \Omega_n\right)^{-1} c_0\Omega_0 \left(\beta_0 - \hat{\beta}^{obs}\right)$, both vectors.

$$\begin{aligned}
\therefore ppp\left(y^{obs}\right) &= \Pr\left[D\left(Y^{rep}, \theta\right) \geq D\left(y^{obs}, \theta\right) |X\right] \\
&= \Pr\left[\chi^2_p \geq \left(U + f\right)' \Omega_n \left(U + f\right)\right]
\end{aligned}$$

This result is shown by Hjort et al. (2006, p. 1166). Now consider the special case: $\Omega_0 = \frac{1}{n}\Omega_n$. In this case the prior variance of β is specified as being proportional to the sample variance of its least squares estimator. The coefficient c_0 then has a more precise interpretation relative to the sample size (Hjort et al. 2006, p. 1166).

$$\begin{aligned}
E\left(\beta|y^{obs}\right) = \tilde{\beta} &= \left(c_0\Omega_0 + \Omega_n\right)^{-1} \left(c_0\Omega_0\beta_0 + \Omega_n\hat{\beta}^{obs}\right) \\
&= \left(\frac{c_0}{n}\Omega_n + \Omega_n\right)^{-1} \left(\frac{c_0}{n}\Omega_n\beta_0 + \Omega_n\hat{\beta}^{obs}\right) \\
&= \left(\frac{c_0 + n}{n}\right)^{-1} \left(\frac{c_0}{n}\beta_0 + \hat{\beta}^{obs}\right) \\
&= \frac{n}{c_0 + n} \left(\frac{c_0}{n}\beta_0 + \hat{\beta}^{obs}\right) \\
&= \frac{c_0}{c_0 + n}\beta_0 + \frac{n}{c_0 + n}\hat{\beta}^{obs}
\end{aligned}$$

Now U is multivariate Normal, as follows:

$$\begin{aligned} U \sim N\left(0, (c_0\Omega_0 + \Omega_n)^{-1}\right) &= N\left(0, \left(\frac{c_n}{n}\Omega_n + \Omega_n\right)^{-1}\right) \\ &= N\left(0, \Omega_n^{-1}\left(\frac{c_0 + n}{n}\right)^{-1}\right) \\ &= N\left(0, \Omega_n^{-1}\frac{n}{c_0 + n}\right) \end{aligned}$$

Then f can be simplified:

$$\begin{aligned} f &= \frac{1}{\sigma}\left(c_0\Omega_0 + \Omega_n\right)^{-1} c_0\Omega_0\left(\beta_0 - \hat{\beta}^{obs}\right) \\ &= \frac{1}{\sigma}\left(\frac{c_0}{n}\Omega_n + \Omega_n\right)^{-1}\frac{c_0}{n}\Omega_n\left(\beta_0 - \hat{\beta}^{obs}\right) \\ &= \frac{1}{\sigma}\left(\frac{n}{c_o + n}\right)\frac{c_0}{n}\left(\beta_0 - \hat{\beta}^{obs}\right) \\ &= \frac{1}{\sigma}\left(\frac{c_0}{c_o + n}\right)\left(\beta_0 - \hat{\beta}^{obs}\right) \end{aligned}$$

For $ppp\left(y^{obs}\right)$ we can now simplify $(U + f)'\,\Omega_n\,(U + f)$:

$$\begin{aligned} &(U + f)'\,\Omega_n\,(U + f) \\ &\quad= \left[U + \frac{c_0}{c_0 + n}\left(\frac{\beta_0 - \hat{\beta}^{obs}}{\sigma}\right)\right]'\Omega_n\left[U + \frac{c_0}{c_0 + n}\left(\frac{\beta_0 - \hat{\beta}^{obs}}{\sigma}\right)\right] \\ &\quad= \frac{n}{c_0 + n}\left[\frac{(c_0 + n)^{\frac{1}{2}}}{n^{\frac{1}{2}}}U + \frac{c_0\left(\beta_0 - \hat{\beta}^{obs}\right)}{n^{\frac{1}{2}}(c_0 + n)^{\frac{1}{2}}\sigma}\right]'\Omega_n\left[\frac{(c_0 + n)^{\frac{1}{2}}}{n^{\frac{1}{2}}}U + \frac{c_0\left(\beta_0 - \hat{\beta}^{obs}\right)}{n^{\frac{1}{2}}(c_0 + n)^{\frac{1}{2}}\sigma}\right] \\ &\quad= \frac{n}{c_0 + n}\left[W + \frac{c_0\left(\beta_0 - \hat{\beta}^{obs}\right)}{n^{\frac{1}{2}}(c_0 + n)^{\frac{1}{2}}\sigma}\right]'\Omega_n\left[W + \frac{c_0\left(\beta_0 - \hat{\beta}^{obs}\right)}{n^{\frac{1}{2}}(c_0 + n)^{\frac{1}{2}}\sigma}\right], \end{aligned}$$

where $W \sim N\left(0, \Omega_n^{-1}\right)$. Further simplifying yields the following:

$$\begin{aligned} &(U + f)'\,\Omega_n\,(U + f) \\ &\quad= \frac{n}{c_0 + n}\left[\Omega_n^{\frac{1}{2}}W + \Omega_n^{\frac{1}{2}}\frac{c_0\left(\beta_0 - \hat{\beta}^{obs}\right)}{n^{\frac{1}{2}}(c_0 + n)^{\frac{1}{2}}\sigma}\right]'\Omega_n^{-\frac{1}{2}}\Omega_n\Omega_n^{-\frac{1}{2}}\left[\Omega_n^{\frac{1}{2}}W + \Omega_n^{\frac{1}{2}}\frac{c_0\left(\beta_0 - \hat{\beta}^{obs}\right)}{n^{\frac{1}{2}}(c_0 + n)^{\frac{1}{2}}\sigma}\right] \\ &\quad= \frac{n}{c_0 + n}\left(\tilde{W} + \tau\right)'\left(\tilde{W} + \tau\right), \end{aligned}$$

where $\tau = \Omega_n^{\frac{1}{2}} \frac{c_0\left(\beta_0 - \hat{\beta}^{obs}\right)}{n^{\frac{1}{2}}(c_0+n)^{\frac{1}{2}}\sigma}$, and $\tilde{W} \sim N(0, I_p)$.

$$\therefore (U+f)' \Omega_n (U+f) \sim \frac{n}{c_0+n} \chi_p^2 (\tau'\tau).$$

The non-central Chi-square distribution has non-centrality parameter $\tau'\tau$, which can be simplified as:

$$\begin{aligned}\tau'\tau &= \left(\Omega_n^{\frac{1}{2}} \frac{c_0\left(\beta_0 - \hat{\beta}^{obs}\right)}{n^{\frac{1}{2}}(c_0+n)^{\frac{1}{2}}\sigma}\right)' \left(\Omega_n^{\frac{1}{2}} \frac{c_0\left(\beta_0 - \hat{\beta}^{obs}\right)}{n^{\frac{1}{2}}(c_0+n)^{\frac{1}{2}}\sigma}\right) \\ &= \frac{c_0^2\left(\beta_0 - \hat{\beta}^{obs}\right)' \Omega_n \left(\beta_0 - \hat{\beta}^{obs}\right)}{\sigma^2 n (c_0+n)}\end{aligned}$$

Now we can solve for the $ppp\left(y^{obs}\right)$ for this special case:

$$\begin{aligned} ppp\left(y^{obs}\right) &= \Pr\left[D\left(Y^{rep}, \theta\right) \geq D\left(y^{obs}, \theta\right) | X\right] \\ &= \Pr\left[\chi_p^2 \geq (U+f)' \Omega_n (U+f)\right] \\ &= \Pr\left[\chi_p^2 \geq \frac{n}{c_0+n}\chi_p^2(\tau'\tau)\right] \\ &= \Pr\left[\frac{\chi_p^2(\tau'\tau)}{\chi_p^2} \leq \frac{c_0+n}{n}\right] \\ &= F_{p,p}\left(1+\frac{c_0}{n}, \tau'\tau\right) \\ &= F_{p,p}\left(1+\frac{c_0}{n}, \frac{c_0^2 K_n^{obs}}{\sigma^2(c_0+n)}\right), \end{aligned}$$

where $K_n^{obs} = \frac{1}{n}\left(\beta_0 - \hat{\beta}^{obs}\right)' \Omega_n \left(\beta_0 - \hat{\beta}^{obs}\right)$.

For $ppp(Y)$, K_n^{obs} is considered to be random, and is written as

$$K_n(Y) = \frac{1}{n}\left(\beta_0 - \hat{\beta}(Y)\right)' \Omega_n \left(\beta_0 - \hat{\beta}(Y)\right).$$

Therefore we have that

$$ppp(Y) = F_{p,p}\left(1+\frac{c_0}{n}, \frac{c_0^2}{c_0+n}\frac{K_n(Y)}{\sigma^2}\right),$$

and

$$\begin{aligned}
cppp\left(y^{obs}\right) &= \Pr\left[ppp\left(Y\right) \leq ppp\left(y^{obs}\right)\right] \\
&= \Pr\left[\frac{K_n\left(Y\right)}{\sigma^2} \geq \frac{K_n^{obs}}{\sigma^2}\right] \\
&= \Pr\left[\frac{\left(\beta_0 - \hat{\beta}\left(Y\right)\right)' \Omega_n \left(\beta_0 - \hat{\beta}\left(Y\right)\right)}{\sigma^2} \geq \frac{\left(\beta_0 - \hat{\beta}^{obs}\right)' \Omega_n \left(\beta_0 - \hat{\beta}^{obs}\right)}{\sigma^2}\right].
\end{aligned}$$

Since $\hat{\beta}\left(Y\right)|\beta, \sigma^2 \sim N\left(\beta, \sigma^2\Omega_n^{-1}\right)$ and $\beta \sim N\left(\beta_0, \sigma^2 \frac{n}{c_0}\Omega_n^{-1}\right)$, it follows that,

$$\hat{\beta}\left(Y\right)|\sigma^2 \ \sim \ N\left(\beta_0, \sigma^2\left(1 + \frac{n}{c_0}\right)\Omega_n^{-1}\right).$$

Therefore,

$$\begin{aligned}
cppp\left(y^{obs}\right) &= \Pr\left[\left(1 + \frac{n}{c_0}\right)\chi_p^2 \geq \frac{\left(\beta_0 - \hat{\beta}^{obs}\right)' \Omega_n \left(\beta_0 - \hat{\beta}^{obs}\right)}{\sigma^2}\right] \\
&= \Pr\left[\chi_p^2 \geq \left(\frac{c_0}{c_0 + n}\right)\frac{\left(\beta_0 - \hat{\beta}^{obs}\right)' \Omega_n \left(\beta_0 - \hat{\beta}^{obs}\right)}{\sigma^2}\right]
\end{aligned}$$

Consider σ^2 unknown. We now consider the regression case where σ^2 is unknown, as was done by Hjort et al. (2006). Suppose that $\lambda = \frac{1}{\sigma^2}$. We have that $\lambda \sim \Gamma\left(\frac{1}{2}a_0, \frac{1}{2}b_0\right)$ and $\beta|\lambda \sim N\left(\beta_0, \lambda^{-1}\left(c_0\Omega_0\right)^{-1}\right)$. We can obtain the joint posterior as follows:

$$\begin{aligned}
p\left(\beta, \lambda|X\right) &\propto L\left(\beta, \lambda|X\right)p\left(\beta, \lambda\right) \\
&\propto \lambda^{\frac{1}{2}p}\exp\left[-\frac{\lambda}{2}\left(\hat{\beta}^{obs} - \beta\right)' \Omega_n \left(\hat{\beta}^{obs} - \beta\right)\right] \\
&\quad \times \lambda^{\frac{1}{2}(n-p)}\exp\left[-\frac{\lambda}{2}\left(y^{obs} - X\hat{\beta}^{obs}\right)'\left(y^{obs} - X\hat{\beta}^{obs}\right)\right] \\
&\quad \times p\left(\beta, \lambda\right),
\end{aligned}$$

with prior $p(\beta,\lambda) \propto \lambda^{\frac{1}{2}a_0-1}\exp\left(\frac{1}{2}b_0\lambda\right)\lambda^{\frac{1}{2}p}\exp\left[-\frac{1}{2}c_0\lambda\left(\beta-\beta_0\right)'\Omega_n\left(\beta-\beta_0\right)\right]$. Completing the square, with respect to β in the exponent, we find that, as before:

$$\begin{aligned}
&\left(\hat{\beta}^{obs}-\beta\right)'\Omega_n\left(\hat{\beta}^{obs}-\beta\right)+c_0\left(\beta-\beta_0\right)'\Omega_0\left(\beta-\beta_0\right)\\
&\quad=\beta'\left(c_0\Omega_0+\Omega_n\right)\beta-2\beta'\left(c_0\Omega_0\beta_0+\Omega_n\hat{\beta}^{obs}\right)+\hat{\beta}'^{obs}\Omega_n\hat{\beta}^{obs}+c_0\beta_0'\Omega_0\beta_0\\
&\quad=\left(\beta-\tilde{\beta}\right)'\left(c_0\Omega_0+\Omega_n\right)\left(\beta-\tilde{\beta}\right)\\
&\qquad-\left(c_0\Omega_0\beta_0+\Omega_n\hat{\beta}^{obs}\right)'\left(c_0\Omega_0+\Omega_n\right)^{-1}\left(c_0\Omega_0\beta_0+\Omega_n\hat{\beta}^{obs}\right)\\
&\qquad+\hat{\beta}'^{obs}\Omega_n\hat{\beta}^{obs}+c_0\beta_0'\Omega_0\beta_0,
\end{aligned}$$

where $\tilde{\beta}=\left(c_0\Omega_0+\Omega_n\right)^{-1}\left(c_0\Omega_0\beta_0+\Omega_n\hat{\beta}^{obs}\right)$.

Therefore, we have that $\beta_0|\lambda,y^{obs}\sim N\left(\tilde{\beta},\frac{1}{\lambda}\left(c_0\Omega_0+\Omega_n\right)^{-1}\right)$.

Since we have,

$$\begin{aligned}
&-\left(c_0\Omega_0\beta_0+\Omega_n\hat{\beta}^{obs}\right)'\left(c_0\Omega_0+\Omega_n\right)^{-1}\left(c_0\Omega_0\beta_0+\Omega_n\hat{\beta}^{obs}\right)+\hat{\beta}'^{obs}\Omega_n\hat{\beta}^{obs}+c_0\beta_0'\Omega_0\beta_0\\
&\quad=\left(\beta_0-\hat{\beta}^{obs}\right)'\Omega_n\left(c_0\Omega_0+\Omega_n\right)^{-1}c_0\Omega_0\left(\beta_0-\hat{\beta}^{obs}\right)\\
&\quad=\left(\beta_0-\hat{\beta}^{obs}\right)'\left(c_0^{-1}\Omega_0^{-1}+\Omega_n^{-1}\right)^{-1}\left(\beta_0-\hat{\beta}^{obs}\right),
\end{aligned}$$

the posterior of λ is then given by,

$$\begin{aligned}
p\left(\lambda|X\right)&\propto\lambda^{\frac{1}{2}(a_0+n)-1}\exp\left\{-\frac{1}{2}\lambda\left[b_0+\left(y^{obs}-X\hat{\beta}^{obs}\right)'\left(y^{obs}-X\hat{\beta}^{obs}\right)\right.\right.\\
&\qquad\left.\left.+\left(\beta_0-\hat{\beta}^{obs}\right)'\left(c_0^{-1}\Omega_0^{-1}+\Omega_n^{-1}\right)^{-1}\left(\beta_0-\hat{\beta}^{obs}\right)\right]\right\}\\
&\propto\lambda^{\frac{1}{2}(a_0+n)-1}\exp\left\{-\frac{1}{2}\lambda\left[b_0+Q_0^{obs}+\left(\beta_0-\hat{\beta}^{obs}\right)'K\left(\beta_0-\hat{\beta}^{obs}\right)\right]\right\},
\end{aligned}$$

where $Q_0^{obs}=\left(y^{obs}-X\hat{\beta}^{obs}\right)'\left(Y_{obs}-X\hat{\beta}^{obs}\right)$ and $K=\left(c_0^{-1}\Omega_0^{-1}+\Omega_n^{-1}\right)^{-1}$. Moving on to the derivation of the *ppp*, we note once again that $D\left(Y^{rep},\beta\right)\sim\chi_p^2$. Moreover,

$$D\left(y^{obs},\theta\right)=\lambda\left(\hat{\beta}^{obs}-\beta\right)'\Omega_n\left(\hat{\beta}^{obs}-\beta\right)$$

Since $\beta|\lambda,X\sim N\left(\tilde{\beta},\left(c_0\Omega_0+\Omega_n\right)^{-1}\right)$, it follows that $\beta|\lambda,X=\tilde{\beta}+\lambda^{-\frac{1}{2}}U$ where U is a

p-dimensional multivariate Normal, $U \sim N\left(0, (c_0\Omega_0 + \Omega_n)^{-1}\right)$.

$$\begin{aligned}
\therefore D\left(y^{obs}, \theta\right) &= \lambda\left(\tilde{\beta} + \lambda^{-\frac{1}{2}}U - \hat{\beta}^{obs}\right)' \Omega_n \left(\tilde{\beta} + \lambda^{-\frac{1}{2}}U - \hat{\beta}^{obs}\right) \\
&= \left(\lambda^{\frac{1}{2}}\left(c_0\Omega_0 + \Omega_n\right)^{-1} c_0\Omega_0 \left(\hat{\beta}^{obs} - \beta_0\right) + U\right)' \Omega_n \\
&\quad \times \left(\lambda^{\frac{1}{2}}\left(c_0\Omega_0 + \Omega_n\right)^{-1} c_0\Omega_0 \left(\hat{\beta}^{obs} - \beta_0\right) + U\right)
\end{aligned}$$

From this expression we can see that it is sufficient to simulate a large number of replicates of (λ_j, U_j) with $\lambda \sim \Gamma\left(\frac{1}{2}a_n, \frac{1}{2}b_n\right)$, with $a_n = a_0 + n$ and,

$$\begin{aligned}
b_n &= b_0 + \left(y^{obs} - X\hat{\beta}^{obs}\right)'\left(y^{obs} - X\hat{\beta}^{obs}\right) + \left(\beta_0 - \hat{\beta}^{obs}\right)'\left(c_0^{-1}\Omega_0^{-1} + \Omega_n^{-1}\right)^{-1}\left(\beta_0 - \hat{\beta}^{obs}\right) \\
&= b_0 + Q_0^{obs} + \left(\beta_0 - \hat{\beta}^{obs}\right)' K \left(\beta_0 - \hat{\beta}^{obs}\right).
\end{aligned}$$

Now consider again the special case: $\Omega_0 = \frac{1}{n}\Omega_n$

$$\begin{aligned}
D\left(y^{obs}, \theta\right) &= \left[\lambda^{\frac{1}{2}}\left(\frac{c_0}{n}\Omega_n + \Omega_n\right)^{-1}\frac{c_0}{n}\Omega_n\left(\hat{\beta}^{obs} - \beta_0\right) + U\right]' \Omega_n \\
&\quad \times \left[\lambda^{\frac{1}{2}}\left(\frac{c_0}{n}\Omega_n + \Omega_n\right)^{-1}\frac{c_0}{n}\Omega_n\left(\hat{\beta}^{obs} - \beta_0\right) + U\right] \\
&= \left[\lambda^{\frac{1}{2}}\frac{c_0}{c_0 + n}\left(\hat{\beta}^{obs} - \beta_0\right) + U\right]' \Omega_n \left[\lambda^{\frac{1}{2}}\frac{c_0}{c_0 + n}\left(\hat{\beta}^{obs} - \beta_0\right) + U\right] \\
&= \left[\lambda^{\frac{1}{2}}\frac{c_0}{c_0 + n}\left(\hat{\beta}^{obs} - \beta_0\right) + \left(\frac{n}{c_0 + n}\right)^{\frac{1}{2}}\Omega_n^{-\frac{1}{2}}Z\right]' \Omega_n \\
&\quad \times \left[\lambda^{\frac{1}{2}}\frac{c_0}{c_0 + n}\left(\hat{\beta}^{obs} - \beta_0\right) + \left(\frac{n}{c_0 + n}\right)^{\frac{1}{2}}\Omega_n^{-\frac{1}{2}}Z\right] \\
&= \left(\frac{n}{c_0 + n}\right)\left[\frac{\lambda^{\frac{1}{2}}c_0\Omega_n^{\frac{1}{2}}}{(c_0 + n)^{\frac{1}{2}}\, n^{\frac{1}{2}}}\left(\hat{\beta}^{obs} - \beta_0\right) + Z\right]' \\
&\quad \times \left[\frac{\lambda^{\frac{1}{2}}c_0\Omega_n^{\frac{1}{2}}}{(c_0 + n)^{\frac{1}{2}}\, n^{\frac{1}{2}}}\left(\hat{\beta}^{obs} - \beta_0\right) + Z\right] \\
&= \left(\frac{n}{c_0 + n}\right)(\tau + Z)'(\tau + Z)
\end{aligned}$$

where $Z \sim N(0, I_p)$, and again we have $\tau = \frac{\lambda^{\frac{1}{2}} c_0 \Omega_n^{\frac{1}{2}}}{(c_0+n)^{\frac{1}{2}} n^{\frac{1}{2}}} \left(\hat{\beta}^{obs} - \beta_0\right)$.

$$\begin{aligned}
\therefore D\left(y^{obs}, \theta\right) &= \left(\frac{n}{c_0+n}\right) \chi_p^2 \left(\tau' \tau\right) \\
&= \left(\frac{n}{c_0+n}\right) \chi_p^2 \left[\frac{c_0^2 \lambda}{n\left(n+c_0\right)} \left(\hat{\beta}^{obs} - \beta_0\right)' \Omega_n \left(\hat{\beta}^{obs} - \beta_0\right)\right] \\
&= \left(\frac{n}{c_0+n}\right) \chi_p^2 \left(\frac{c_0^2 \lambda}{n+c_0} K_n^{obs}\right),
\end{aligned}$$

where $K_n^{obs} = \frac{1}{n} \left(\hat{\beta}^{obs} - \beta_0\right)' \Omega_n \left(\hat{\beta}^{obs} - \beta_0\right)$ once more.

$$\therefore ppp\left(y^{obs}\right) = \Pr\left[\chi_p^2 \geq \frac{n}{c_0+n} \chi_p^2 \left(\lambda \frac{c_0}{c_0+n} K_n^{obs}\right)\right] \tag{5.3}$$

$$ppp\left(y^{obs}\right) = \int_0^\infty F_{p,p}\left(\frac{c_0+n}{n}, \lambda \frac{c_0^2}{c_0+n} K_n^{obs}\right) g_n(\lambda)\, d\lambda, \tag{5.4}$$

where $g_n(\lambda)$ is a Gamma density with parameters $a_n = a_0 + n$ and

$$b_n = b_0 + \left(y^{obs} - X\hat{\beta}^{obs}\right)' \left(y^{obs} - X\hat{\beta}^{obs}\right) + \frac{c_0}{c_0+n} n K_n^{obs}.$$

To derive $ppp(Y)$, K_n^{obs} in Equation (5.4) must be considered random. Therefore define $K_n = \frac{1}{n} \left(\hat{\beta}(Y) - \beta_0\right)' \Omega_n \left(\hat{\beta}(Y) - \beta_0\right)$. The marginal distribution of Y, or more specifically, the marginal distribution of $\hat{\beta}(Y)$ must be used to determine the marginal distribution of K_n. Now $\hat{\beta}(Y) | \beta, \sigma^2 \sim N\left(\beta, \sigma^2 \Omega_n^{-1}\right)$. Since $\beta | \sigma^2 \sim N\left(\beta_0, \sigma^2 \left(\frac{n}{c_0}\right) \Omega_n^{-1}\right)$, it follows that $\hat{\beta}(Y) | \sigma^2 \sim N\left(\beta_0, \sigma^2 \Omega_n^{-1} \left(1 + \frac{n}{c_0}\right)\right)$, and $\tilde{Z} = \left(\frac{1}{n}\right)^{\frac{1}{2}} \Omega_n^{-\frac{1}{2}} \left(\hat{\beta}(Y) - \beta_0\right) \sim N\left(0, \sigma^2 \left(\frac{1}{n} + \frac{1}{c_0}\right)\right)$. Therefore,

$$\tilde{Z}' \tilde{Z} = K_n \sim \sigma^2 \left(\frac{1}{n} + \frac{1}{c_0}\right) \chi_p^2 = \sigma^2 \left(\frac{1}{n} + \frac{1}{c_0}\right) Z \tag{5.5}$$

Substitute Equation (5.5) in (5.3), and using the fact that $\lambda = \frac{1}{\sigma^2}$ it follows that,

$$\begin{aligned}
Sppp\left(Y\right) &= \Pr\left\{\chi_p^2 \geq \chi_p^2\left[\frac{1}{\sigma^2}\frac{c_0^2}{c_0+n}\sigma^2\left(\frac{1}{n}+\frac{1}{c_0}\right)Z\right]\right\}\\
&= \Pr\left[\chi_p^2 \geq \frac{n}{c_0+n}\chi_p^2\left(\frac{c_0}{n}Z\right)\right]\\
&= \Pr\left[F_{p,p}\left(\frac{c_0}{n}Z\right) \leq \frac{c_0+n}{n}\right]\\
&= \Pr\left[F_{p,p}\left(\frac{c_0}{n}Z\right) \leq 1+\frac{c_0}{n}\right]
\end{aligned}$$

Therefore, to make probability statements about the *ppp* measure in this special case, we can use the following function:

$$\begin{aligned}
G\left(u\right) &= \Pr\left[ppp\left(Y\right) \leq u\right]\\
&= \Pr\left[F_{p,p}\left(1+\frac{c_0}{n},\frac{c_0}{n}Z\right) \leq u\right].
\end{aligned}$$

$$\begin{aligned}
\therefore cppp\left(y^{obs}\right) &= \Pr\left[ppp\left(Y\right) \leq ppp\left(y^{obs}\right)\right]\\
&= \Pr\left[F_{p,p}\left(1+\frac{c_0}{n},\frac{c_0}{n}Z\right)\right.\\
&\quad \left.\leq \int_0^\infty F_{p,p}\left(1+\frac{c_0}{n},\lambda\frac{c_0}{c_0+n}K_n^{obs}\right)g_n\left(\lambda\right)d\lambda\right]
\end{aligned}$$

5.2.2 Mathematical derivation — new discrepancy

Now consider a new discrepancy measure, expanding on the work done by Hjort et al. (2006). Suppose that $D\left(y,\theta\right) = \frac{1}{\sigma^2}\left(Y-X\beta\right)'\left(Y-X\beta\right)$.

Consider σ^2 known. We have that $Y = X\beta+\epsilon$, where $\epsilon \sim N\left(0,I_p\right)$. Furthermore, since $\hat{\beta} = \left(X'X\right)^{-1}\left(X'Y\right)$,

$$\begin{aligned}
D\left(y,\theta\right) &= \frac{1}{\sigma^2}\left(Y-X\beta\right)'\left(Y-X\beta\right)\\
&= \frac{1}{\sigma^2}\left[\left(Y-X\hat{\beta}\right)'\left(Y-X\hat{\beta}\right)+\left(\beta-\hat{\beta}\right)'X'X\left(\beta-X\hat{\beta}\right)\right]\\
&= \frac{1}{\sigma^2}\left(Y-X\hat{\beta}\right)'\left(Y-X\hat{\beta}\right)+\frac{1}{\sigma^2}\left(\beta-\hat{\beta}\right)'\Omega_n\left(\beta-X\hat{\beta}\right),
\end{aligned}$$

because the cross product term is zero.

Suppose now that the prior for β is $\beta \sim N\left(\beta_0, \sigma^2 \left(c_0\Omega_0\right)^{-1}\right)$. We consider only the special case where $\Omega_0 = \frac{1}{n}\Omega_n$. Then $\beta|\sigma^2 \sim N\left(\beta_0, \sigma^2 \left(\frac{c_0}{n}\Omega_n\right)^{-1}\right)$. Now we determine the posterior distribution of $\frac{1}{\sigma^2}\left(\beta - \hat{\beta}\right)' \Omega_n \left(\beta - X\hat{\beta}\right)$.

$$\begin{aligned}\frac{1}{\sigma^2}\left(\beta - \hat{\beta}\right)' \Omega_n \left(\beta - X\hat{\beta}\right) &\sim \frac{n}{n+c_0}\chi^2_p \left[\frac{c_0^2}{n\left(c_0+n\right)} \frac{\left(\hat{\beta}^{obs} - \beta_0\right)' \Omega_n \left(\hat{\beta}^{obs} - \beta_0\right)}{\sigma^2}\right] \\ &\sim \frac{n}{n+c_0}\chi^2_p \left(\frac{c_0}{c_0+n}\frac{K_n^{obs}}{\sigma^2}\right),\end{aligned}$$

where $K_n^{obs} = \frac{1}{n}\left(\hat{\beta}^{obs} - \beta_0\right)' \Omega_n \left(\hat{\beta}^{obs} - \beta_0\right)$. Define $\left(Y - X\hat{\beta}\right)'\left(Y - X\hat{\beta}\right) = RSS$, the residual sum of squares.

$$D\left(y^{obs}, \theta\right) = \frac{1}{\sigma^2}RSS + \frac{n}{n+c_0}\chi^2_p\left(\frac{c_0}{c_0+n}\frac{K_n^{obs}}{\sigma^2}\right).$$

To obtain $D\left(Y, \theta\right)$ the marginal distribution of Y must be used: $RSS \sim \sigma^2\chi^2_{n-p}$, and,

$$\begin{aligned}D\left(Y, \theta\right) &= \chi^2_{n-p} + \frac{n}{n+c_0}\chi^2_p\left[\frac{c_0\sigma^2\left(\frac{1}{n} + \frac{1}{c_0}\right)Z}{\left(n+c_0\right)\sigma^2}\right] \\ &= \chi^2_{n-p} + \frac{n}{n+c_0}\chi^2_p\left(\frac{c_0}{n}Z\right), \text{ where} Z \sim \chi^2_p. \qquad (5.6)\end{aligned}$$

Equation (5.6) follows from the fact that $\frac{RSS}{\sigma^2} \sim chi^2_{n-p}$ and $\hat{\beta}\left(Y\right)|\sigma^2, \beta, X \sim N\left(\beta, \sigma^2\Omega_n^{-1}\right)$. But, $\beta \sim N\left(\beta_0, \sigma^2\frac{n}{c_0}\Omega_n^{-1}\right)$,

$$\begin{aligned}\therefore \hat{\beta}\left(Y\right)|\sigma^2, X, \beta_0 &\sim N\left(\beta_0, \sigma^2\Omega_n^{-1} + \sigma^2\frac{n}{c_0}\Omega_n^{-1}\right) \\ &\sim N\left(\beta 0, \sigma^2\Omega_n^{-1}\left(1 + \frac{n}{c_0}\right)\right).\end{aligned}$$

From this it follows that,

$$\begin{aligned}
\hat{\beta}(Y) - \beta_0 &\sim N\left(0, \sigma^2 \Omega_n^{-1}\left(1 + \frac{n}{c_0}\right)\right) \\
\therefore \frac{\Omega_n^{\frac{1}{2}}}{\sigma}\left(1 + \frac{n}{c_0}\right)^{-\frac{1}{2}}\left(\hat{\beta}(Y) - \beta_0\right) &\sim N(0, I_p) \\
\therefore \frac{1}{\sigma^2}\left(1 + \frac{n}{c_0}\right)^{-1}\left(\hat{\beta}(Y) - \beta_0\right)' \Omega_n \left(\hat{\beta}(Y) - \beta_0\right) &\sim \chi_p^2 \\
\therefore \frac{1}{\sigma^2}\left(\frac{c_0}{n + c_0}\right)\left(\hat{\beta}(Y) - \beta_0\right)' \Omega_n \left(\hat{\beta}(Y) - \beta_0\right) &\sim \chi_p^2 \\
\therefore \frac{n}{\sigma^2}\left(\frac{c_0}{n + c_0}\right) K_n &\sim \chi_p^2 \\
\therefore \frac{c_0^2}{n + c_0} K_n &\sim \frac{c_0}{n}\chi_p^2 = \frac{c_0}{n} Z
\end{aligned}$$

Now we can move on to the *cppp* formulation.

$$\begin{aligned}
cppp\left(y^{obs}\right) &= \Pr\left[ppp(Y) \leq ppp\left(y^{obs}\right)\right] \\
&= \Pr\left\{\Pr\left[\chi_p^2 \geq \chi_{n-p}^2 + \frac{n}{n + c_0}\chi_p^2\left(\frac{c_0}{n}Z\right)\right]\right. \\
&\qquad \left.\leq \Pr\left[\chi_p^2 \geq \frac{RSS}{\sigma^2} + \frac{n}{n + c_0}\chi_p^2\left(\frac{c_0^2}{c_0 + n}\frac{K_n^{obs}}{\sigma^2}\right)\right]\right\} \\
&= \Pr\left[\chi_{n-p}^2 + \frac{n}{n + c_0}\chi_p^2\left(\frac{c_0}{n}Z\right) \geq \frac{RSS}{\sigma^2} + \frac{n}{n + c_0}\chi_p^2\left(\frac{c_0}{c_0 + n}\frac{K_n^{obs}}{\sigma^2}\right)\right] \\
&= \Pr\left[\chi_{n-p}^2 \geq \frac{RSS}{\sigma^2} + \frac{n}{n + c_0}\chi_p^2\left(\frac{c_0}{c_0 + n}\frac{K_n^{obs}}{\sigma^2}\right) + \frac{n}{n + c_0}\chi_p^2\left(\frac{c_0}{n}Z\right)\right]
\end{aligned}$$

Now $\chi_p^2\left(\frac{c_0}{c_0+n}\frac{K_n^{obs}}{\sigma^2}\right)$ is a non-central chi-square distribution with p degrees of freedom and non-centrality parameter $\left(\frac{c_0}{c_0+n}\frac{K_n^{obs}}{\sigma^2}\right)$. Furthermore, $\chi_p^2\left(\frac{c_0}{n}Z\right)$ is a non-central chi-square distribution with p degrees of freedom and non-centrality parameter $\frac{c_0}{n}Z$, where $Z \sim \chi_p^2$. So the non-centrality parameter is also a random variable.

Consider σ^2 unknown. Once more, $D(y, \theta) = \frac{1}{\sigma^2}(Y - X\beta)'(Y - X\beta)$, with the same prior for β as before, namely, $\beta \sim N\left(\beta_0, \sigma^2\left(c_0\Omega_0\right)^{-1}\right)$. We consider again only the special case where $\Omega_0 = \frac{1}{n}\Omega_n$. The results will be similar to those found in Hjort et al. (2006, Section 5.2).

Let $\lambda = \frac{1}{\sigma^2}$. Then,

$$D\left(Y^{rep}, \theta\right) \sim \chi_n^2$$

and,

$$D\left(y^{obs}, \theta\right) = \lambda RSS + \frac{n}{n+c_0}\chi_p^2\left(\lambda \frac{c_0^2}{n+c_0} K_n^{obs}\right).$$

$$\begin{aligned}
\therefore ppp\left(y^{obs}\right) &= \Pr\left[D\left(Y^{rep}, \theta\right) \geq D\left(y^{obs}, \theta\right)\right] \\
&= \Pr\left[\chi_n^2 \geq \lambda RSS + \frac{n}{n+c_0}\chi_p^2\left(\lambda \frac{c_o^2}{n+c_0} K_n^{obs}\right)\right] \\
&= \Pr\left\{\chi_n^2 \geq \int_0^{\infty}\left[\lambda RSS + \frac{n}{n+c_0}\chi_p^2\left(\lambda \frac{c_o^2}{n+c_0} K_n^{obs}\right)\right] g_n\left(\lambda\right) d\lambda\right\},
\end{aligned}$$

where $g_n\left(\lambda\right)$ is a Gamma density with parameters $a_n = a_0 + n$, and $b_n = b_0 + Q_0^{obs} + \frac{c_0 n K_n^{obs}}{c_0+n}$.

$$\begin{aligned}
\therefore ppp\left(Y\right) &= \Pr\left[D\left(Y^{rep}, \theta\right) \geq D\left(Y, \theta\right)\right] \\
&= \Pr\left[\chi_n^2 \geq \chi_{n-p}^2 + \frac{n}{n+c_0}\chi_p\left(\frac{c_0}{n} Z\right)\right],
\end{aligned}$$

where $Z \sim chi_p^2$.

$$\begin{aligned}
\therefore cppp\left(y^{obs}\right) &= \Pr\left[ppp\left(Y \leq ppp\left(y^{obs}\right)\right)\right] \\
&= \Pr\left\{\Pr\left[\chi_n^2 \geq \chi_{n-p}^2 + \frac{n}{n+c_0}\chi_p^2\left(\frac{c_0}{n} Z\right)\right]\right. \\
&\qquad \left.\leq \Pr\left[\chi_n^2 \geq \int_0^{\infty}\left[\lambda RSS + \frac{n}{n+c_0}\chi_p^2\left(\lambda \frac{c_o^2}{n+c_0} K_n^{obs}\right)\right] g_n\left(\lambda\right) d\lambda\right]\right\} \\
&= \Pr\left\{\chi_{n-p}^2 + \frac{n}{n+c_0}\chi_p^2\left(\frac{c_0}{n} Z\right)\right. \\
&\qquad \left.\geq \int_0^{\infty}\left[\lambda RSS + \frac{n}{n+c_0}\chi_p^2\left(\lambda \frac{c_o^2}{n+c_0} K_n^{obs}\right)\right] g_n\left(\lambda\right) d\lambda\right\}
\end{aligned}$$

We now have a usable formulation for the *cppp* statistic for a new discrepancy measure. However, the problem with this formulation of the *cppp* is that the prior distribution has to be known. Needing to know the exact prior is not practical, so we turn to the method of Cabras et al. (2011), using a minimal training sample in order to transform a vague prior into a proper one. However, before moving to this MCMC approximation method, we simplify into a non-regression context, with σ^2 known, so that we can illustrate the distribution of the *cppp*.

Non-regression context

Consider σ^2 known. Suppose that $y_i|\theta \sim N(\theta, \sigma^2)$, $i = 1, 2, \ldots, n$, and that $\theta|\theta_0, \sigma_0^2 \sim N(\theta_0, \sigma_0^2)$. Also, consider the discrepancy measure $D(Y, \theta) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2$. Therefore, we know that $D(Y^{rep}, \theta) \sim \chi_n^2$.

Now, $D(Y, \theta) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2$, which can be written as:

$$\begin{aligned} D(Y, \theta) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left[\left(y_i - \bar{y}^{obs}\right) - \left(\theta - \bar{y}^{obs}\right)\right]^2 \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i - \bar{y}^{obs}\right)^2 + \frac{n\left(\bar{y}^{obs} - \theta\right)}{\sigma^2}, \end{aligned}$$

since the cross product is zero.

$$\therefore D(Y, \theta) = \frac{RSS}{\sigma^2} + \frac{n\left(\bar{y}^{obs} - \theta\right)^2}{\sigma^2}$$

The last term is the discrepancy measure considered by Hjort et al. (2006).

We need to consider the posterior distribution of θ in order to obtain $D\left(y^{obs}, \theta\right)$. It can be shown that the posterior distribution of θ is as follows:

$$\theta|\sigma^2, \theta_0, \sigma_0^2, y^{obs} \sim N\left[(1 - P_n)\theta_0 + P_n\bar{y}^{obs}, \frac{P_n\sigma^2}{n}\right],$$

where

$$P_n = \frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2}.$$

Thus, if $z_1 \sim N(0, 1)$, then we have that:

$$\theta|\sigma^2, \theta_0, \sigma_0^2, y^{obs} \sim (1 - P_n)\theta_0 + P_n\bar{y}^{obs} + z_1\frac{P_n^{\frac{1}{2}}\sigma}{\sqrt{n}}.$$

$$\begin{aligned}
\therefore D\left(y^{obs},\theta\right) &= \frac{RSS}{\sigma^2} + \frac{n\left\{\bar{y}^{obs} - \left[(1-P_n)\,\theta_0 + P_n\bar{y}^{obs} + z_1\frac{P_n^{\frac{1}{2}}\sigma}{\sqrt{n}}\right]\right\}^2}{\sigma^2} \\
&= \frac{RSS}{\sigma^2} + \left[P_n^{\frac{1}{2}}z_1 - (1-P_n)\,\frac{\sqrt{n}\left(\bar{y}^{obs}-\theta_0\right)}{\sigma}\right]^2
\end{aligned}$$

$$\begin{aligned}
\therefore ppp\left(y^{obs}\right) &= \Pr\left[D\left(Y^{rep},\theta\right) \geq D\left(y^{obs},\theta\right)\right] \\
&= \Pr\left\{\chi^2_n \geq \frac{RSS}{\sigma^2} + \left[P_n^{\frac{1}{2}}z_1 - (1-P_n)\,\frac{\sqrt{n}\left(\bar{y}^{obs}-\theta_0\right)}{\sigma}\right]^2\right\} \\
&= \Pr\left\{\chi^2_n \geq \frac{RSS}{\sigma^2} + P_n\left[z_1 - \frac{\sqrt{n}\,(1-P_n)}{\sigma P_n^{\frac{1}{2}}}\left(\bar{y}^{obs}-\theta_0\right)\right]^2\right\}
\end{aligned}$$

We now need the marginal distribution of $\bar{y}$ for the null distribution of $ppp\,(Y)$. We know that $\bar{y}|\sigma^2,\theta_0,\sigma_0^2 \sim N\left(\theta_0,\frac{\sigma^2}{n}+\sigma_0^2\right)$, which means that if $z_2 \sim N\,(0,1)$ independent of z_1, then $\bar{y}|\sigma^2,\theta_0,\sigma_0^2 \sim \theta_0+\left(\frac{\sigma^2}{n}+\sigma_0^2\right)^{\frac{1}{2}}z_2$. We also know that $\frac{RSS}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^n\left(y_i-\bar{y}\right)^2 \sim \chi^2_{n-1}$. So, we have that:

$$D\,(Y,\theta) = \chi^2_{n-1} + P_n\left[z_1 - \frac{\sqrt{n}\,(1-P_n)}{P_n}\frac{\sigma_0}{\sigma}z_2\right]^2$$

$$\begin{aligned}
\therefore ppp\,(Y) &= \Pr\left[D\left(Y^{rep},\theta\right) \geq D\,(Y,\theta)\right] \\
&= \Pr\left\{\chi^2_n \geq \chi^2_{n-1} + P_n\left[z_1 - \frac{\sqrt{n}\,(1-P_n)}{P_n}\frac{\sigma_0}{\sigma}z_2\right]^2\right\}
\end{aligned}$$

$$
\begin{aligned}
\therefore cppp\left(y^{obs}\right) &= \Pr\left[ppp\left(Y\right) \leq ppp\left(y^{obs}\right)\right] \\
&= \Pr\left\{ \Pr\left[\chi^2_n \geq \chi^2_{n-1} + P_n\left[z_1 - \frac{\sqrt{n}\left(1-P_n\right)}{P_n}\frac{\sigma_0}{\sigma}z_2\right]^2\right]\right. \\
&\qquad \left. \leq \Pr\left[\chi^2_n \geq \frac{RSS}{\sigma^2} + P_n\left[z_1 - \frac{\sqrt{n}\left(1-P_n\right)}{\sigma P_n^{\frac{1}{2}}}\left(\bar{y}^{obs}-\theta_0\right)\right]^2\right]\right\} \\
&= \Pr\left\{\chi^2_{n-1} + P_n\left[z_1 - \frac{\sqrt{n}\left(1-P_n\right)}{P_n}\frac{\sigma_0}{\sigma}z_2\right]^2\right. \\
&\qquad \left. \geq \frac{RSS}{\sigma^2} + P_n\left[z_1 - \frac{\sqrt{n}\left(1-P_n\right)}{\sigma P_n^{\frac{1}{2}}}\left(\bar{y}^{obs}-\theta_0\right)\right]^2\right\} \\
&= \Pr\left\{\chi^2_{n-1} \geq \frac{RSS}{\sigma^2} + P_n\left[z_1 - \frac{\sqrt{n}\left(1-P_n\right)}{\sigma P_n^{\frac{1}{2}}}\left(\bar{y}^{obs}-\theta_0\right)\right]^2\right. \\
&\qquad \left. - P_n\left[z_1 - \frac{\sqrt{n}\left(1-P_n\right)}{P_n}\frac{\sigma_0}{\sigma}z_2\right]^2\right\} \qquad (5.7)
\end{aligned}
$$

With this formulation of the *cppp* using our new discrepancy measure in a non-regression context, with σ^2 known, we can illustrate the distribution of the *cppp*.

Example. Suppose that $\theta \sim N\left(\theta_0, \sigma_0^2\right)$, where $\theta_0 = 5$ and $\sigma_0^2 = 9$. Also assume that $y_i|\theta, \sigma^2 \sim N\left(\theta, \sigma^2\right)$, so that $\bar{y}|\theta, \sigma^2 \sim N\left(\theta, \frac{\sigma^2}{n}\right)$, and that $n = 16$ and $\sigma^2 = 4$ (recall we assume that σ^2 is known). We want to calculate the *cppp* $\left(y^{obs}\right)$ from Equation (5.7). The easiest way is through simulation, adhering to the following steps:

1. Draw a value of θ from $\theta \sim N\left(5, 9\right)$. Then draw a value of $\bar{y}|\theta^*, \sigma^2$. Now draw a value of $\frac{RSS}{\sigma^2} \sim \chi^2_{n-1}$. Calculate $P_n = \frac{n\sigma_0^2}{\sigma^2+n\sigma_0^2}$. These are all draws for a single data set, and for a single *cppp* $\left(y^{obs}\right)$ value. These values stay constant over Step 2.

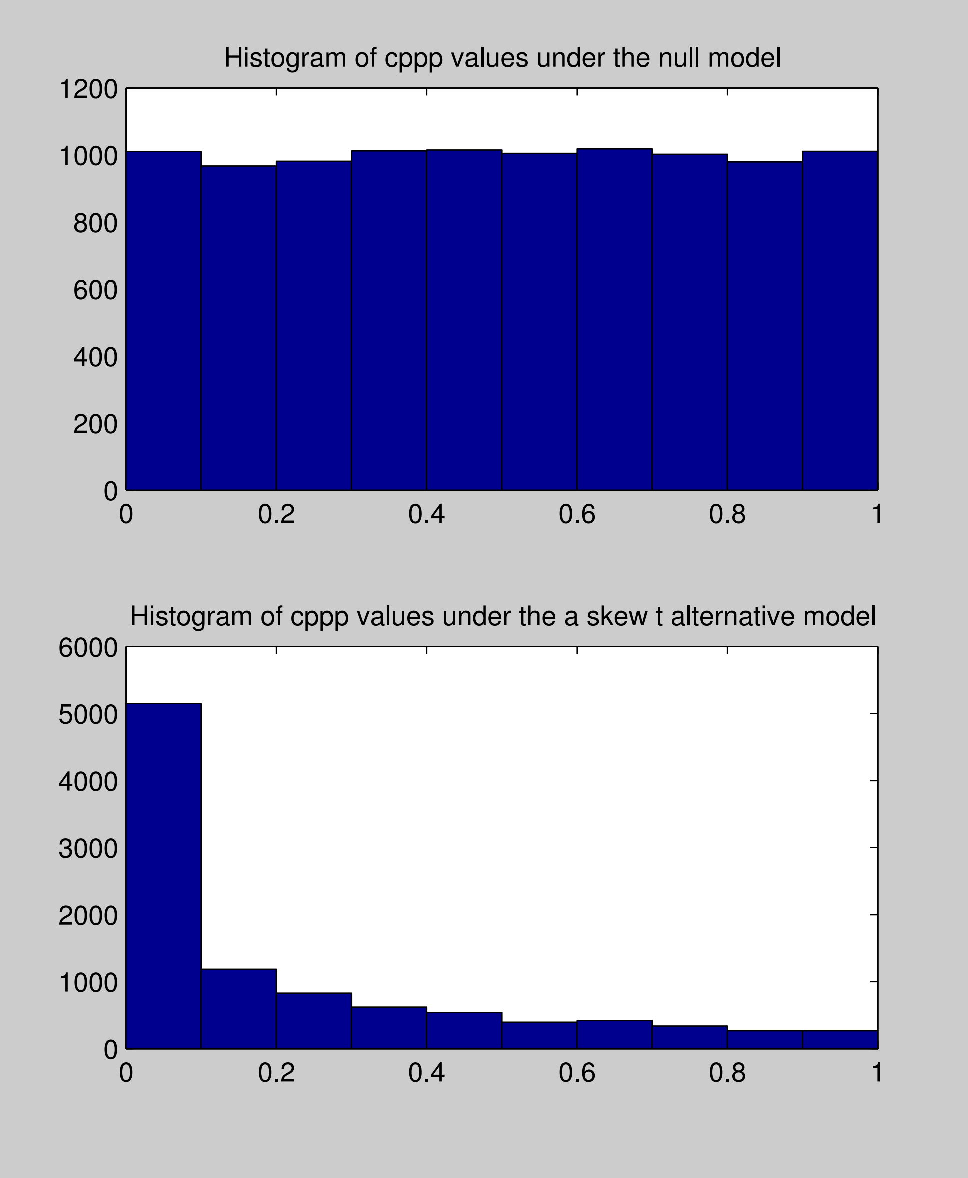
2. Draw, say, 10 000 values of z_1 and independently, 10 000 values of z_2, from $N\left(0, 1\right)$ distributions. Draw 10 000 draws from a χ^2_{n-1} distribution. Calculate the right-hand side of the inequality for *cppp* $\left(y^{obs}\right)$ from Equation (5.7) for each of the 10 000 sets of $\left(z_1, z_2\right)$ draws, and determine what proportion of these right-hand sides are greater than their corresponding χ^2_{n-1} draw. This gives a single *cppp* $\left(y^{obs}\right)$ value for the given simulated data.

3. Repeat Steps 1 and 2, say, 10 000 times, to obtain the distribution of the *cppp* $\left(y^{obs}\right)$ under the null model.

4. Repeat the entire process for alternative data (calculating the RSS and $\bar{y}$ under each simulation for each alternative) to determine distribution of the *cppp* $\left(y^{obs}\right)$ under alternative data scenarios (*i.e.* to investigate power).

Following these steps, we obtain a uniform distribution for the *cppp* $\left(y^{obs}\right)$ under the null model. If we assume $y_i|\theta, \sigma^2 = \theta + Z + \sigma t_5$, where $Z \sim N\left(0,1\right) I_{Z>0}$, then we have the following distribution of the *cppp*. These results are shown below in Figure 5.1.

Figure 5.1: Distribution of the *cppp* statistic under the null model, and a skew *t* alternative



With this simple result, it is clear that further research into the mathematics behind the *cppp* measure is warranted. Note that additional work not reported in this thesis was completed in this regard, with no success in finding a *cppp* formulation that was not reliant on the prior *and* uniform under the null distribution. As far as we can determine, the problem of not knowing the prior distribution is not alleviated using the suggestion by Hjort et al. (2006), O'Hagan's (1995) fractional Bayes prior, or Zellner's g-prior (as used in Liang, Paulo, Molina, Clyde & Berger 2008).

5.2.3 Monte Carlo approximation of the *cppp*

Since the *cppp* measure requires knowledge of the prior distribution on β, Cabras et al. (2011) use training samples to turn vague priors into proper priors for use in the *cppp* formulation. To approximate the *cppp*, we follow the steps outlined by Cabras et al. (2011, 434):

1. Draw a minimum training sample $\{y_t, X_t\}$, where y_t is a random sample of size $n_t = p + 1$ drawn from the observed data y^{obs}, and X_t are the corresponding rows of X.

2. Calculate *ppp* $\left(y^{obs}\right)$ using the following Monte Carlo sum:

   $$ppp\left(y^{obs}\right) = \frac{1}{J}\sum_{j=1}^{J} I\left\{D\left(Y_j^{rep}, \beta_j\right) \geq D\left(y^{obs}, Y_{m,j}, \beta_j\right)\right\} \tag{5.8}$$

   where β_j is drawn from the posterior distribution of β given the observed data, while the imputed values $Y_{m,j}$, and the replicated data Y^{rep}, are both drawn from the posterior predictive distribution.

   In essence, this step calculates how often the discrepancy measure D of the completed variable is greater than that of simulated responses.

3. Approximate *cppp* according to:

   $$cppp\left(y^{obs}\right) = \frac{1}{K}\sum_{k=1}^{K} I\left\{ppp\left(Y_{obs,k}\right) \leq ppp\left(y^{obs}\right)\right\} \tag{5.9}$$

   where $Y_{obs,k}$ are drawn from the posterior predictive distribution given the observed covariates X_{obs}, and β_k is simulated from the trained prior for β. For each $Y_{obs,k}$, calculate $ppp\left(Y_{obs,k}\right)$ using the second step above. First draw response observations from the prior predictive distribution for the cases with observed responses and replace these observed responses with the prior-drawn responses. Then generate (many times over) response observations for all cases (and response imputations for missing cases) from the posterior predictive distribution with observed responses equal to these prior-drawn responses. We count the proportion that the D from the replicated data is larger than that of the completed data. This gives us one $ppp\left(Y_{obs,k}\right)$. We compare this value to the proportion of times that the R^2 from the replicated data is smaller than that of the completed data when the completed data has its original responses (and not prior-drawn responses). Doing this comparison K times, each time D calculated from new prior-drawn 'observed' responses D compared to D calculated from actual observed responses, allows the approximation of the *cppp* (step 3).

The process can be filtered down to a programmable algorithm. The first three steps involve creating the $ppp\left(y\right)$ statistic which will be compared against the numerous $ppp\left(Y_{obs,k}\right)$ statistics.

1. From the posterior predictive distribution, generate a complete set of responses Y^{rep} and calculate D for these responses regressed on the explanatory variables.

2. From the posterior predictive distribution, generate a set of imputations to complete the response variable. Calculate D for this completed variable regressed on the explanatory variables.

3. Repeat steps 1 and 2, counting the proportion of times that the D from step 2 is less than D from step 1. This gives $ppp\left(y\right)$.

The following steps generate K separate prior-draws which allow multiple *ppp* comparisons with the *ppp* generated in step 3.

4. Draw a minimum training sample from the observed data

5. Create a proper prior distribution from the minimum training sample for the parameters, and then a proper prior predictive distribution, and draw from this distribution to replace all the observed responses.

Now repeat the initial three steps with the prior-replaced observed responses.

6. From the posterior predictive distribution, generate a complete set of responses Y^{rep} and calculate the D for these responses regressed on the explanatory variables.

7. From the posterior predictive distribution, generate a set of imputations to complete the response variable. Calculate the D for this completed variable regressed on the explanatory variables.

8. Repeat this process, counting the proportion of times that D from step 7 is less than D from step 6. This gives *ppp* $(Y_{o,1})$.

The last few step involves the iterating procedure that calibrates the *ppp*.

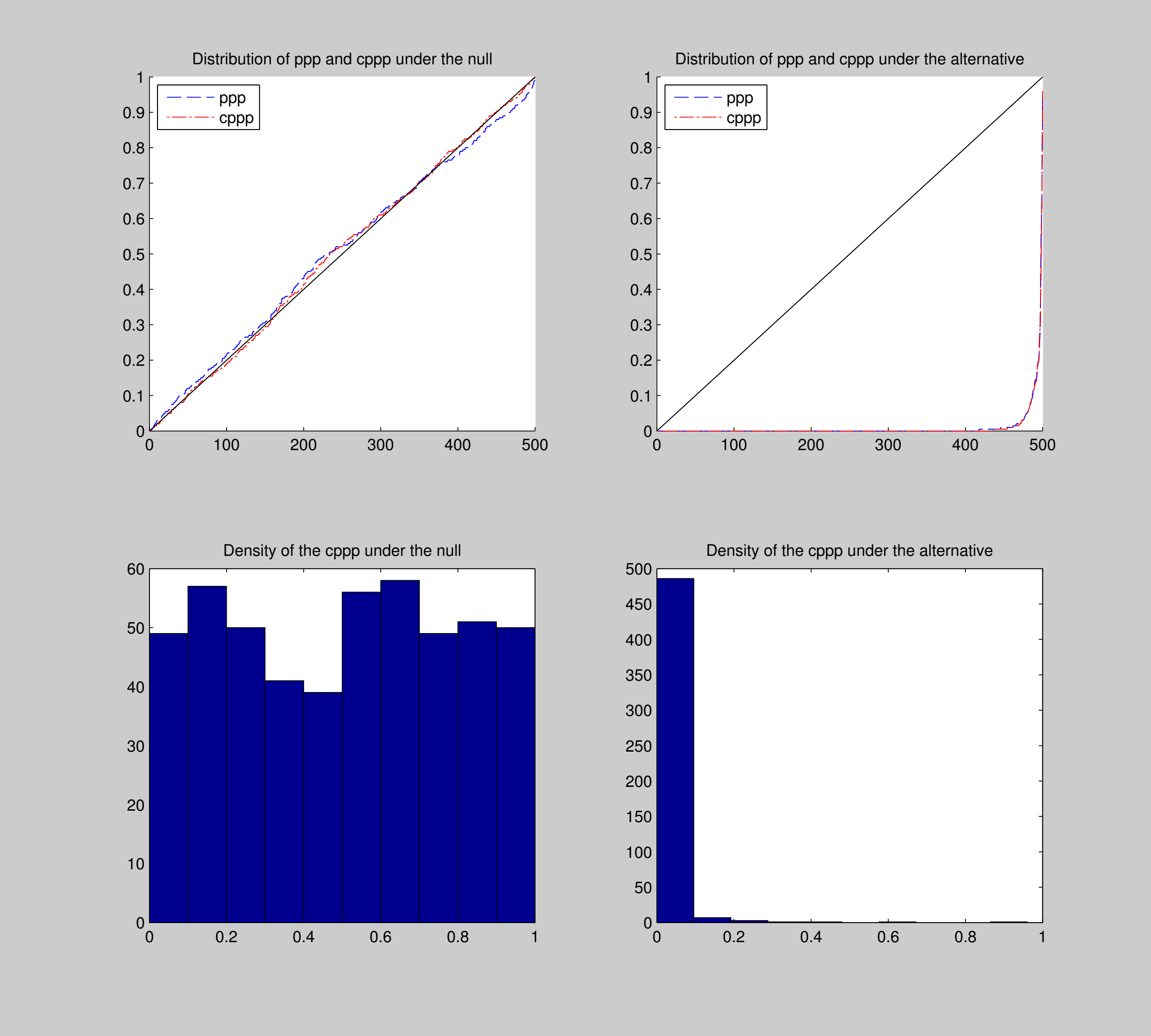
9. Repeat steps 4–8 K times, counting the proportion in which the *ppp* in step 8 is greater than the *ppp* in step 3. This proportion is the *cppp*.

One of the problems associated with this approximation is that the discrepancy measure must be something observable, such as the R^2 statistic, a function of the residuals, *etc.* We cannot easily use the new discrepancy measure suggested Subsection 5.2.2, *i.e.* $D(y, \theta) = \frac{1}{\sigma^2}(Y - X\beta)'(Y - X\beta)$. However, as was done in the previous section, we illustrate the distribution of the *cppp* under a null model and an alternative using the approximation given by Cabras et al. (2011). In this way, we can gauge whether there is merit in following up this line of research in the future.

Example. Suppose that $Y = 1 + X + \epsilon$, where $X \sim N(0,1)$, and $\epsilon \sim N(0,1)$ under the null model, and $\epsilon = 0.5Z + W$ with $Z \sim N(0,1)\, I_{Z>0}$ and $W \sim t_3$ under the alternative model. If we set the discrepancy measure to be the MSE of the deviation from Normality of the regression residuals, *i.e.* mean (Normal quantiles $-$ Empirical quantiles$)^2$, then we obtain the distribution of *ppp* and *cppp* as shown in Figure 5.2. It should be noted that this method of obtaining the *cppp* statistic is simulation-intensive, and requires hours of simulation for 500 *cppp* values as opposed to the minutes required for the 10 000 *cppp* values obtained in the previous example.

Immediately one will note that, although the *cppp* statistic is well calibrated and has high power, the *ppp* measure is also Uniform, implying that no calibration seems to have

Figure 5.2: Distribution of the MCMC *cppp* statistic under the null model, and a skew t alternative



been necessary. This is also true for the discrepancy measure that subtracts the smallest residual from the largest residual.[2]

Furthermore, it is clear that only discrepancy measures that are actually observable are candidates for this methodology. This is somewhat limiting, given the fact that the discrepancy measure used in the mathematical example is $D\left(y,\theta\right)=\frac{1}{\sigma^2}\left(Y-X\beta\right)'\left(Y-X\beta\right)$, and not just a function of the residuals, or some such observed quantity.

Clearly, more research needs to be done concerning this method of *cppp* approximation. The *ppp* statistic should be investigated in order to determine why it is Uniform in this instance.

[2] For brevity, this very similar result is not illustrated here.

5.3 Using the *cppp* in Multiple Imputation

The goal of this chapter was to find a way to determine objectively whether the Normal model within SRMI should be substituted for the more robust skew t alternative. In theory, this could be done by calculating the *cppp* statistic on both the incomplete continuous variable, assuming Normality, and after completing the variable using the Normal model. This two-stage *cppp* calculation would test the fit of the proposed model before and after imputation. If the statistic were above a particular threshold in either step, then the variable would have to be imputed or re-imputed using the more robust skew t alternative that is the focus of this thesis.

Of course, one could also try to develop the *cppp* statistic for the robust alternative. However, preliminary research into this area, using Cabras et al.'s (2011) *cppp* approximation algorithm, has not yielded any positive results yet. So far, we have not been able to obtain a Uniform *cppp* statistic for the skew t model. This means that there is potential for future research in this area; we are not suggesting that it cannot be accomplished — we are merely suggesting that additional methods of calibration of the *ppp* statistic need to be investigated.

5.4 Conclusion

This chapter has seen some mixed results. The *cppp* research field is rich with opportunity. Hjort et al. (2006) have provided the mathematical basis for further work in obtaining the distributional properties of the *cppp*, provided that we can somehow create proper priors. Cabras et al. (2011) also provide grounding for more work on MCMC approximation of the *cppp*, but the example shown in this section shows that the *ppp* measure is just as Uniform as the *cppp*. A more thorough investigation of the *cppp* approximation methodology is thus warranted as well.

One can note that the discovery of a method for finding a mathematical function or statistical distribution of the *cppp* statistic without a known proper prior would be a boon for this thesis. In essence it would allow imputers to use the Normal SRMI model on continuous data until the statistic moved out of its limits, implying a more robust model is necessary. In these cases, the faster Normal regression models could be replaced by the much slower, more flexible skew t model introduced in Chapter 4.

However, not having the statistic's distribution for the time being is not a disaster. As was shown in Chapter 4, the skew t SRMI model performs very well when the data is Normally distributed. This means that the robust model can be used as a default model until the *cppp* calculation is practicable.

Chapter 6

Skew Robit Model

6.1 Introduction

A categorical response model is a regression model in which the dependent variable can take on one of a set of values. The probit model, one type of binary response model, assumes that there is an underlying, latent variable (not observed), which indicates in which category each observation belongs. This underlying variable can be a function of the observed covariates, with a Normally distributed error. There are other models that can map $(-\infty, \infty)$ data to the $(0, 1)$ space, for example the logit link function, and the complementary log-log. This chapter builds on the Bayesian estimation processes of probit, set out by Albert & Chib (1993), the logit, set out by Groenewald & Mokgatlhe (2005), and the robit, set out by Liu (2005), and introduces a more robust model for the underlying latent variable, namely the model based on the skew adaptation of Student's t-distribution, this thesis' distribution of interest. Since the categorical response model is based on a (robust) skew t-distribution, it will henceforth be referred to as the strobit model. This study is concerned with estimation of the strobit model for both binary responses and ordinal categorical responses, the latter being an extension of the former.

The Bayesian estimation procedure does not actually model a categorical response variable as a function of the predictors. Rather, it models the latent variable as a function of the predictors. This implies that the estimated regression parameters have no meaning, except for classification and prediction purposes. For this study, however, this is of no concern, because the model can be used for the prediction of a category for a new observation (or of an observation with missing response category). Thus, the estimation method of this regression model is suitable for SRMI, for example. The goal of this chapter's study is to determine whether or not a more robust model for the underlying latent variable leads to better classification of new observations (observations with missing binary or ordinal responses) when the underlying latent variable is misspecified. Although this chapter does not directly deal with incomplete data, it is with little extra effort that the model introduced can be incorporated into an SRMI algorithm, even for multivariate missingness. Moreover, if the model works well in a univariate classification setting, then it will work well for an imputation setting, the latter simply being a classification of a case with unobserved response variable into a particular category for that response variable.

This chapter first reviews estimation procedures of the Bayesian probit model for binary and ordinal responses, as constructed by Albert & Chib (1993). We will then introduce methodology to estimate the parameters of a skew t-distribution, and incorporate this process into the estimation of the latent variable in the binary and ordinal response strobit models. Some practicalities will be discussed, after which the strobit model will be tested on simulated data. Conclusions will be drawn based on the comparison between the probit and strobit models after these models are applied to categorical data that is built on both latent Normal and non-Normal assumptions.

6.2 Bayesian Estimation of the Probit Model

In this section, we review probit and ordered probit estimation as laid out by Albert & Chib (1993). It is important to understand the MCMC simulation procedure for this method, since it will be adapted for estimation of the strobit and ordered strobit models. Additionally, the probit and ordered probit models are compared to the strobit and ordered strobit models, respectively, in the simulation study.

6.2.1 Two-category probit model

Consider a binary outcome vector Y, and covariate matrix X with rows $x_1, \ldots, x_n$. Introduce n latent variables (one for each observation), $W_1, \ldots, W_n$, where the W_i are independent $N(x_i'\beta, 1)$, and define $Y_i = 2$ if $W_i > 0$ and $Y_i = 1$ otherwise. It can be shown that the Y_i are independent Bernoulli r.v. with $p_i = P\left(Y_i = 2\right) = \Phi\left(x_i'\beta\right)$. So the joint posterior of the unobservables is:

$$\pi\left(\beta, W|y\right) \propto \pi\left(\beta\right) \prod_{i=1}^{n} \left(I_{W_i>0} I_{y_i=2} + I_{W_i\leq 0} I_{y_i=1}\right) \phi\left(W_i; x_i'\beta, 1\right),$$

where the vector y represents the observed categorical data, $\pi\left(\beta\right)$ is the prior on β, I is an indicator function that takes the value 1 on the subscripted condition, and 0 otherwise, and $\phi\left(W_i; x_i'\beta, 1\right)$ is the Normal density function for the variable W_i with mean $x_i'\beta$, and variance 1.

The conditional posterior distributions (using diffuse priors) are as follows:

$$\beta|y, W \sim N\left(\left(X'X\right)^{-1}\left(X'W\right), \left(X'X\right)^{-1}\right) \tag{6.1}$$

$$\begin{aligned} W_i|y, \beta &\sim N\left(x_i'\beta, 1\right) \text{ truncated at the left by 0 if } y_i = 2 \\ W_i|y, \beta &\sim N\left(x_i'\beta, 1\right) \text{ truncated at the right by 0 if } y_i = 1 \end{aligned} \tag{6.2}$$

Thus, for a Gibbs sampler to estimate draws from the joint posterior we use the following sequential procedure:

1. Initialise $\beta^{(0)}$ using the least squares estimate $\left(X'X\right)^{-1}\left(X'y\right)$.

2. Generate a vector W from Equation (6.2), given the preceding draw of β.

3. Generate a new vector β from Equation (6.1), given the preceding draw of W.

4. Repeat steps 2 and 3 until convergence of W and β.

6.2.2 Ordinal probit model

Albert and Chib (1993) also described an approach for Bayesian estimation of an ordered probit, similar to the two-category estimation procedure. The first category split (between categories 1 and 2), γ_1, is pinned down on the latent variable at 0, as before. The second split, γ_2 (to differentiate between categories 2 and 3), becomes an additional parameter to be estimated in the Gibbs sampler. Similarly, if there are more than three categories, each additional boundary, γ_j, on the underlying latent variable is another parameter to estimate within the Gibbs sampler.

In the case of the ordered probit, given that γ is a vector of the J category boundaries on the latent variable, the joint posterior of the unobservables is (with diffuse priors):

$$\pi\left(\beta, \gamma, W|y\right) \propto \prod_{i=1}^{n} \left\{ \left[\sum_{j=1}^{J} I_{Y_i=j} I_{\gamma_{j-1}<W_i<\gamma_j} \right] \phi\left(W_i; x_i'\beta, 1\right) \right\}$$

The conditional distributions for the $\gamma_j|W, Y$ are then Uniformly (UNF) distributed as follows:

$$UNF\left\{\max\left[\max\left(W_i : Y_i = j\right), \gamma_{j-1}\right], \min\left[\min\left(W_i : Y_i = j+1\right), \gamma_{j+1}\right]\right\} \tag{6.3}$$

The $\gamma_j|W, Y$ parameters are drawn before the W_i and the parameter estimates in the Gibbs sampler: So for a Gibbs sampler to estimate draws from the joint posterior we simply perform the following steps:

1. Initialise β using the least squares estimate $\left(X'X\right)^{-1}\left(X'y\right)$.

2. Generate category splits from Equation (6.3), with γ_1 fixed at a latent value of 0, given the previously generated W.

3. Generate a new vector W from Equation (6.2), given the preceding draw of W and the draws of the γ_j.

4. Generate a new vector β from Equation (6.1), given the preceding draw of W.

5. Repeat steps 2–4 until convergence of W and the parameters.

6.3 The Skew Student t-Distribution

We again follow the structure and estimation of the skew t-distribution presented in Chapter 4. For the sake of completeness, the base distribution and the required conditional distributions used in the Gibbs sampler for the strobit estimation are reiterated here.

Consider a linear regression model in which an observation vector $y = (y_1, \dots, y_n)'$ satisfies

$$y = X\beta + Z\delta + \epsilon$$

where $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ are the regression coefficients, δ is a skewness parameter, Z is a vector with elements $z_i > 0$, $i = 1, 2, \dots, n$ as skewness coefficients, $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$ is the error vector and $\epsilon_1, \dots, \epsilon_n$ are i.i.d. according to the Student-t distribution with location zero, scale parameter σ and ν degrees of freedom. Here $X = [x_1, \dots, x_n]'$ is the $n \times p$ matrix of explanatory variables and is taken to be full rank p. We denote the model parameters by $\theta = (\beta, \delta, \sigma, \nu) \in \mathbb{R}^{p+1} \times (0, \infty)^2$. The likelihood function is given by:

$$L\left(\beta, \sigma, \nu | y, X\right) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)^n \nu^{n\nu/2}}{\Gamma\left(\frac{\nu}{2}\right)^n \pi^{n/2}\sigma^n} \prod_{i=1}^{n} \left[\nu + \left(\frac{y_i - x_i'\beta - \delta z_i}{\sigma}\right)^2\right]^{-(\nu+1)/2}. \tag{6.4}$$

The likelihood for the t-distribution given in Equation (6.4) can be restructured as follows:

$$L \propto \prod_{i=1}^{n} \left(\frac{\lambda_i \tau}{2\pi}\right)^{\frac{1}{2}} \exp\left[-\frac{\tau}{2}\left(y_i - x_i'\beta - \delta z_i\right)^2\right] \times \prod_{i=1}^{n} \left[\frac{(\nu/2)^{\nu/2}}{\Gamma\left(\nu/2\right)} \lambda_i^{\nu/2-1} \exp\left(\frac{-\nu\lambda_i}{2}\right)\right] \tag{6.5}$$

where $\tau = \sigma^{-2}$ and the λ_i are weights indicating the influence of each observation on ν. Integrating out the λ_i in Equation (6.5) yields Equation (6.4).

6.3.1 Fitting the skew t-distribution

As was discussed in Section 4.2.1, when the t-distribution is used for errors on the posterior predictive distribution, generating the imputations is simply a matter of applying the posterior-drawn regression parameters to the covariates and adding an appropriate t error. As before, the challenge is to find the degrees of freedom for this error. This involves a

Gibbs sampling process for the parameters β, τ, $z_i, i = 1, \ldots, n$, δ, $\lambda_i, i = 1, \ldots, n$, and ν, while ν itself is drawn via a Metropolis-Hastings algorithm in each step of the Gibbs sampler. The Gibbs sampler requires the formulation of the conditional posterior distributions for each of the parameters of the model.

For each observation $i, i = 1, \ldots, n$, and covariate $q, q = 0, 1 \ldots, p$, $\tilde{y}_{iq} = y_i - \beta_{-q} X_{-q} - \delta z_i$, with $-q$ representing all variables in X besides variable q. In other words, for $q = 0$:

$$\tilde{y}_{i0} = y_i - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_p x_{ip} - \delta z_i$$

For $q = 1$:

$$\tilde{y}_{i1} = y_i - \beta_0 - \beta_2 x_{i2} - \beta_3 x_{i3} - \ldots - \beta_p x_{ip} - \delta z_i$$

For $q = 2, \ldots, p$:

$$\tilde{y}_{iq} = y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_{q-1} x_{i(q-1)} - \beta_{q+1} x_{i(q+1)} - \ldots - \beta_p x_{ip} - \delta z_i$$

Finally, for $q = p$:

$$\tilde{y}_{ip} = y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_{p-1} x_{i(p-1)} - \delta z_i$$

We also define $\tilde{\tilde{y}}_i = y_i - \beta x_i - \delta z_i$ separate to $\hat{y}_i = y_i - \beta x_i$, where x_i is the i^{th} row of the data matrix, corresponding to the covariates for observation i.

With skewness a part of the $\tilde{y}_{iq}$, the same conditional distributions exist for the β_q:

$$\begin{aligned}&\beta_q | y, \beta_{-q}, \tau, \Lambda \sim \\ &N\left\{\left(\tau \sum_{i=1}^{n} \lambda_i x_{iq}^2 + \frac{1}{\sigma_{\beta_q}^2}\right)^{-1} \left(\tau \sum_{i=1}^{n} \lambda_i x_{iq} \tilde{y}_{iq} + \frac{\mu_{\beta_q}}{\sigma_{\beta_q}^2}\right), \left(\tau \sum_{i=1}^{n} \lambda_i x_{iq}^2 + \frac{1}{\sigma_{\beta_q}^2}\right)^{-1}\right\},\end{aligned} \tag{6.6}$$

where x_{iq} is element (i, q) of the data matrix X (and when $q = 0$, $x_{i0} = 1$ for all i), and μ_{β_q} and $\sigma_{\beta_q}^2$ are the conjugate Normal prior mean and variance for β_q respectively. Once again, $\mu_{\beta_q} = 0$ and $\sigma_{\beta_q}^2 = 10000$.

For τ, we have that:

$$\tau|y,\beta,\Lambda \sim \Gamma\left\{\frac{n}{2}+a_\tau, \left(\frac{1}{2}\sum_{i=1}^{n}\lambda_i\tilde{\tilde{y}}_i^2+2b_\tau\right)^{-1}\right\}, \tag{6.7}$$

where a_τ and b_τ are the conjugate Gamma prior parameters for τ, and the matrix Λ is the diagonal matrix with diagonal elements $\lambda_1, \lambda_2, \ldots, \lambda_n$. However, for the case of the strobit and ordered strobit models, without loss of generality, τ is fixed at 1, just as σ is fixed at 1 in the formulation of the probit estimation of Albert & Chib (1993).

The conditional posterior for the $z_i, i = 1, \ldots, n$ is derived to be:

$$z_i|y,\beta,\tau,\delta,\Lambda \sim N\left\{\left(\tau\lambda_i\delta^2+1\right)^{-1}\tau\lambda_i\delta\hat{y}_i, \left(\tau\lambda_i\delta^2+1\right)^{-1}\right\}I_{Z_i>0}, \tag{6.8}$$

where $I_{Z_i>0}$ is an indicator function to ensure that only positive z_i exist (in order to make sense of the sign of the skewness parameter δ).

The conditional posterior distribution of the skewness parameter, δ, is given can be shown to be:

$$\begin{aligned}&\delta|y,\beta,\tau,\Lambda,z_1,\ldots,z_n \sim \\ &N\left\{\left(\tau\sum_{i=1}^{n}\lambda_i z_i^2+\frac{1}{\sigma_\delta^2}\right)^{-1}\left(\tau\sum_{i=1}^{n}\lambda_i z_i\hat{y}_i+\frac{\mu_\delta}{\sigma_\delta^2}\right), \left(\tau\sum_{i=1}^{n}\lambda_i z_i^2+\frac{1}{\sigma_\delta^2}\right)^{-1}\right\},\end{aligned} \tag{6.9}$$

where μ_δ and σ_δ^2 are the conjugate Normal prior parameters for δ.

For the λ_i, it can be shown that

$$\lambda_i|y,\beta,\tau,\nu,\delta,z_1,\ldots,z_n \sim \Gamma\left\{\frac{1}{2}(\nu+1), \left[\frac{1}{2}\left(\tau\tilde{\tilde{y}}_i^2+\nu\right)\right]^{-1}\right\}, \tag{6.10}$$

with the skewness built into the distribution by replacing $\hat{y}_i$ with $\tilde{\tilde{y}}_i$.

The posterior for ν, conditional on Λ, and its priors, are given in the following equations.

$$p(\nu|y,\Lambda) \propto \frac{\nu^{\frac{1}{2}\nu n}}{2^{\frac{1}{2}\nu n}\left[\Gamma\left(\frac{\nu}{2}\right)\right]^n}|\Lambda|^{\frac{1}{2}\nu-1}\exp\left[-\frac{1}{2}\nu\sum_{i=1}^{n}\lambda_i\right]p(\nu), \tag{6.11}$$

with the prior on ν taking one of four forms, namely the truncated exponential, the Independence Jeffrey's prior, the probability-matching prior or reference priors for the orders (ν, μ, σ^2), (ν, σ^2, μ), and (μ, ν, σ^2), and the reference priors for the orders (μ, σ^2, ν), (σ^2, μ, ν), and (σ^2, ν, μ). In this study the Independence Jeffrey's prior is used. We find that this prior is less restrictive on the degrees on freedom than the well-established exponential prior, within the context of the strobit estimation.[1] It is shown by Fonseca et al. (2008) that the independence Jeffreys prior is

$$p_{IJEFF}(\nu, \beta, \sigma) \propto \sigma^{-1} \left(\frac{\nu}{\nu+3}\right)^{\frac{1}{2}} \left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+3)}{\nu(\nu+1)^2}\right]^{\frac{1}{2}}$$

assuming that the marginal priors for β and (σ, ν) are independent *a priori*. Once again, $\psi'(\cdot)$ is the trigamma function.

Working with the natural log posterior and log priors is easier:

$$\begin{aligned}\log(p(\nu|y,\lambda)) \propto{}& \frac{1}{2}\nu n \log(\nu) - \frac{1}{2}\nu n \log(2) - n \log\left(\Gamma\left(\frac{\nu}{2}\right)\right) \\ & - \left(\frac{1}{2}\nu - 1\right) \sum_{i=1}^{n} \log(\lambda_i) - \left(\frac{1}{2}\nu - 1\right) \sum_{i=1}^{n} \lambda_i - \log\left[p_{IJEFF}(\nu, \beta, \sigma)\right].\end{aligned} \tag{6.12}$$

The algorithm for the Gibbs sampler (and Metropolis sampler for ν) when we wish to incorporate skewness into the imputation model utilises the conditional distributions listed above.

6.3.2 Strobit and ordered strobit model estimation

Now that we have the conditional distributions of the parameters of the skew t-distribution, we can use these draws in the place of the draws of probit parameters, namely the β.

Once more, after initialising the parameters, if the strobit is estimating a two-category response variable, the category split on the latent variable is set at 0. Otherwise, similarly

[1] Of the priors derived in this thesis, the independence Jeffreys prior seemed to allow for ν draws closer to the true degrees of freedom. In the overall simulation study, however, the choice of prior made little difference, if any, to the final results.

to the ordered probit estimation, the first category split, γ_1, is set at 0, while the remaining category splits, the γ_j, become parameters to be estimated in the same way as for the ordered probit, namely their conditional distribution follows Equation (6.3).

Given all the other unknowns, we can draw bounded latent variables as follows:

$$\begin{aligned} W_i|y,\beta,\tau=1,\delta,Z,\nu &\sim t_\nu + X\beta + Z\delta \\ &\quad \text{truncated at the left by 0 if } y_i = 2 \\ W_i|y,\beta,\tau=1,\delta,Z,\nu &\sim t_\nu + X\beta + Z\delta \\ &\quad \text{truncated at the right by 0 if } y_i = 1 \end{aligned} \tag{6.13}$$

Thus, for a Gibbs sampler to estimate draws from the joint posterior we us the following algorithm:

1. Initialise β using the Gibbs sampler with y as the dependent variable, follow up with initialisation of all the other parameters.

2. Set the category split at a latent value of 0 in the case of 2-category response variable, or, in the case of the ordered response variable, draw the γ_j variables from Equation (6.3), given the preceding draw of W.

3. Generate a vector W from Equation (6.13).

4. Step through the Gibbs sampler for conditional draws from Equations (6.6)-(6.11), each paramater based on preceding draws of the other parameters.

5. Repeat steps 2–4 until the draws for W and the parameters converge.

Some practicalities

If we follow the algorithm above, then the draws for the parameters (β in particular) vary widely from round to round. Theoretically, the draws should be stable, but the variance in the draws makes any prediction based on a single draw in Step 4 rather unreliable. In other words, since the drawn parameter values vary widely from one draw to the next, a prediction based on one particular draw might differ drastically from a prediction based on the very next draw in the sampling sequence. In order to stabilise the draws, Step 4 of the above procedure is repeated several times, say 200 times, until a conservative set of draws for the parameters of the skew t-distribution is evident.

While this modification of the algorithm considerably increases its running time, the modification is necessary if the fitting algorithm is to be used for prediction. In prediction, only a single draw from the end of the Gibbs sampler is used, and if the variation from one draw to the next is very high, one is likely to obtain drastically different coefficient estimates from one run of the fitting procedure to the next.

Through thorough investigation, we are satisfied that this extra smoothing step does not detract from the implementation of the strobit model except in the case where there are time constraints for the fitting procedure.

6.4 Simulation Study Methodology

6.4.1 Simulated data

In order to assess the robustness of the probit and strobit models, and their ordered counterparts, four difference latent data construction scenarios are examined: Normal, skew t, Exponential and Uniform. We assume $U = -1 + 4x + \xi$, where U is the true latent variable, $x \sim N(0.5, 1)$, and ξ is an error that depends on the data scenario under question:

1. Normal data: $\xi_i \sim N(0,1), i = 1, \ldots, n$;
2. Skew t data: $\xi_i = -2z_i + 0.5w, i = 1, \ldots, n$, where $z_i \sim N(0,1)I_{z_i>0}$, in other words, the z_i are positively truncated Normal random variables, and $w_i \sim t_5$;
3. Exponential data: $\xi_i \sim \text{Exp}(1), i = 1, \ldots, n$ or $\xi_i = -\ln(1-u_i)$, where $u_i \sim U(0,1)$;
4. Uniform data: $\xi_i, i = 1, \ldots, n$, is a random integer between 0 and 5.

Once the latent data is generated, the observations are allocated to categories based on latent data using random category splits in the full simulation analysis, or splitting point(s) -2 (and 2) for the two-category (three-category) single simulation discussion.

Two sample sizes are considered in the full simulation study, namely $n = 200$ and $n = 1000$, but for the single simulation analysis, the review is restricted to $n = 1000$.

6.4.2 Assessment Methods

The primary method of assessing the probit and strobit models, as well as their ordered counterparts, is using the mean absolute deviation (MAD) of the predicted category values from their actual category values given a new sample for a particular data scenario. This criteria is essentially a summary of the classification matrices across multiple simulations within each data scenario. In brief, we proceed using the following steps:

1. Generate latent data dependent on an exogenous Normal random variable, x, an intercept, and an error appropriate to the data scenario under examination.

2. Split the latent data at random points to generate a categorical variable (ensuring that each category contains at least 2% of the sample).

3. Estimate parameter values for the (ordered) probit and strobit models on the given simulated data, using the average of 300 draws from the Gibbs sampler, after a burn-in of 300 draws. Within the strobit estimation, the smoothing process also burns in 50 draws of the skew t-distribution parameters within each of the 600 strobit Gibbs sampler runs.

4. Generate a new sample according to the same latent data scenario of step 1.

5. Using the random splits generated in step 2, re-split the new sample into categories.[2] These categories are the ‘correct’ categories for the new sample.

6. Using the estimated parameter values for the regression model estimated in step 3, predict a latent value for each observation in the new sample, drawing random Normal errors for the probit and ordered probit predictions, and skew t errors for the stobit and ostrobit models. [3]

[2] It can be noted that in some instances, this procedure led to one category containing all the observations. These cases were not eliminated, since the model could still theoretically predict an observation outside of the category bounds containing all these observations, leading to classification error.

[3] Symmetric t errors combined with a aero-truncated Normal error for skewness

7. Using the latent predictions and the estimated category splits from the model estimation, re-categorise the new sample. These categories are the predicted categories of the new sample.

8. Calculate the MAD for a model by averaging the absolute difference between actual and predicted categories of the new sample.

9. Repeat steps 1–8 for a total of 200 simulations.

6.5 Simulation Study Analysis

In this section, a single simulation across all data scenarios is scrutinised, and then the process is repeated for a total of 200 runs for a thorough assessment of the methodology.

6.5.1 Single-run analysis

In order to understand the simulation analysis, the histograms of the data, as well as histograms for the errors that are added to the exogenous covariates, are presented in Figures 6.1 and 6.2, for a two- and three-category simulation, respectively. From these figures, it is clear that the latent data is modelled as a regression on a Normal covariate and an intercept, and is coupled with varying errors, including Normality (scenario 1), negative skewness (scenario 2), positive skewness (scenario 3), and uniformity (scenario 4). The probit and strobit Gibbs sampler draws for the two-category model estimation (after burn-in) are shown in Figures 6.3 and 6.4. These parameter draws are particularly stable, except for the degrees of freedom, ν, for the strobit estimation. The probit and strobit Gibbs sampler draws for the three-category model estimation (after burn-in) are shown in Figures 6.5 and 6.6. One will notice in these figures that there is sometimes drift in both the γ_1 value and a β value. This drift is not much of a concern as long as the draws drift together — one cannot pin more than one category boundary down without severely limiting the estimation procedure. One might argue that for three categories, one could fix the category boundaries and hope that the sampler is long enough to squeeze and move the underlying latent model to correctly fit the data, but beyond three categories this would be unrealistically strict. In any case, the drift of the strobit parameter pairs is not entirely a problem, since we are not using the estimation procedure for interpretation of fit parameters, but merely for prediction (and classification). Parameter pair drift will not affect this goal.

Figure 6.1: Latent data under the four data scenarios for the 2-category analysis, with embedded errors; $n = 1000$.

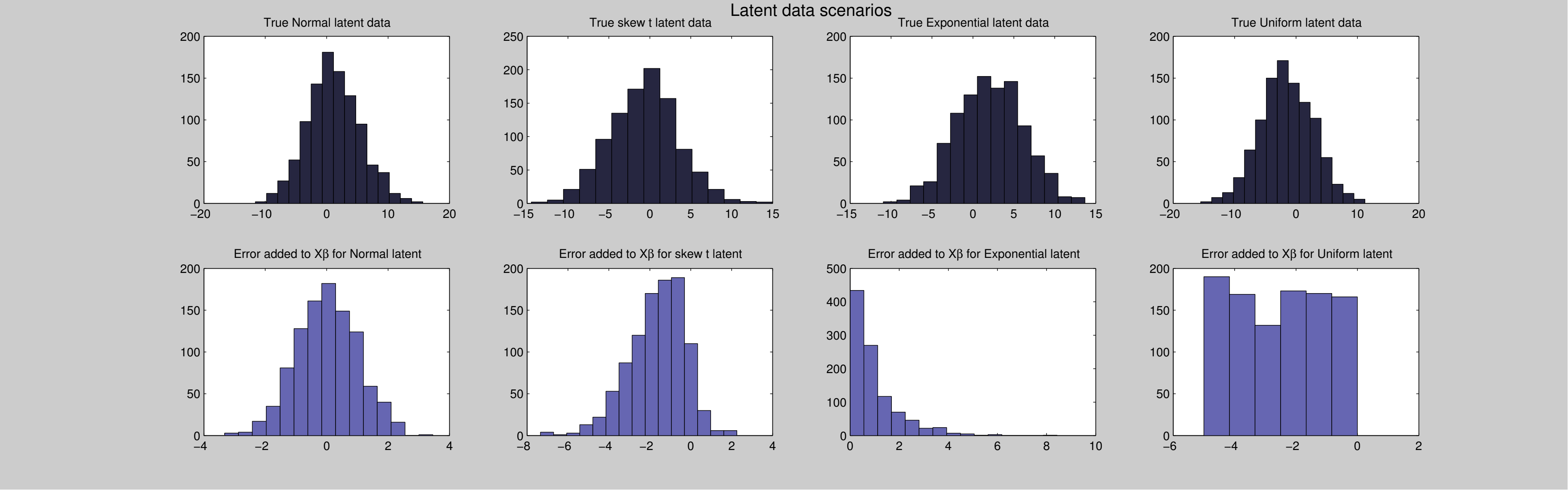


Figure 6.2: Latent data under the four data scenarios for the 3-category analysis, with embedded errors; $n = 1000$.

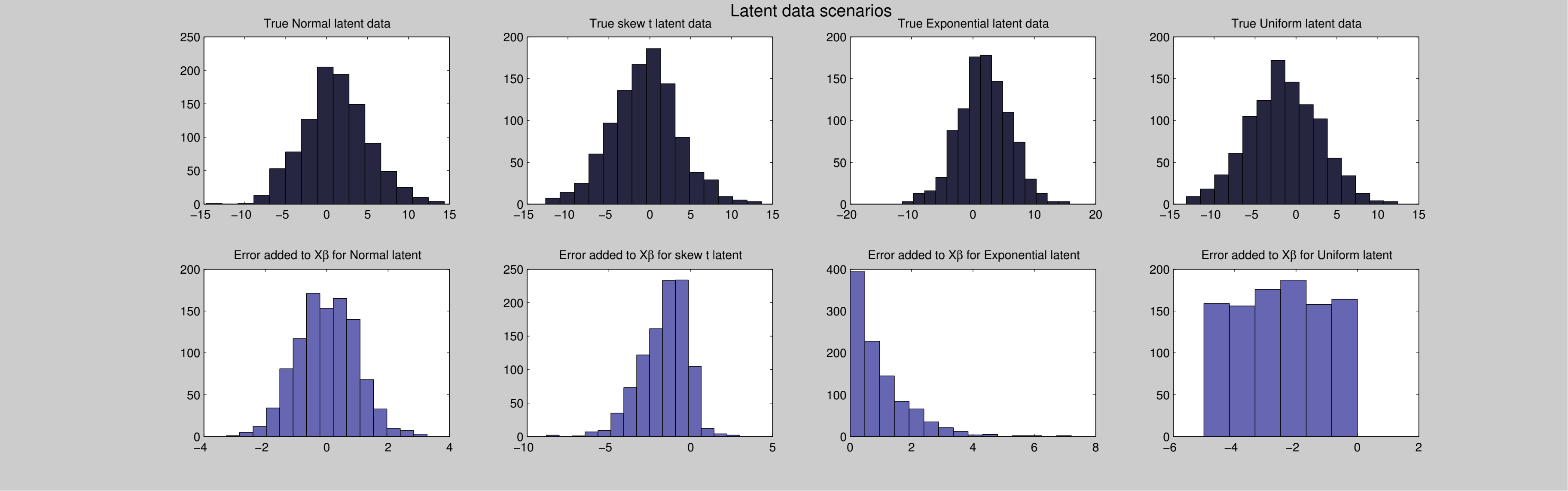


Figure 6.3: Probit Gibbs sampler draws after burn-in for the 2-category data

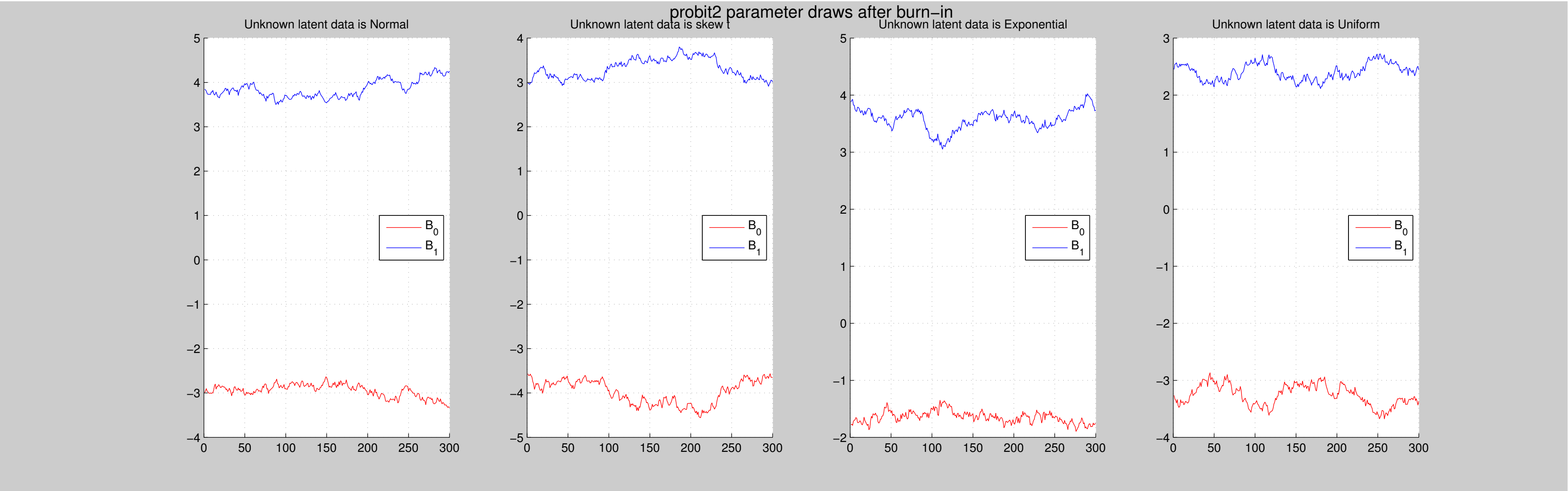


Figure 6.4: Strobit Gibbs sampler draws after burn-in for the 2-category data

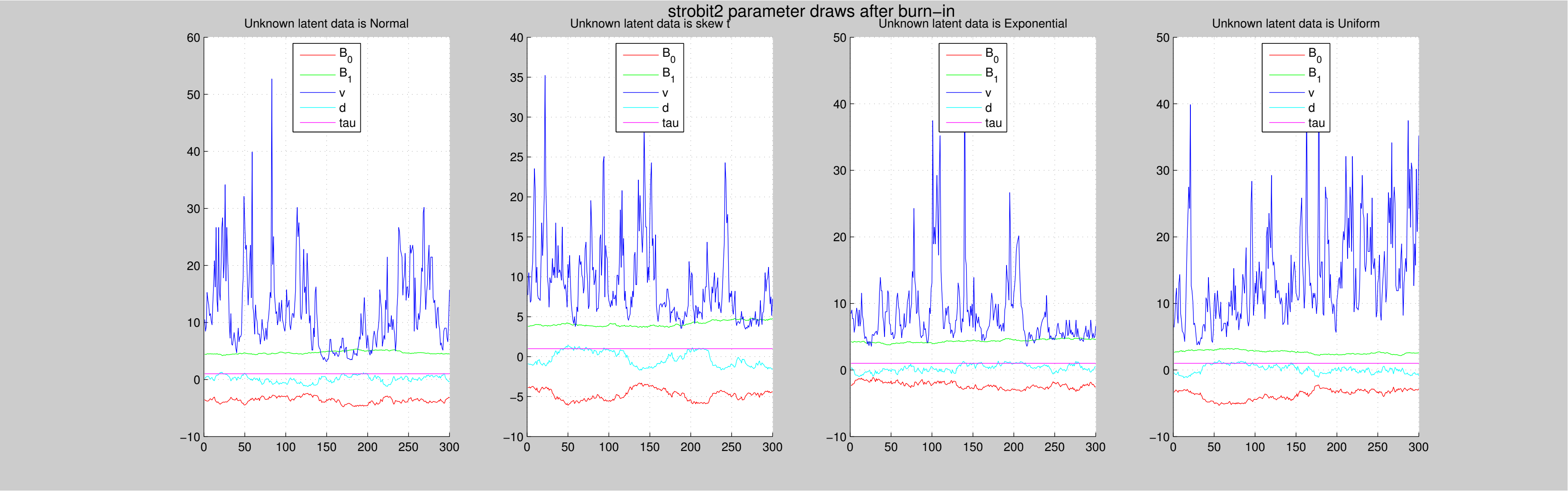


Figure 6.5: Probit Gibbs sampler draws after burn-in for the 3-category data

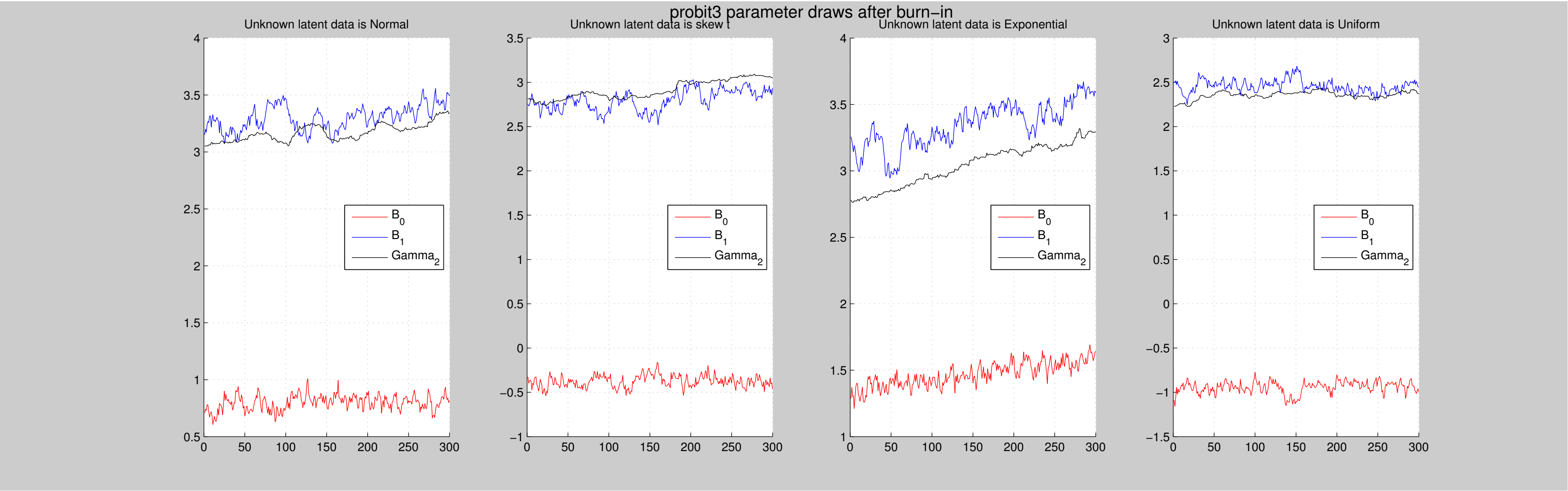
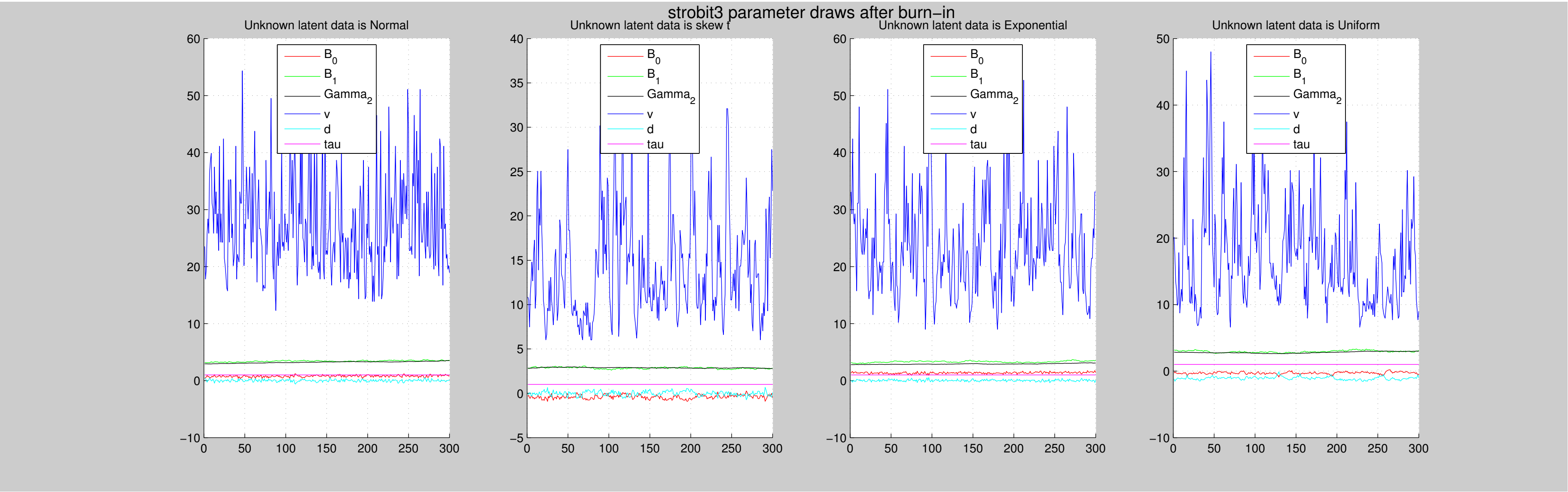


Figure 6.6: Strobit Gibbs sampler draws after burn-in for the 3-category data



Once the probit and strobit models are fitted under each data scenario, the fitted latent distributions are graphed in Figure 6.7 for two categories and Figure 6.8 for three categories. The different shades indicate the different sequentially observed categories. Note that the fitted latent data is forced to be separated by category, leading to multi-modal distributions. One would hope that the estimation algorithms would lead to smooth, uni-modal fitted distributions, but this is not the case, even for the probit on Normal data.

Once the models are estimated, a new sample is drawn according to the appropriate data scenario, and the estimated models are used to predict a new distribution of the latent data. Histograms of these distributions are given in Figure 6.9 for two categories, and Figure 6.10 for three categories, and are shaded according to the categories that the new sample's observations would be assigned to had the underlying model been known. It is clear from these figures that there is no way of splitting all the new observations using their predicted latent data into their correct categories. This leads to classification error. A visual representation of the classification matrix for the three-category simulation is given in Figure 6.11.

For the two simulations represented in the graphs, we have the following classification errors for the new samples: for two categories, the probit has MAD errors of 18%, 18.8%, 20.6% and 14.9% for the Normal, skew t, Exponential and Uniform data scenarios respectively, while the srobit has MAD errors of 13%, 9.9%, 13.7% and 15.6% for the four data scenarios, respectively; for three categories the probit has MAD errors of 32.1%, 37.7%, 27.2% and 41% for the Normal, skew t, Exponential and Uniform data scenarios respectively, while the srobit has MAD errors of 21.6%, 24.5%, 19.6% and 28% for the four data scenarios, respectively. These figures have little value without repeating the simulation process multiple times, as is carried out in the next section.

6.5.2 Multiple-run analysis

The initial simulation analysis, summarised in Table 6.1, seems promising for the strobit model. In all simulation scenarios, across two and three categories, sample sizes of both 200 and 1000, and across all four data scenarios, the strobit model's MAD error is more often than not lower than that of the probit model's MAD error.

Table 6.1: MAD error superiority proportions, by category, sample size, and data scenario

Categories	Sample size	Data scenario	Probit better	Models equal	Strobit better
2	**200**	Normal	20.5%	33.5%	46.0%
		skew *t*	12.5%	41.0%	46.5%
		Exponential	16.5%	40.5%	43.0%
		Uniform	15.5%	43.0%	41.5%
	1000	Normal	28.0%	27.5%	44.5%
		skew *t*	18.5%	31.0%	50.5%
		Exponential	18.5%	38.0%	43.5%
		Uniform	20.5%	39.0%	40.5%
3	**200**	Normal	20.5%	33.5%	46.0%
		skew *t*	12.5%	41.0%	46.5%
		Exponential	16.5%	40.5%	43.0%
		Uniform	15.5%	43.0%	41.5%
	1000	Normal	41.0%	4.5%	54.5%
		skew *t*	39.5%	7.0%	53.5%
		Exponential	42.5%	9.5%	48.0%
		Uniform	29.5%	8.5%	62.0%

However, upon further analysis, the strobit model loses some of its favour. The first problem becoming evident is the number of times within the multiple simulation procedure that MAD errors from the probit and strobit models are the same. In Figures 6.12 – 6.15, we plot the difference between probit and strobit MAD errors against a measure of tail-category sparseness or observation-scarcity, namely the negative sum of the natural logs of the proportions of observations in the tail categories.[4] We find that in the two-category case the probit and strobit models are misclassifying the same proportions when tail scarcity is high, *i.e.* the strobit model is not doing better than the probit in classifying observations into the correct categories when those categories are sparsely-populated tail categories. This is quite a concern, since the strobit model, with an underlying heavy-tailed skew distribution, might naturally be thought of as more capable in this context.

Another issue becomes apparent when one examines the difference in MAD errors between the probit and strobit models across multiple simulations: this difference is not significantly greater than zero, *i.e.* while it is evident that the strobit model's MAD error is often smaller than that of the probit, the average strobit MAD error is not significantly lower than that of the probit. This is illustrated in Figures 6.16 – 6.19. The empirical 95% intervals for the probit minus strobit MAD error crosses over zero for all data scenarios

[4] Or simply negative sum of the natural logs of the proportions of observations in both categories in the two-category case.

under both the two- and three-category cases.

Apart from a general review of the analyses performed, one can note a few interesting results from this study. It is clear that the strobit model performs better classifications that the probit when the latent data is Normally distributed. This is a strange result, but can be partially explained by the fact that the Normal distribution is a special case of the skew t. Also, for the two-category scenarios, the strobit does perform better than the probit model when the scarcity measure is not extremely high, but moderately large (Figures 6.12 and 6.14).

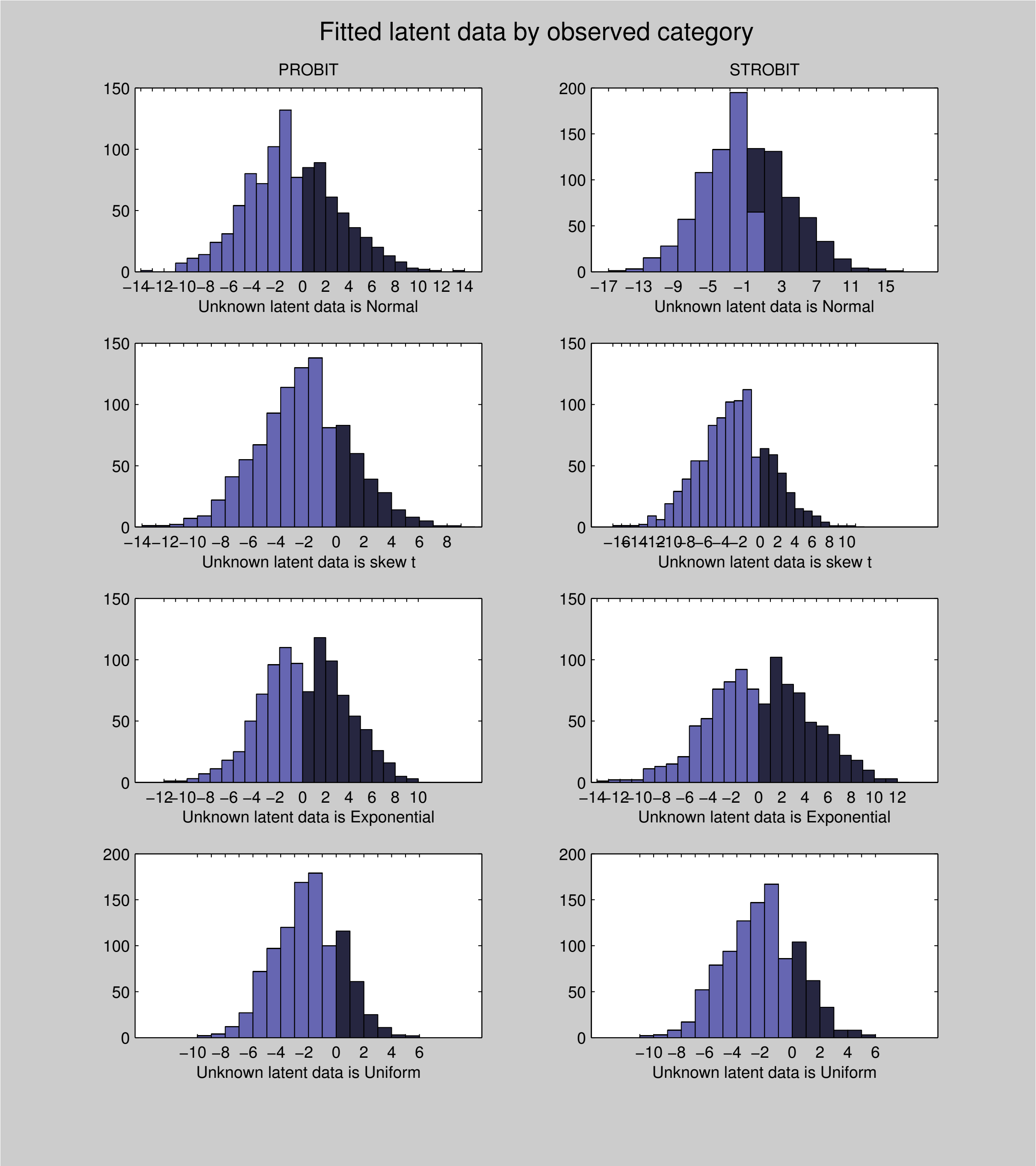
6.6 Conclusion

This chapter introduces a new robust Bayesian procedure for modelling ordinal categorical response data as a function of exogenous covariates. The work is based on the Bayesian estimation processes of probit, as explained by Albert & Chib (1993), the logit, as explained by Groenewald & Mokgatlhe (2005), and the robit, as explained by Liu (2005). The modelling procedure expands on the existing literature by assuming that the (ordinal) categorical responses are linked to skew t-distributed latent data — this model is then called the strobit model. Procedures are introduced to estimate the parameters of this Bayesian strobit model. Since the strobit model fits more parameters than the probit and logit, the estimation procedure can be quite time consuming, and some practicalities associated with this process are discussed. It is noted that the Bayesian estimation procedures of categorical response models such as the probit, logit and strobit, produce parameters that are linked to unknown underlying latent data, and are thus not useful for interpretation, but only for prediction or classification of new observations.

The probit and strobit model are compared under two-category and three-category binary responses based on simulated latent data with various characteristics. The strobit model performs marginally better than the probit under all data situations (even when the latent data is Normally distributed), but the difference in performance between the two models is not significant. However, since the Normal distribution is a special case of the skew t-distribution, and hence the probit is a special case of the strobit, and since the strobit performs marginally better than the probit model under varying data scenarios (including Normality), the authors recommend that if computing time is not of great concern to a modeller, the Bayeian estimation of the strobit model be used in place of the Bayesian estimation of the probit.

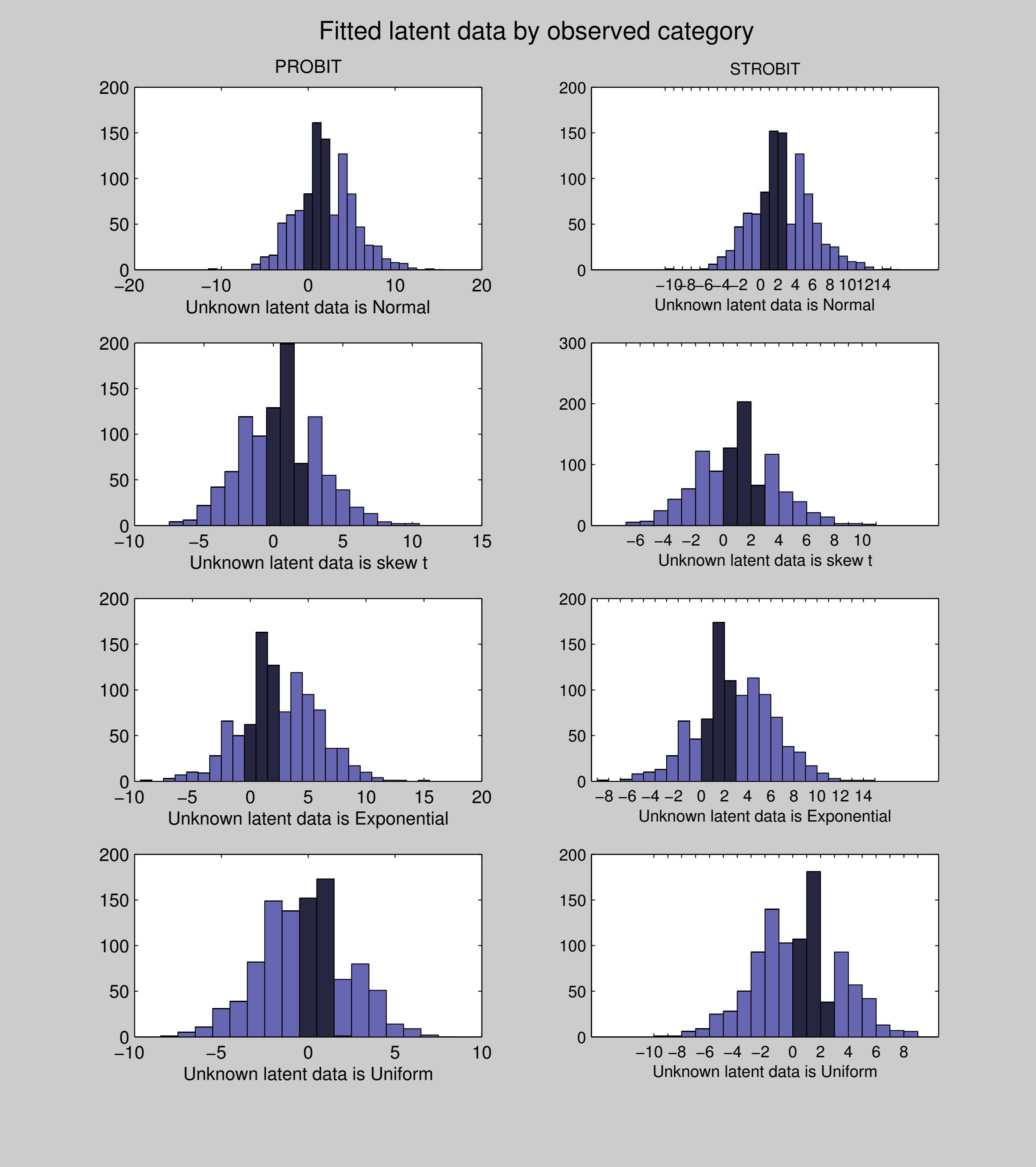
Naturally, this study opens up various topics for further research. The strobit model is built for implementation in sequential regression multiple imputation (SRMI), and thus further research into the applicability of this model in that context is warranted. Moreover, this model should be compared with other categorical imputation procedures, such as the multinomial model that is commonly used in SRMI, or other categorical response models. Also, the fact that the strobit model, as it is defined in this chapter, does not significantly improve the classification results on tail categories with low counts, is a concern. Perhaps a skew t model with more allowance for skewness could be examined (*i.e.* $z_i \sim t_3 I_{z_i>0}$ instead of $z_i \sim N(0,1)I_{z_i>0}$ could be used). As far as the estimation procedure itself is concerned, work should be done to speed up the Gibbs sampler, as well as stabilise the actual sampling. This research required sampling-in-sampling to obtain stabilised parameter estimates, and research can be done on the suitability and efficiency of such a procedure.

Figure 6.7: Fitted latent data (2 categories)



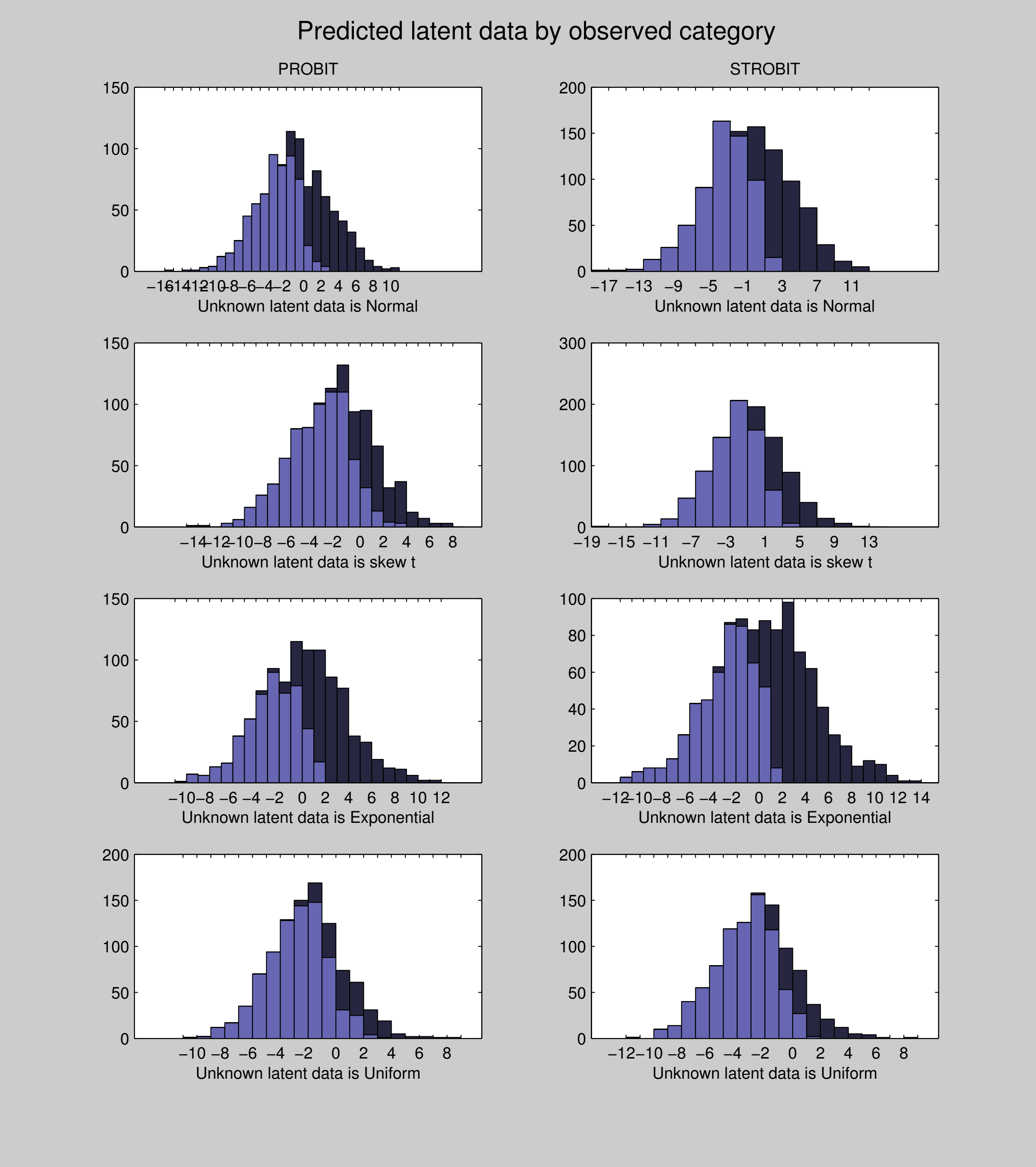
The different shades represent the two different observed categories. The latent data is separated for each category by the fitting algorithm.

Figure 6.8: Fitted latent data (3 categories)



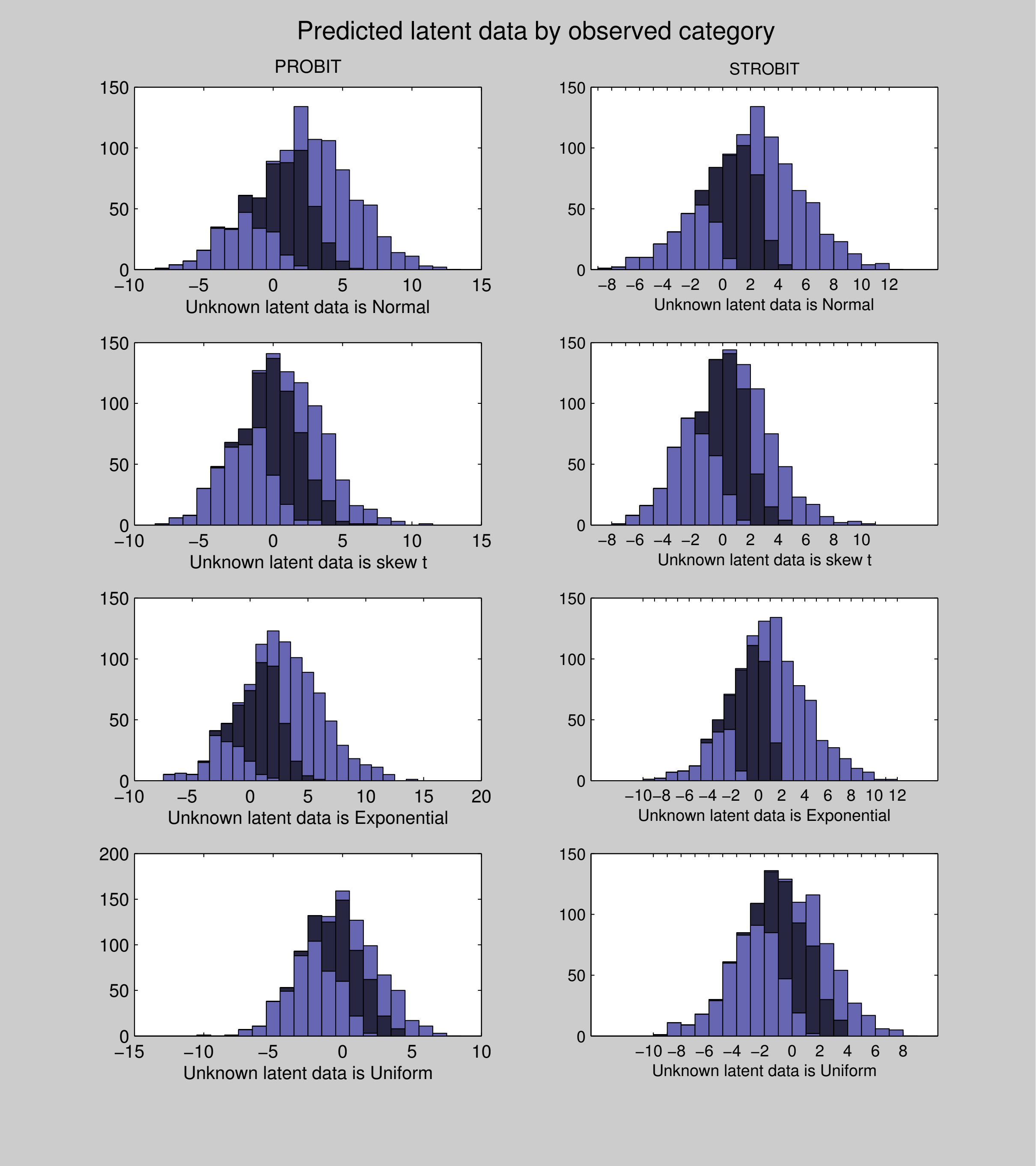
The different shades represent the three different observed categories. The latent data is separated for each category by the fitting algorithm.

Figure 6.9: Predicted latent data (2 categories)



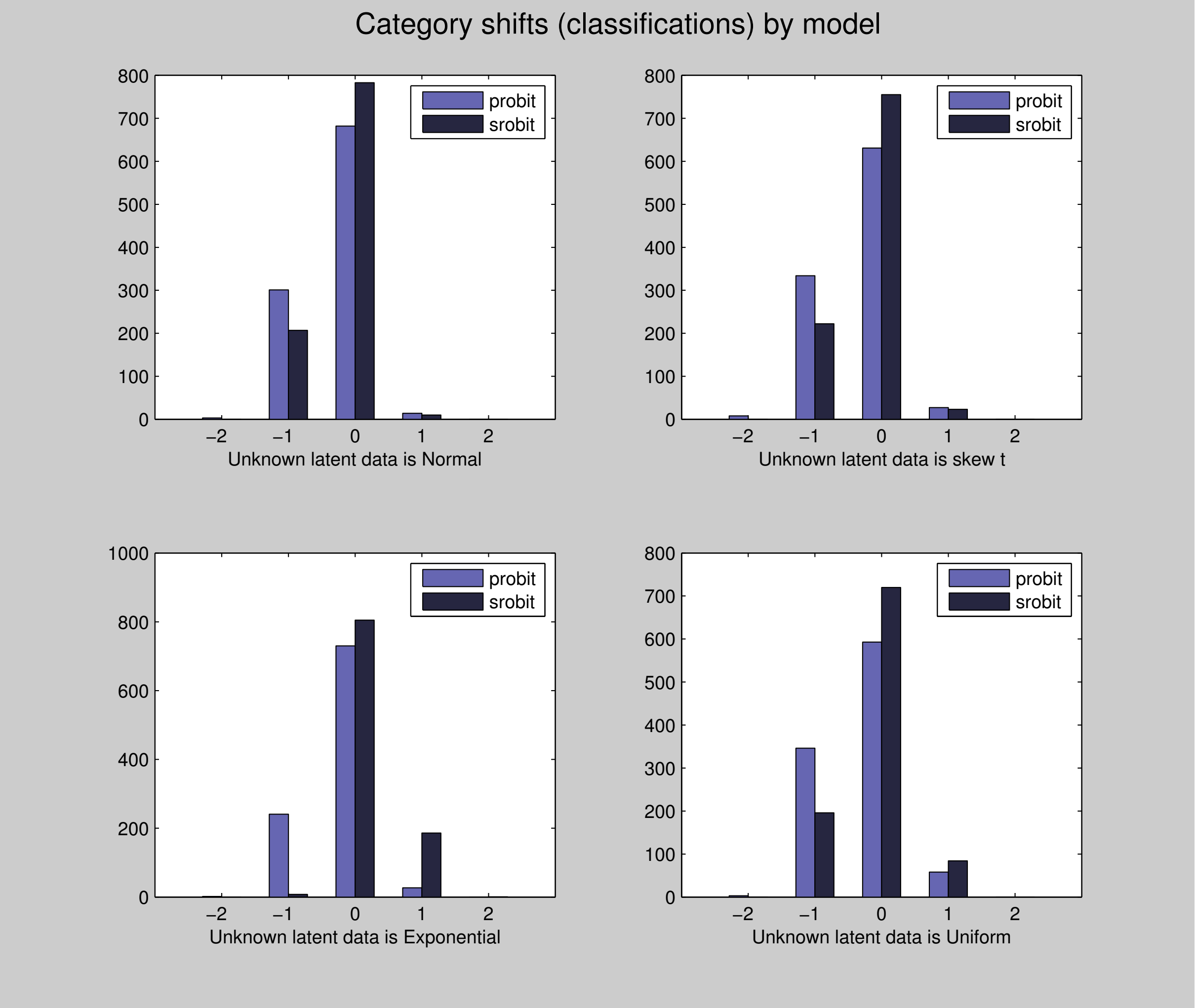
The different shades represent the two different actual observed categories, and not the categories that are chosen off the predicted latent data.

Figure 6.10: Predicted latent data (3 categories)



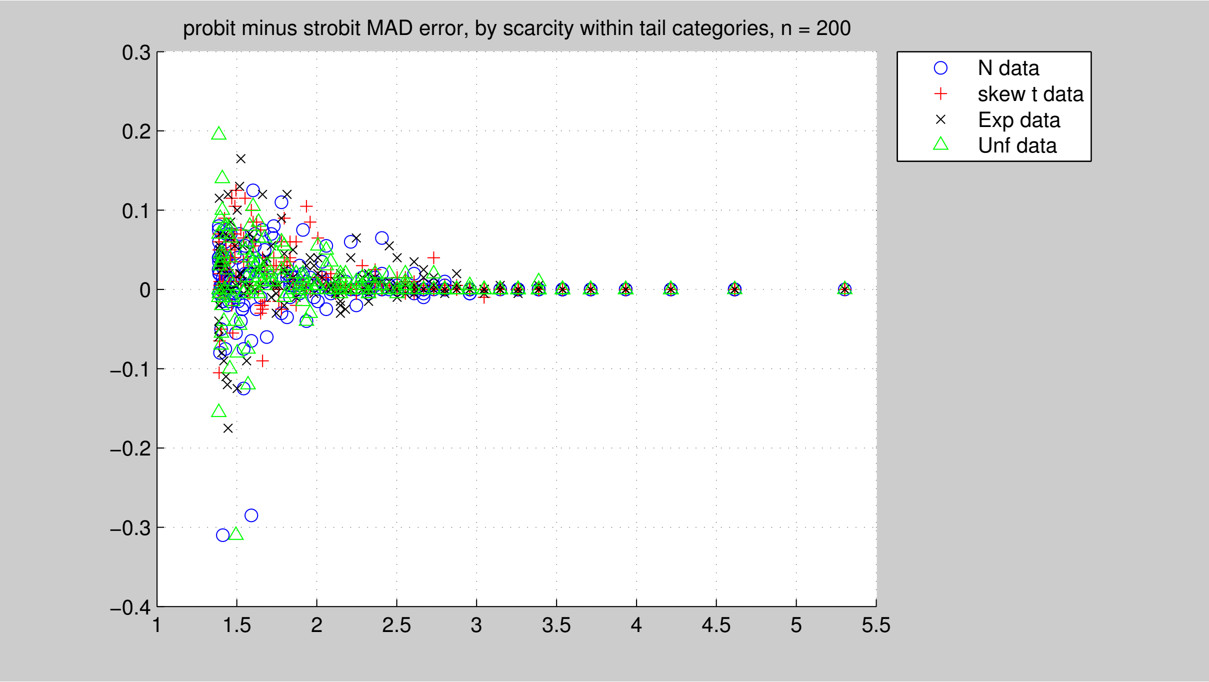
The different shades represent the two different actual observed categories, and not the categories that are chosen off the predicted latent data.

Figure 6.11: Classification errors for 3-category simulation



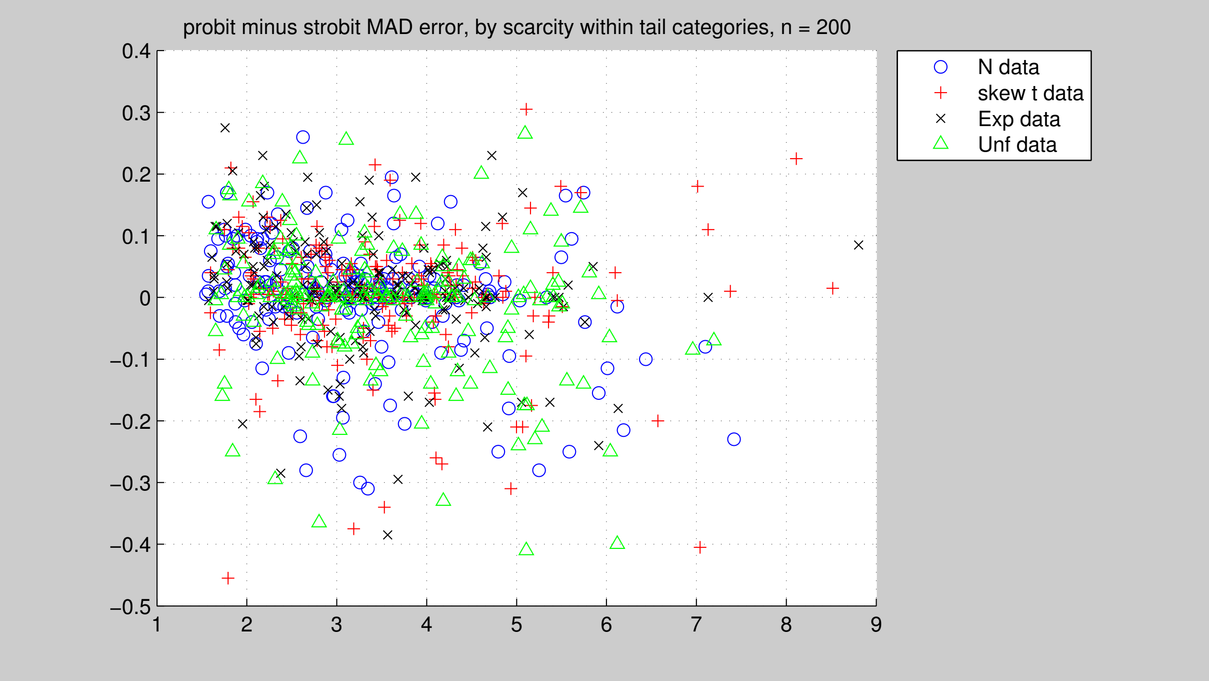
This set of histograms represents a visual take on a classification matrix. Off-diagonal elements of the classification matrix (*i.e.* classification errors) are to the right and left of the '0' column.

Figure 6.12: Two-category MAD error difference on category sparseness, by data scenario, $n = 200$



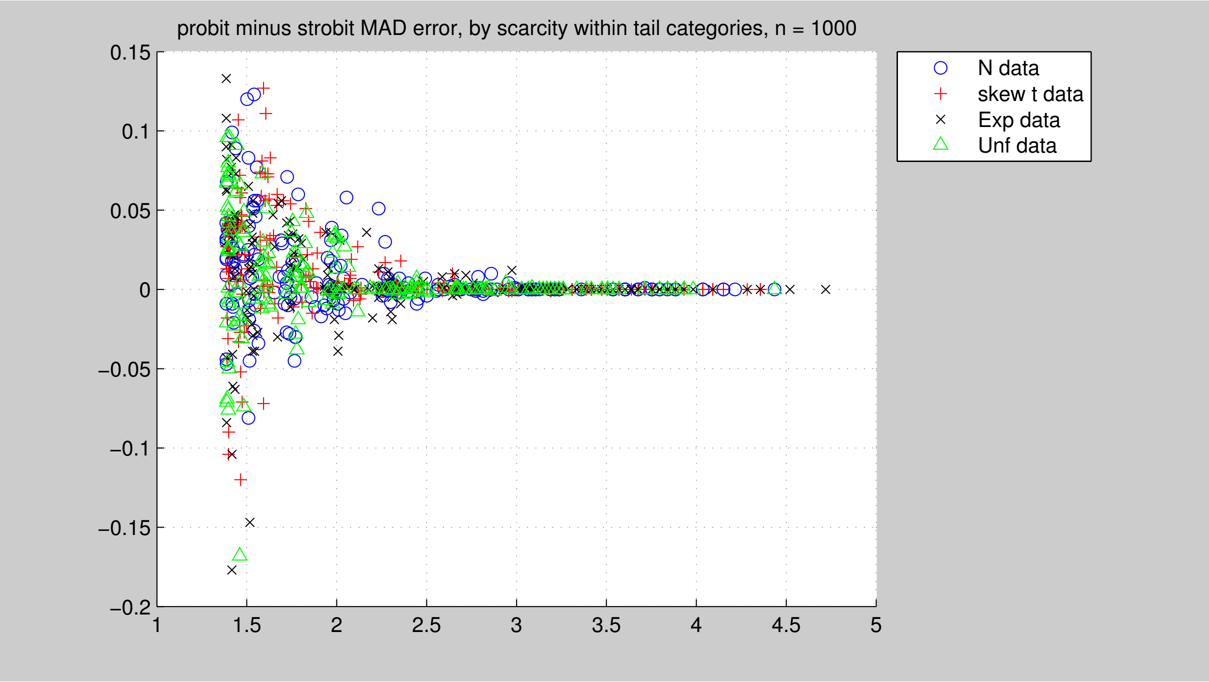
Scatterplot of probit minus strobit MAD error by negative sum of logs of category proportions for the two-category multiple simulation procedure.

Figure 6.13: Three-category MAD error difference on tail category sparseness, by data scenario, $n = 200$



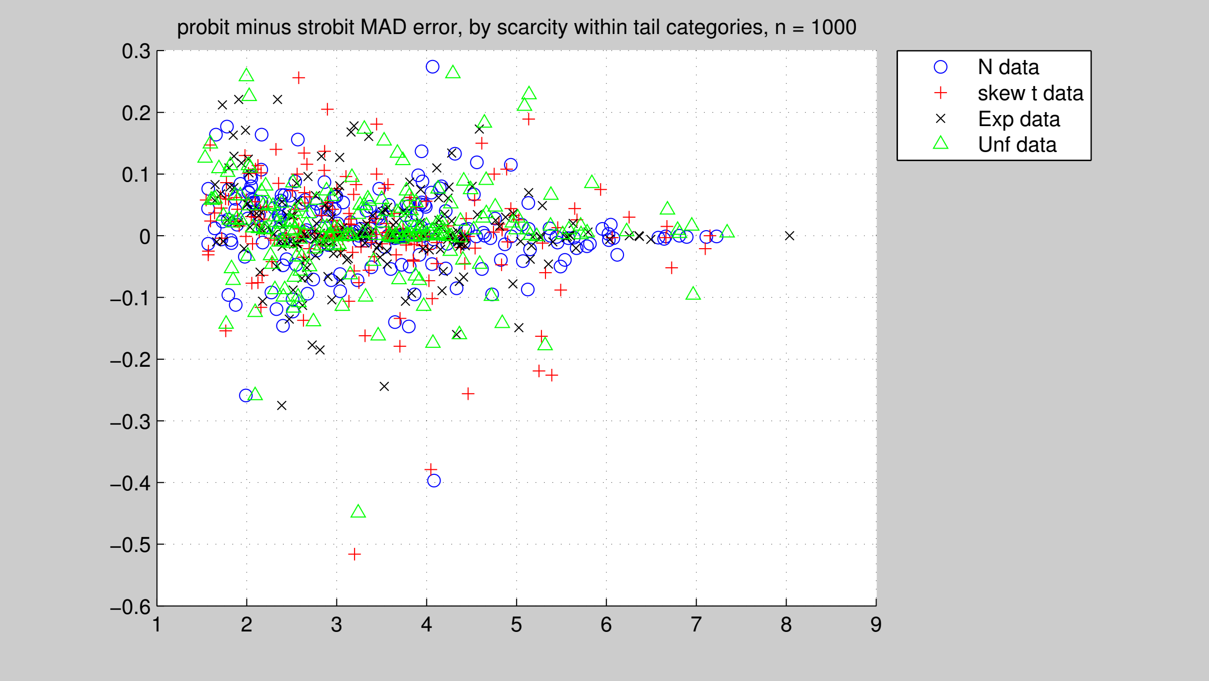
Scatterplot of probit minus strobit MAD error by negative sum of logs of outer category proportions for the three-category multiple simulation procedure.

Figure 6.14: Two-category MAD error difference on category sparseness, by data scenario, $n = 1000$



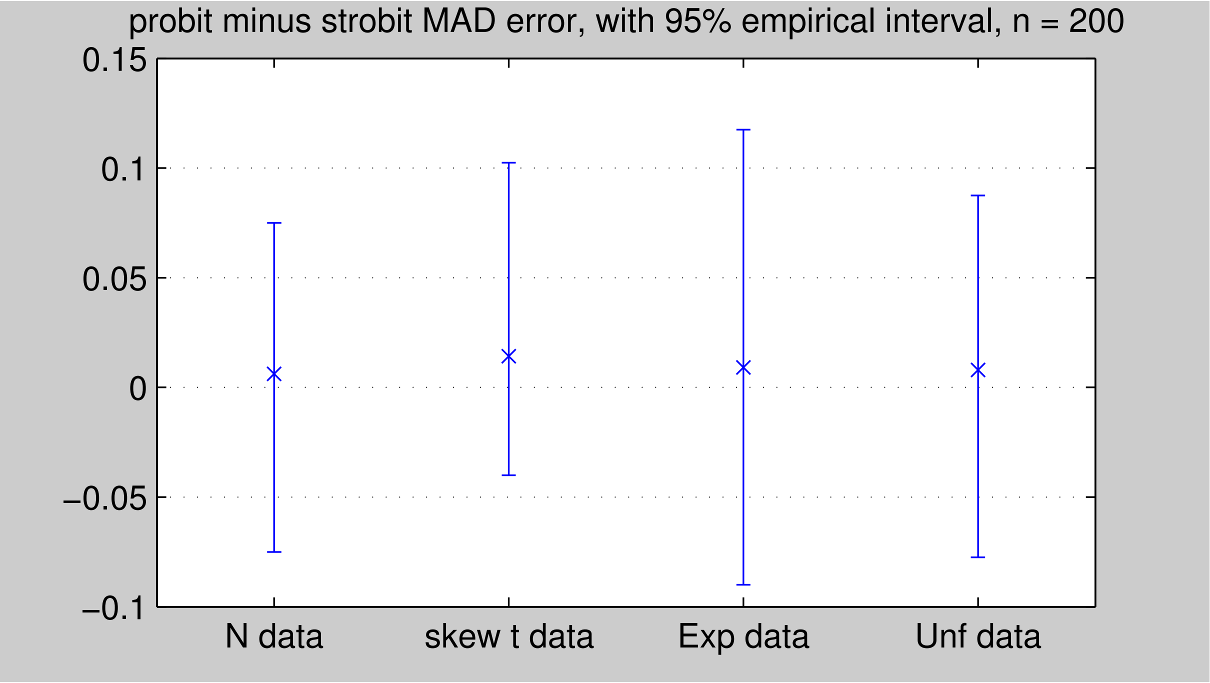
Scatterplot of probit minus strobit MAD error by negative sum of logs of category proportions for the two-category multiple simulation procedure.

Figure 6.15: Three-category MAD error difference on tail category sparseness, by data scenario, $n = 1000$



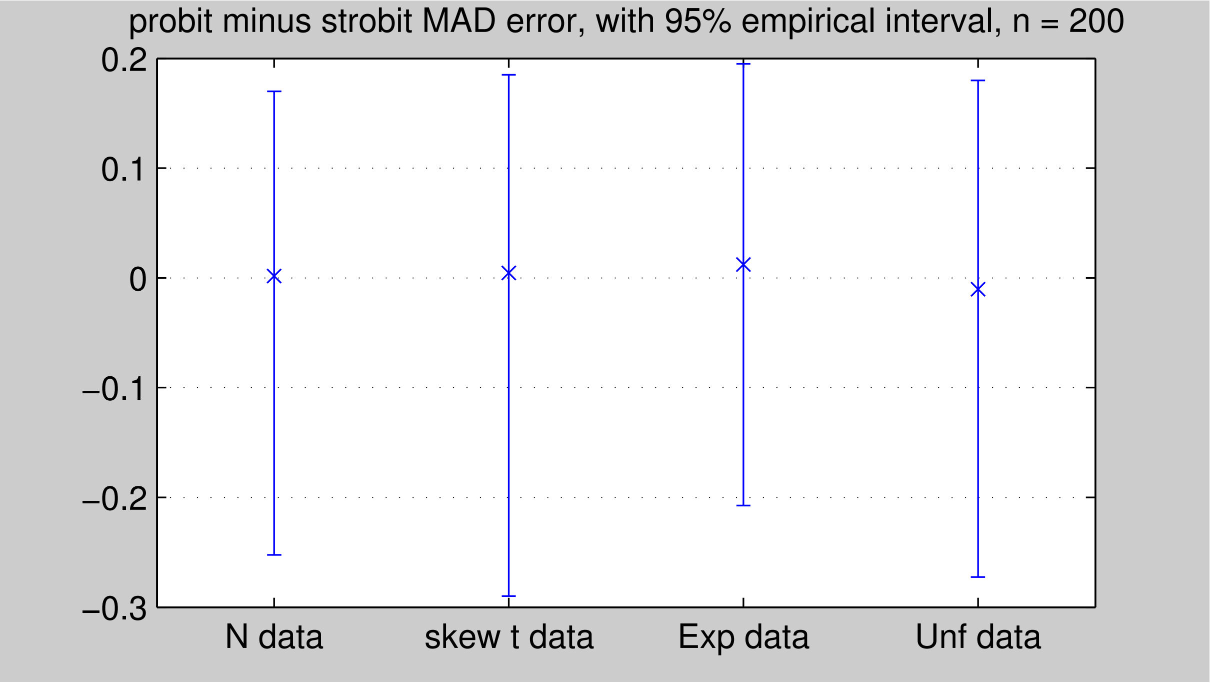
Scatterplot of probit minus strobit MAD error by negative sum of logs of outer category proportions for the three-category multiple simulation procedure.

Figure 6.16: Two-category mean MAD error difference with 95% interval, by data scenario, $n = 200$



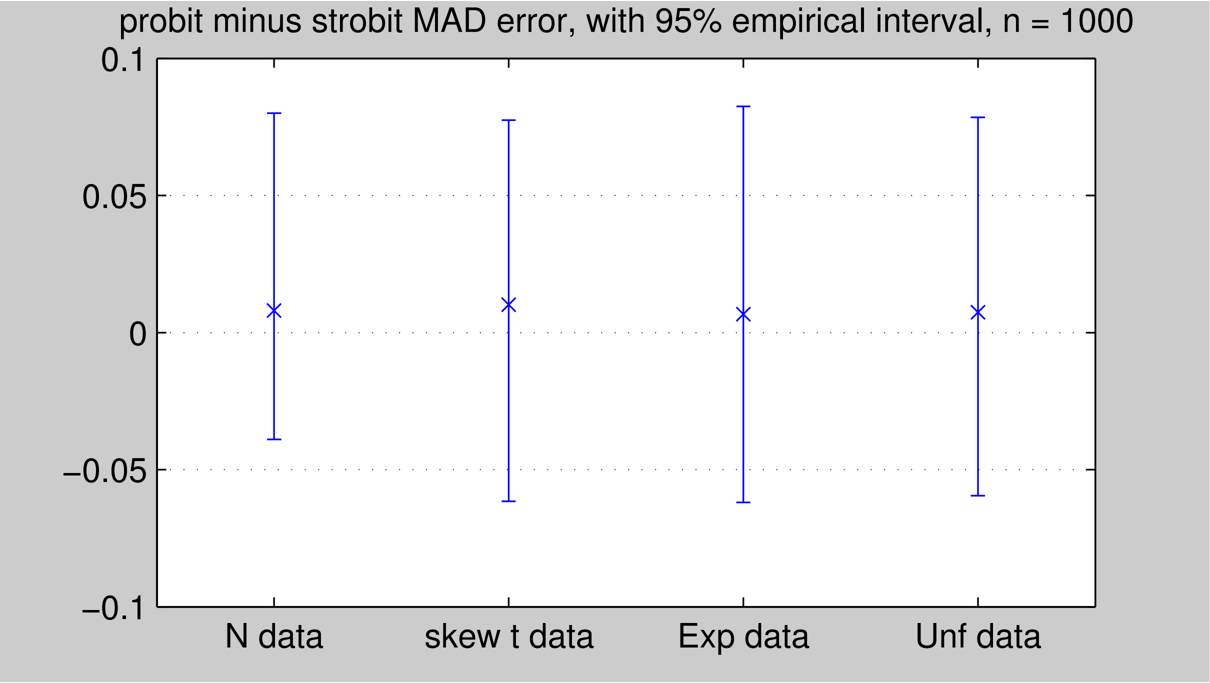
Mean probit minus strobit MAD error, with associated 95% empirical intervals, by data scenario, for the two-category multiple simulation procedure.

Figure 6.17: Three-category mean MAD error difference with 95% interval, by data scenario, $n = 200$



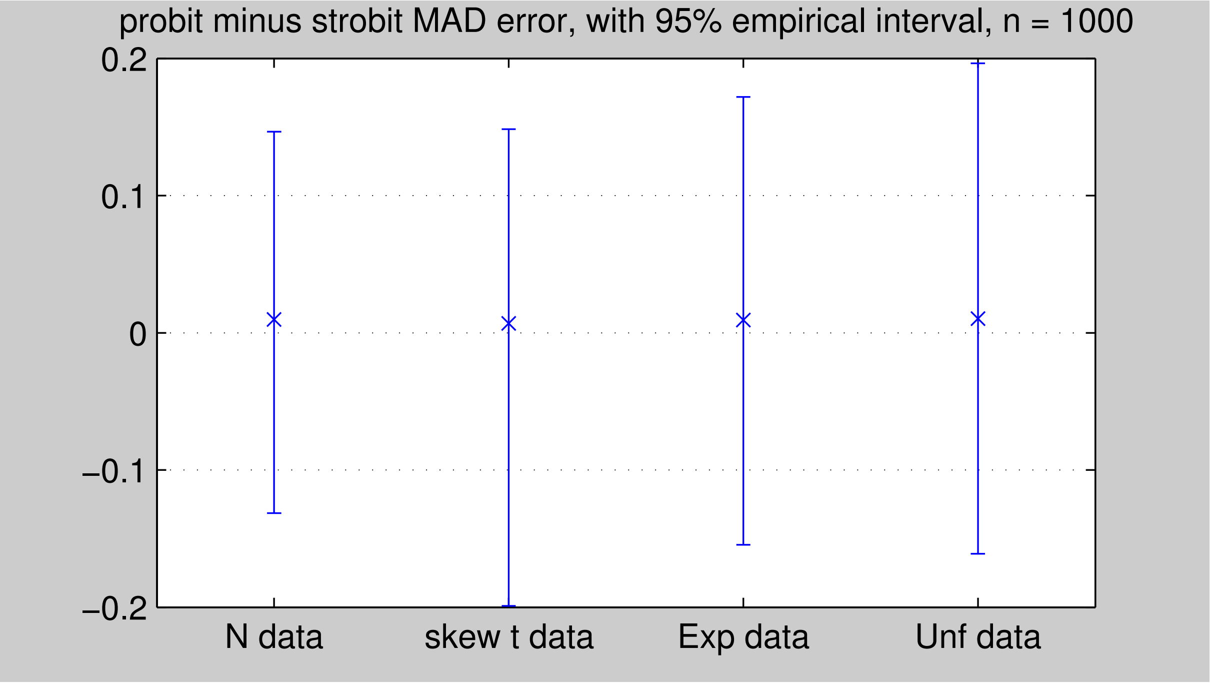
Mean probit minus strobit MAD error, with associated 95% empirical intervals, by data scenario, for the three-category multiple simulation procedure.

Figure 6.18: Two-category mean MAD error difference with 95% interval, by data scenario, $n = 1000$



Mean probit minus strobit MAD error, with associated 95% empirical intervals, by data scenario, for the two-category multiple simulation procedure.

Figure 6.19: Three-category mean MAD error difference with 95% interval, by data scenario, $n = 1000$



Mean probit minus strobit MAD error, with associated 95% empirical intervals, by data scenario, for the three-category multiple simulation procedure.

Chapter 7

Conclusion

7.1 Review

Incomplete data, or the presence of missing data, is a prevalent problem in large survey data sets. Missing data is almost always multivariate in nature, and occurs for numerous reasons, from non-response to drop-out to purposeful deletion. There are several archaic and inappropriate methods for handling missing data, such as re-weighting and single imputation, that need to be avoided when completing a data set for public use. Multiple imputation (MI) was developed as a successor to these methods. The MI family of methods predict several plausible values for each missing datum; in this way several completed data sets are created, ready for complete-case analysis methods. The inferences from the same analysis on the multiple data sets are combined using particular rules to obtain a single overall inference. These rules were formulated based on asymptotic theory, or an infinite number of completed data sets. However, the combining rules were subsequently defined for a finite number of imputations, and the guidelines for proper, valid, MI were established. The beauty of MI lies in the fact that the three types of uncertainty associated with missing data are incorporated into the final complete-data analysis: the uncertainty associated with the missing data mechanism (MDM); the uncertainty associated with the imputation model; and, the uncertainty associated with the sampled units.

It is assumed that a random process, the MDM, causes missingness in data sets. The MDM can be completely random (MCAR), random but based on observed data (MAR), or not random, *i.e.* based on unobserved data (MNAR). The MCAR and MAR MDMs are ignorable in multiple imputation, meaning that one need not explicitly model the MDM in the joint posterior distribution of missingness and data model. The assumption of an ignorable MDM is not a very strong one, and researchers believe it must at least be used even if the missing data is MNAR. The flexibility that this assumption provides is useful, and has allowed for numerous advances in specifying the imputation model without having the complexity of the MDM confounding the process. Bayesian statistics has proved useful in MI modelling, since model and sampling uncertainty (the two uncertainties that cannot be ignored) can easily be incorporated in a posterior predictive distribution.

There has been a vast amount of research in the area of MI. Topics of interest have included uncongeniality and efficiency, adaptations of the combining rules, research into nonignorable MDMs, and comparisons between MI inferences and non-MI inferences. There is consensus that MI represents one of the best solutions (and certainly the most parsimonious) for the missing data problem when the imputation task and the analysis task are performed by different experts.

One significant development in the field of MI has been sequential regression multiple imputation (SRMI). It is easy to see that joint modelling of an entire survey data set can be a complex (if not impossible) task when the variables in the data set are of different types (continuous versus discrete, for example), or follow different univariate distributions. For this reason, SRMI was developed. In SRMI, a univariate multiple regression is performed separately and sequentially on each incomplete variable, predicting the missing data points through the corresponding univariate posterior predictive models. The process is refined into an approximate Gibbs sampler, so that after some iteration, draws from the univariate predictive posteriors approximate draws from the joint model, as is required in proper MI.

The SRMI algorithm has proved useful, because any type of Bayesian regression model (or approximately Bayes adaptation of a model) can be incorporated in the chained equations. In this way, variables are completed using plausible models appropriate to those variables. Research in SRMI has shown that it is at least equivalent to fully Bayesian joint model MI, and considerably better than single imputation methods for solving the missing data problem. The SRMI algorithm has proved useful under different circumstances, from its intended purpose, namely solving the missing data problem, to disclosure limitation of public use data. There has been research into the evaluation of model selection for the chained equation models in SRMI, but this is an area that is still rather open. Of particular interest to this thesis is the need for a robust model in SRMI that can handle both Normal-type data and data with heavy tails and/or skewness.

The need for a robust SRMI model is met through the implementation of the Bayesian skew Student's t-distribution for regression model errors. Estimation procedures and algorithms are designed and summarised in this thesis, and the model is compared with both the Normal SRMI model, a symmetric t model, and predictive mean matching (PMM), local residual draw (LRD) and expanded residual draw (ERD) adaptations designed to incorporate skewness into the symmetric models. The skew t model proves to be useful as a flexible SRMI model alternative, handling both Normal and non-Normal incomplete variables well.

The calibrated posterior predictive p-value (*cppp*), designed to test model fit in Bayesian statistics, is also examined in this thesis. The *cppp* is mathematically derived under the assumptions of Normality and complete data, to serve as a test statistic for deviation from Normality. The complete data assumption is merely a formality, knowing that a well defined *cppp* model can easily be extended to the Normal SRMI algorithm. However, this thesis shows that the *cppp* is not without its own problems. The measure requires known proper priors for the regression parameters — a task that is not trivial in practice. Various attempts at determining or approximating these priors were unsuccessful in this thesis. One possible solution is a Markov Chain Monte Carlo (MCMC) approximation to the *cppp*. However, this method also has drawbacks, and requires further study. Thus, while the thesis attempted to shed light on an SRMI evaluation measure, several obstacles were encountered, preventing a deeper advance into this specific research area. However, these problems should not detract from the positive results displayed in this thesis, since the robust model can be used when the data is Normal as well; the *cppp* evaluation statistic would merely speed up the SRMI algorithm in practice, by allowing the Normal model to be used when the robust model is not entirely necessary.

Finally, the robust skew t model is implemented in the context of Bayesian estimation of ordinal categorical data, in a new model called the strobit model. Again, the assumption of complete data is used without loss of generality to implementation in SRMI, since the estimation procedure is only defined with prediction and classification in mind. The goal of the study was to better predict observations in scarcely populated ordinal data tail categories. While this goal was not achieved, general prediction is slightly improved over the probit model, so that the strobit model seems is a viable alternative to the probit model in SRMI.

7.2 Further Research

Several topics for research have been uncovered in the pursuit of the two objectives of this thesis: MI and SRMI review; and robust SRMI model development. These topics can be divided into two broad categories, namely, open topics that are not related to the work in this thesis (which were mentioned in passing during the literature review), and topics that have been discovered in the development of the novel methods in this thesis.

Open topics in MI

- The possibility of using the R^2 statistic from binary-outcome regression models explaining variable missingness might be developed into a measure of the MAR nature of the MDM. Low R^2 values could simply be used to measure the extent to which the MDM may be MNAR. Additionally, the effect of spuriously making a MNAR mechanism more MAR by adding additional covariates could be investigated.
- Extending the work of Zhou & Reiter (2010), one might consider a more extensive comparison of using smaller numbers of imputed data sets versus larger numbers of imputed data sets when the completed-data inferences are not based on the Normal distribution. Additionally, one could compare the inferences based on the regular combining rules with the inferences based on the empirical distributions of the large number of completed data sets. The rationale for such a study is simple — to determine under the robustness of the regular combining rules.
- It was noted that research into the interdependency of imputed data sets is warranted, as the Gibbs sampler might not produce imputation draws that display

enough independence in the MI algorithm.

- One could investigate the effect of donor proportions in PMM and LRD adaptations of symmetrical models used in SRMI.

Topics related to the development of the robust model for SRMI

- The robust skew t model is in its infancy. For this reason, numerous related topics could be expanded on, including (but not limited to):
  - refining and speeding up the estimation procedure;
  - investigating the effect of discretising the posterior distribution for the degrees of freedom in the skew t estimation procedure;
  - comparing actual inferences on data completed using the Normal model against inferences on completed using the robust model;
  - investigating the effect of a t-distributed Z value — *i.e.* a heavy-tailed exposure to the skewness parameter, δ; and,
  - introducing into SRMI a possible light-tailed alternative to the skew t, namely the skew Normal.
- Extensive work can be done on the *cppp* SRMI evaluation procedure. Besides attempting to find a way of correctly specifying the prior distribution of the regression parameters, work on approximate methods can also be extended. An investigation into the seemingly already-calibrated nature of the *ppp* displayed in this thesis is also warranted.
- One could compile a comparison of the strobit model with the regular multinomial formulation in SRMI, and with other correspondence analysis based imputation methods. Additionally, research is still required to find a Bayesian model that might better predict missing ordinal data when the tail categories are sparsely populated.

7.3 Final Note

In summary, this thesis has reviewed the existing research on MI and SRMI, and has successfully developed and implemented a robust model for use in SRMI, for both continuous and ordinal data. While this development may seem like a small step for a statistician, imputers would like other statisticians to remember that all statistical theory is, in fact, the study of incomplete data, and that any advance in the field of missing data is an advance for all ‘statistician-kind’.

References

Abayomi, K., Gelman, A. & Levy, M. (2008), 'Diagnostics for multivariate imputations', *Journal for the Royal Statistical Society: Series C (Applied Statistics)* **57**(3), 273–291.

Albert, J. & Chib, S. (1993), 'Bayesian analysis of binary and polychotomous response data', *Journal of the American Statistical Association* **88**(422), 669–679.

Ambler, G., Omar, R. Z. & Royston, P. (2007), 'A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome', *Statistical Methods in Medical Research* **16**, 277–298.

Ardington, C., Lam, D., Leibbrandt, M. & Welch, M. (2006), 'The sensitivity to key data imputations of recent estimates of income poverty and inequality in south africa', *Economic Modelling* **23**, 822–835.

Barnard & Rubin, D. B. (1999), 'Small-sample degrees of freedom with multiple imputation', *Biometrika* **86**, 949–955.

Barnes, H., Gutierrez-Romero, R. & Noble, M. (2006), Multiple imputation of missing data in the 2001 south african census, Working Paper 4, Centre for the Analysis of South African Social Policy, University of Oxford.

Berger, J. & Bernardo, J. (1992), 'On the development of reference priors', *Bayesian Statistics* **4**, 35–60.

Bishop, Y. M. M., Feinberg, S. E. & Holland, P. W. (1975), *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge, MA.

Box, G. E. P. & Cox, D. R. (1964), 'An analysis of transformations', *Journal of the Royal Statistical Society* **Ser. B**(26), 211–252. (With discussion).

Brand, J. P. L. (1998), Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets, PhD thesis, Erasmus University, Rotterdam.

Cabras, S., Castellanos, M. E. & Quirós, A. (2011), 'Goodness-of-fit of conditional regression models for multiple imputation', *Bayesian Analysis* **6**(3), 429–456.

Carpenter, J. R. & Kenward, M. G. (2007), 'Sensitivity analysis after multiple imputation under missing at random: a weighting approach', *Statistical Methods in Medical Research* **16**, 259–275.

Cleveland, W. S. (1979), 'Locally weighted regression and smoothing scatterplots', *Journal of the American Statistical Association* **74**, 829–836.

Cochran, W. G. (1977), *Sampling Techniques*, 3 edn, John Wiley, New York.

Collins, L. M., Schafer, J. L. & Kam, C. M. (2001), 'A comparison of inclusive and restrictive strategies in modern missing-data procedures', *Psychological Methods* **6**, 330–351.

Datta, G. & Ghosh, J. (1995), 'On priors providing frequentist validity for bayesian inference', *Biometrika* **82**(1), 37–45.

de Jong, R., van Buuren, S. & Spiess, M. (2014), 'Multiple imputation of predictor variables using generalized additive models', *Communications in Statistics - Simulation and Computation* .

Dobson, A. J. (2002), *An Introduction to Generalised Linear Models*, 2 edn, Chapman & Hall/CRC, Boca Raton.

Faris, P. D., Ghali, W. A., Brant, R., Norris, C. M., Galbraith, P. D. & Knudtson, M. L. (2002), 'Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses', *Journal of Clinical Epidemiology* **55**, 184–191.

Fay, R. E. (1992), When are inferences from multiple imputation valid?, *in* 'Proceedings of the Survey Research Methods Section', American Statistical Association, Alexandria, VA., pp. 227–232.

Fonseca, T. C. O., Ferreira, M. A. R. & Migon, H. S. (2008), 'Objective bayesian analysis for the student-t regression model', *Biometrika* **95**(2), 325–333.

Gelman, A. (2004), 'Parameterization and bayesian modeling', *Journal of the American Statistical Association* **99**(466), 537–545.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2004), *Bayesian Data Analysis*, Chapman & Hall, London.

Gelman, A., Meng, X.-L. & Stern, H. (1996), 'Posterior predictive assessment of model fitness via realized discrepancies', *Statistica Sinica* **6**, 733–807. With discussions.

Gelman, A. & Rubin, D. B. (1992), 'Inference from iterative simulation using multiple sequences', *Statistical Science* **7**(4), 457–472.

Glynn, R., Laird, N. & Rubin, D. B. (1993), 'The performance of mixture models for nonignorable nonresponse with follow ups', *Journal of the American Statistical Association* **88**(423), 984–993.

Groenewald, P. & Mokgatlhe, L. (2005), 'Bayesian computation for logistic regression', *Computational Statistics and Data Analysis* **48**, 857–868.

He, Y. & Raghunathan, T. E. (2006), 'Tukey's *gh* distribution for multiple imputation', *The American Statistician* **60**, 251–256.

He, Y. & Raghunathan, T. E. (2009), 'On the performance of sequential regression multiple imputation methods with non normal error distributions', *Communications in Statistics - Simulation and Computation* **38**, 856–883.

Hjort, N. L., Dahl, F. A. & Steinbakk, G. H. (2006), 'Post-processing predictive p values', *Journal of the American Statistical Association* **101**(475), 1157–1174.

Kennickell (1991), Imputation of the 1989 survey of consumer finances: Stochastic relaxation and multiple imputation, *in* 'Proceedings of the Survey Research Methods Section', American Statistical Association, pp. 112–121.

Kenward, M. G. & Carpenter, J. (2007), 'Multiple imputation: current perspectives', *Statistical Methods in Medical Research* **16**, 199–218.

Lavori, P. W., Dawson, R. & Shera, D. (1995), 'I multiple imputation strategy for clinical trials with truncation of patient data', *Statistics in Medicine* **14**, 1913–1925.

Lee, K. J. & Carlin, J. B. (2010), 'Multiple imputation for missing data: Fully conditional specification vesrsus multivariate normal imputation', *American Journal of Epidemiology* **171**(5), 624–632.

Liang, F., Paulo, R., Molina, G., Clyde, M. & Berger, J. O. (2008), 'Mixtures of g priors for bayesian variable selection', *Journal of the American Statistical Association* **103**(481), 410–423.

Little, R. J. A. & Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, 2 edn, John Wiley & Sons.

Liu, C. (2005), Robit regression: A simple robust alternative to logistic and probit regression, *in* A. Gelman & X.-L. Meng, eds, 'Applied Bayesian Modeling and Causal Inference from Income-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family', John Wiley & Sons, Chichester, U.K., pp. 227–238.

MATLAB (2012), *version 7.14.0.739 (R2012a)*, The MathWorks Inc., Natick, Massachusetts.

McBride, M. (2001), 'Relative-income effects on subjective well-being in the cross-section', *Journal of Economic Behavior & Organization* **45**, 251–278.

Meng, X.-L. (1993), Coherent multiple-imputation inference under incoherent models, Technical Report 359, Dept. Statistics, University of Chicago.

Meng, X.-L. (1994), 'Multiple-imputation inferences with uncongenial sources of input', *Statistical Science* **9**(4), 538–558.

Meng, X.-L. & Romero, M. (2003), 'Discussion: Efficiency and self-efficiency with multiple imputation inference', *International Statistical Review* **71**(3), 607–618.

Nielson, S. F. (2003), 'Proper and improper multiple imputation', *International Statistical Review* **71**(3), 593–607.

O'Hagan, A. (1995), 'Fractional bayes factors for model comparison', *Journal for the Royal Statistical Society: Series B (Methodological)* **57**(1), 99–138.

Paiva, T., Chakraborty, A., Reiter, J. P. & Gelfand, A. (2014), 'Imputation of confidential data sets with spatial locations using disease mapping models', *Statistics in Medicine* **33**, 1928–1945.

Pearn, W. & Wu, C. (2005), 'A bayesian approach for assessing process precision based on multiple samples', *European Journal of Operational Research* **165**(3), 685–695.

Penn, D. A. (2007), 'Estimating missing values from the general social survey: An application of multiple imputation', *Social Science Quarterly* **88**(2), 573–584.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J. & Solenberger, P. (2001), 'A multivariate technique for multiply imputing missing values using a sequence of regression models', *Survey Methodology* **27**(1), 85–95.

Reiter, J. P. (2005), 'Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study', *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **168**(1), 185–205.

Reiter, J. P. (2009), 'Using multiple imputation to integrate and disseminate confidential microdata', *International Statistical Review* **77**(2), 179–195.

Reiter, J. P. (2012), 'Statistical approached to protecting confidentiality for microdata and their effects on the quality of statistical inferences', *Public Opinion Quarterly* **76**(1), 163–181.

Reiter, J. P. & Raghunathan, T. E. (2007), 'The multiple adaptations of multiple imputation', *Journal of the American Statistical Association* **102**(480), 1462–1471.

Rubin, D. B. (1976), 'Inference and missing data', *Biometrika* **63**(3), 581–592.

Rubin, D. B. (1978), Multiple imputation in sample surveys — a phenomenological bayesian approach to nonresponse, *in* 'Proceedings of the Survey Research Methods Section', American Statistical Association, Washington, D.C., pp. 20–34.

Rubin, D. B. (1981), 'The bayesian bootstrap', *The Annals of Statistics* **9**, 130–134.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York.

Rubin, D. B. (1996), 'Multiple imputation after 18+ years', *Journal of the American Statistical Association* **91**(434), 473–489.

Rubin, D. B. (2003*a*), 'Discussion on multiple imputation', *International Statistical Review* **71**(3), 619–625.

Rubin, D. B. (2003*b*), 'Nested multiple imputation of nmes via partially incompatible mcmc', *Statistica Neerlandica* **57**(1), 3–18.

Rubin, D. B. & Schenker, N. (1986), 'Multiple imputation for interval estimation from samples with ignorable nonresponse', *Journal of the American Statistical Association* **81**(394), 366–374.

Rubin, D. B. & Schenker, N. (1991), 'Multiple imputation in health-care databases: An overview and some applications', *Statistics in Medicine* **10**(4), 585–598.

Sahu, S. K., Dey, D. K. & Branco, M. D. (2003), 'A new class of multivariate skew distributions with applications to bayesian regression models', *The Canadian Journal of Statistics* **31**(2), 129–150.

Saunders, J. A., Morrow-Howell, N., Spitznagel, E., Doré, P., Proctor, E. K. & Pescarino, R. (2006), 'Imputing missing data: A comparison of methods for social work researchers', *Social Work Research* **30**(1), 19–35.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, CRC Press, New York.

Schafer, J. L. (2003), 'Multiple imputation in multivariate problems when the imputation and analysis methods differ', *Statistica Nederlandica* **57**(1), 19–35.

Schafer, J. L. & Graham, J. W. (2002), 'Missing data: Our view of the state of the art', *Psychological Methods* **7**(2), 147–177.

Schenker, N., Raghunathan, T. E., Chiu, P.-L., Makuc, D. M., Zhang, G. & Cohen, A. J. (2006), 'Multiple imputation of missing income data in the national health interview survey', *Journal of the American Statistical Association* **101**(475), 924–933.

Schenker, N. & Taylor, J. M. G. (1996), 'Partially parametric techniques for multiple imputation', *Computational Statistics and Data Analysis* **22**(425–446).

Shen, Z. (2000), Nested Multiple Imputations, PhD thesis, Harvard University, Cambridge, MA.

Siddique, J. & Belin, T. R. (2008), 'Using an approximate bayesian bootstrap to multiply impute nonignorable missing data', *Computational Statistics and Data Analysis* **53**, 405–415.

Tukey, J. W. (1977), Modern techniques in data analysis, MSF-Sponsored Regional Research Conference at Southeastern Massachusetts University, North Dartmouth, MA.

van Buuren, S. (2007), 'Multiple imputation of discrete and continuous data by fully conditional specification', *Statistical Methods in Medical Research* **16**, 219–242.

van Buuren, S., Boshuizen, H. C. & Knook, D. (1999), 'Multiple imputation of missing blood pressure covariates in survival analysis', *Statistics in Medecine* **18**, 681–694.

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M. & Rubin, D. B. (2006), 'Fully conditional specification in multivariate imputation', *Journal of Statistical Computation and Simulation* **76**(12), 1049–1064.

van der Heijden, G. J. M. G., Donders, A. R. T., Stijnen, T. & Moons, K. G. M. (2006), 'Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example', *Journal of Clinical Epidemiology* **59**, 1102–1109.

Wang, H. & Reiter, J. P. (2012), 'Multiple imputation for sharing precise geographies in public use data', *Annals of Applied Statistics* **6**(1), 229–252.

Zhang, P. (2003), 'Multiple imputation: Theory and method', *International Statistical Review* **71**(3), 581–592.

Zhou, X. & Reiter, J. P. (2010), 'A note on bayesian inference after multiple imputation', *The American Statistician* **64**(2), 159–163.

Appendix A

Priors for the Student *t*-distribution

This appendix provides the derivations (or references for these derivations) for the Jeffreys, reference, and probability-matching priors for the symmetric Student t-distribution. Reference and probability-matching priors generally lead to procedure with properties frequentists can relate to while still retaining Bayesian validity.

The Jeffreys Prior. It is shown by Fonseca et al. (2008) that the independence Jeffreys prior is

$$P_{IJEFF}\left(\nu,\beta,\sigma\right)\propto\sigma^{-1}\left(\frac{\nu}{\nu+3}\right)^{\frac{1}{2}}\left[\psi'\left(\frac{\nu}{2}\right)-\psi'\left(\frac{\nu+1}{2}\right)-\frac{2\left(\nu+3\right)}{\nu\left(\nu+1\right)^{2}}\right]^{\frac{1}{2}}$$

assuming that the marginal priors for β and (σ,ν) are independent *a priori*. Note that $\psi'\left(\cdot\right)$ is the trigamma function, the derivative of the digamma function.

Reference Priors. We derive the reference prior of Berger & Bernardo (1992) for the parameters of the Student t-distribution. The derivation depends on the ordering of the parameters and how the parameter vector is divided into sub-vectors. The reference prior maximises the difference in information about the parameters provided by the prior and the posterior (Pearn & Wu 2005); *i.e.* the reference prior provides as little information as possible about the parameters of interest.

$$I(\beta,\sigma,\nu)=\begin{bmatrix}\beta^2 & \beta\sigma & \beta\nu\\ \sigma\beta & \sigma^2 & \sigma\nu\\ \nu\beta & \nu\sigma & \nu^2\end{bmatrix} \text{ and } I(\nu,\beta,\sigma)=\begin{bmatrix}\nu^2 & \nu\beta & \nu\sigma\\ \beta\nu & \beta^2 & \beta\sigma\\ \sigma\nu & \sigma\beta & \sigma^2\end{bmatrix}$$

In general,

$$\{I(\theta)\}_{ij} \;=\; E_{Y|\theta}\left[-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log\{L(\theta;y,x)\}\right]$$

The Fisher information matrix for the ordering $\{\nu,\beta,\sigma\}$ is therefore given by,

$$\begin{aligned} & I(\nu,\beta,\sigma)\\ &= \begin{bmatrix} \frac{n}{4}\left[\psi'\left(\frac{\nu}{2}\right)-\psi'\left(\frac{\nu+1}{2}\right)-\frac{2(\nu+5)}{\nu(\nu+1)(\nu+3)}\right] & 0 & \frac{-2n}{\sigma(\nu+1)(\nu+3)}\\ 0 & \frac{\nu+1}{\sigma^2(\nu+3)}\sum_{i=1}^n x_i x_i' & 0\\ \frac{-2n}{\sigma(\nu+1)(\nu+3)} & 0 & \frac{2n\nu}{\sigma^2(\nu+3)} \end{bmatrix} \quad (A.1)\\ &= \begin{bmatrix} F_{11} & F_{12} & F_{13}\\ F_{21} & F_{22} & F_{23}\\ F_{31} & F_{32} & F_{33}\end{bmatrix} \end{aligned}$$

To calculate the reference prior for the ordering $\{\nu,\beta,\sigma\}$, we must first calculate:

$$h_1 = F_{11} - \begin{bmatrix} F_{12} & F_{13}\end{bmatrix}\begin{bmatrix} F_{22} & F_{23}\\ F_{32} & F_{33}\end{bmatrix}^{-1}\begin{bmatrix} F_{21}\\ F_{31}\end{bmatrix}$$

from the information matrix in Equation (A.1).

$$\therefore h_1 = \frac{n}{4}\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2\left(\nu+5\right)}{\nu\left(\nu+1\right)\left(\nu+3\right)}\right]$$

$$- \begin{bmatrix} 0 & \frac{-2n}{\sigma(\nu+1)(\nu+3)} \end{bmatrix} \begin{bmatrix} \frac{\sigma^2(\nu+3)}{\nu+1}\left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} & 0 \\ & \\ 0 & \frac{\sigma^2(\nu+3)}{2n\nu} \end{bmatrix} \begin{bmatrix} 0 \\ \\ \frac{-2n}{\sigma(\nu+1)(\nu+3)} \end{bmatrix}$$

$$= \frac{n}{4}\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2\left(\nu+5\right)}{\nu\left(\nu+1\right)\left(\nu+3\right)}\right] - \begin{bmatrix} 0 & \frac{-\sigma}{\nu(\nu+1)} \end{bmatrix} \begin{bmatrix} 0 \\ \\ \frac{-2n}{\sigma(\nu+1)(\nu+3)} \end{bmatrix}$$

$$\begin{aligned}
\therefore h_1 &= \frac{n}{4}\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2\left(\nu+5\right)}{\nu\left(\nu+1\right)\left(\nu+3\right)}\right] - \frac{2n}{\nu\left(\nu+1\right)^2\left(\nu+3\right)} \\
&= \frac{n}{4}\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right)\right] - \frac{n\left(\nu+5\right)}{2\nu\left(\nu+1\right)\left(\nu+3\right)} - \frac{2n}{\nu\left(\nu+1\right)^2\left(\nu+3\right)} \\
&\propto \frac{1}{4}\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right)\right] - \frac{\left(\nu+1\right)\left(\nu+5\right)+2\left(2\right)}{2\nu\left(\nu+1\right)^2\left(\nu+3\right)} \\
&\propto \frac{1}{4}\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right)\right] - \frac{\nu^2+6\nu+9}{2\nu\left(\nu+1\right)^2\left(\nu+3\right)} \\
&\propto \frac{1}{4}\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right)\right] - \frac{\left(\nu+3\right)^2}{2\nu\left(\nu+1\right)^2\left(\nu+3\right)} \\
&\propto \frac{1}{4}\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right)\right] - \frac{\nu+3}{2\nu\left(\nu+1\right)^2} \\
&\propto \psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2\left(\nu+3\right)}{\nu\left(\nu+1\right)^2} \\
\therefore h_1^{\frac{1}{2}} &\propto \left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2\left(\nu+3\right)}{\nu\left(\nu+1\right)^2}\right]^{\frac{1}{2}}
\end{aligned}$$

Now, $p\left(\nu\right) \propto h_1^{\frac{1}{2}}$. Also, since it does not contain β, $h_2^{\frac{1}{2}} \propto c$. So $p\left(\beta|\nu\right) \propto c$. Further, $h_3 = F_{33} = \frac{2n\nu}{\sigma^2(\nu+3)}$, and $p\left(\sigma^2|\nu,\beta\right) \propto h_3^{\frac{1}{2}} = \sigma^{-1}$,

$$\begin{aligned}
\therefore P_{REF}^{1}\left(\nu,\beta,\sigma\right) &= p\left(\nu\right)p\left(\beta|\nu\right)p\left(\sigma^2|\nu,\beta\right) \\
&\propto \sigma^{-1}\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2\left(\nu+3\right)}{\nu\left(\nu+1\right)^2}\right]^{\frac{1}{2}}.
\end{aligned}$$

Similarly, we can find the Reference Prior for the ordering (β, σ, ν):

$$P^2_{REF}(\beta,\sigma,\nu) \propto \sigma^{-1}\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+5)}{\nu(\nu+1)(\nu+3)}\right]^{\frac{1}{2}}$$

The Fisher information matrix for the ordering $\{\nu, \sigma, \beta\}$ is,

$$I(\nu,\sigma,\beta) \\ = \begin{bmatrix} \frac{n}{4}\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+5)}{\nu(\nu+1)(\nu+3)}\right] & \frac{-2n}{\sigma(\nu+1)(\nu+3)} & 0 \\ \frac{-2n}{\sigma(\nu+1)(\nu+3)} & \frac{2n\nu}{\sigma^2(\nu+3)} & 0 \\ 0 & 0 & \frac{\nu+1}{\sigma^2(\nu+3)}\sum_{i=1}^n x_i x_i' \end{bmatrix}$$

$$\begin{aligned} \therefore h_1 &= \frac{n}{4}\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+5)}{\nu(\nu+1)(\nu+3)}\right] \\ &\quad - \begin{bmatrix} \frac{-2n}{\sigma(\nu+1)(\nu+3)} & 0 \end{bmatrix} \begin{bmatrix} \frac{\sigma^2(\nu+3)}{2n\nu} & 0 \\ 0 & \frac{\sigma^2(\nu+3)}{\nu+1}\left(\sum_{i=1}^n x_i x_i'\right)^{-1} \end{bmatrix} \begin{bmatrix} \frac{-2n}{\sigma(\nu+1)(\nu+3)} \\ 0 \end{bmatrix} \\ &= \frac{n}{4}\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+5)}{\nu(\nu+1)(\nu+3)}\right] \\ &\quad - \begin{bmatrix} \frac{-\sigma}{\nu(\nu+1)} & 0 \end{bmatrix} \begin{bmatrix} \frac{-2n}{\sigma(\nu+1)(\nu+3)} \\ 0 \end{bmatrix} \\ &= \frac{n}{4}\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+5)}{\nu(\nu+1)(\nu+3)}\right] - \frac{2n}{\nu(\nu+1)^2(\nu+3)} \end{aligned}$$

$$\therefore h_1^{\frac{1}{2}} \propto \left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+3)}{\nu(\nu+1)^2}\right]^{\frac{1}{2}}$$

$$\begin{aligned}\therefore P^{3}_{REF}(\nu,\sigma,\beta) &= P^{1}_{REF}(\nu,\beta,\sigma)\\ &\propto \sigma^{-1}\left[\psi'\left(\frac{\nu}{2}\right)-\psi'\left(\frac{\nu+1}{2}\right)-\frac{2(\nu+3)}{\nu(\nu+1)^{2}}\right]^{\frac{1}{2}}\end{aligned}$$

The Fisher information matrix for the ordering $\{\beta,\nu,\sigma\}$ is:

$$I(\beta,\nu,\sigma)=\begin{bmatrix}\frac{\nu+1}{\sigma^{2}(\nu+3)}\sum_{i=1}^{n}x_i x_i' & 0 & 0\\ 0 & \frac{n}{4}\left[\psi'\left(\frac{\nu}{2}\right)-\psi'\left(\frac{\nu+1}{2}\right)-\frac{2(\nu+5)}{\nu(\nu+1)(\nu+3)}\right] & \frac{-2n}{\sigma(\nu+1)(\nu+3)}\\ 0 & \frac{-2n}{\sigma(\nu+1)(\nu+3)} & \frac{2n\nu}{\sigma^{2}(\nu+3)}\end{bmatrix}$$

It is clear that $h_1 \propto c$, therefore $p(\beta)\propto h_1^{\frac{1}{2}}\propto c$. Define H as follows:

$$\begin{aligned}H &= \begin{bmatrix}\frac{\nu+1}{\sigma^{2}(\nu+3)}\sum_{i=1}^{n}x_i x_i' & 0\\ 0 & \frac{n}{4}\left[\psi'\left(\frac{\nu}{2}\right)-\psi'\left(\frac{\nu+1}{2}\right)-\frac{2(\nu+5)}{\nu(\nu+1)(\nu+3)}\right]\end{bmatrix}\\ &\quad -\frac{\sigma^{2}(\nu+3)}{2n\nu}\begin{bmatrix}0\\ \frac{-2n}{\sigma(\nu+1)(\nu+3)}\end{bmatrix}\begin{bmatrix}0 & \frac{-2n}{\sigma(\nu+1)(\nu+3)}\end{bmatrix} \qquad \text{(A.2)}\\ &= \begin{bmatrix}H_{11} & H_{12}\\ H_{21} & H_{22}\end{bmatrix}.\end{aligned}$$

Since $h_2 = H_{22}$ it follows that,

$$\therefore h_2 = \frac{n}{4}\left[\psi'\left(\frac{\nu}{2}\right)-\psi'\left(\frac{\nu+1}{2}\right)-\frac{2(\nu+5)}{\nu(\nu+1)(\nu+3)}\right]-\frac{4n^{2}\sigma^{2}(\nu+3)}{2n\sigma^{2}\nu(\nu+1)^{2}(\nu+3)^{2}}$$

$$\therefore h_2^{\frac{1}{2}} \propto \sigma^{-1}\left[\psi'\left(\frac{\nu}{2}\right)-\psi'\left(\frac{\nu+1}{2}\right)-\frac{2(\nu+3)}{\nu(\nu+1)^{2}}\right]^{\frac{1}{2}}$$

Further, $h_3^{\frac{1}{2}} = \left[\frac{2n\nu}{\sigma^2(\nu+3)}\right]^{\frac{1}{2}}$, and $p\left(\sigma|\nu,\beta\right) \propto h_3^{\frac{1}{2}} = \sigma^{-1}$. So

$$\begin{aligned}\therefore P_{REF}^4 &= P_{REF}^3\left(\nu,\sigma,\beta\right)\\ &= P_{REF}^1\left(\nu,\beta,\sigma\right)\\ &\propto \sigma^{-1}\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2\left(\nu+3\right)}{\nu\left(\nu+1\right)^2}\right]^{\frac{1}{2}}\end{aligned}$$

The Fisher information matrix for the ordering $\{\sigma,\nu,\beta\}$ is given by,

$$\begin{aligned}&I\left(\sigma,\nu,\beta\right)\\ &= \begin{bmatrix} \frac{2n\nu}{\sigma^2(\nu+3)} & \frac{-2n}{\sigma(\nu+1)(\nu+3)} & 0 \\ \frac{-2n}{\sigma(\nu+1)(\nu+3)} & \frac{n}{4}\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+5)}{\nu(\nu+1)(\nu+3)}\right] & 0 \\ 0 & 0 & \frac{\nu+1}{\sigma^2(\nu+3)}\sum_{i=1}^n x_i x_i' \end{bmatrix}\end{aligned}$$

$$\begin{aligned}\therefore h_1 &= \frac{2n\nu}{\sigma^2\left(\nu+3\right)}\\ &- \begin{bmatrix} \frac{-2n}{\sigma(\nu+1)(\nu+3)} & 0 \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & \frac{\sigma^2(\nu+3)}{(\nu+1)\sum_{i=1}^n x_i x_i'} \end{bmatrix} \begin{bmatrix} 0 \\ \frac{-2n}{\sigma(\nu+1)(\nu+3)} \end{bmatrix}\end{aligned} \tag{A.3}$$

where $A = \frac{4}{n}\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+5)}{\nu(\nu+1)(\nu+3)}\right]^{-1}$

$$\therefore h_1 = \frac{2n\nu}{\sigma^2\left(\nu+3\right)} - \frac{4n^2}{\sigma^2\left(\nu+1\right)^2\left(\nu+3\right)}\left(\frac{4}{n}\right)\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2\left(\nu+5\right)}{\nu\left(\nu+1\right)\left(\nu+3\right)}\right]^{-1}$$

$$\therefore h_1 \propto \sigma^{-2}$$

$$\therefore h_1^{\frac{1}{2}} \propto \sigma^{-1} = p\left(\sigma\right)$$

$$\begin{aligned} H &= \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \\ &= \begin{bmatrix} \frac{2n\nu}{\sigma^2(\nu+3)} & \frac{-2n}{\sigma(\nu+1)(\nu+3)} \\ \frac{-2n}{\sigma(\nu+1)(\nu+3)} & \frac{n}{4}\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+5)}{\nu(\nu+1)(\nu+3)}\right] \end{bmatrix} \\ &\qquad - \frac{\sigma^2(\nu+3)}{(\nu+1)\sum_{i=1}^{n} x_i x_i'} \begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \end{bmatrix} \end{aligned}$$

$$\therefore h_2^{\frac{1}{2}} = H_{22}^{\frac{1}{2}} \propto \left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+5)}{\nu(\nu+1)(\nu+3)}\right]^{\frac{1}{2}}$$

We also have that $h_3^{\frac{1}{2}} \propto c$.

$$\therefore P_{REF}^5(\sigma,\nu,\beta) \propto \sigma^{-1}\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+5)}{\nu(\nu+1)(\nu+3)}\right]^{\frac{1}{2}}$$

Finally, the Fisher information matrix for the ordering $\{\sigma,\beta,\nu\}$ is given by,

$$\begin{aligned} &I(\sigma,\beta,\nu) \\ &= \begin{bmatrix} \frac{2n\nu}{\sigma^2(\nu+3)} & 0 & \frac{-2n}{\sigma(\nu+1)(\nu+3)} \\ 0 & \frac{\nu+1}{\sigma^2(\nu+3)}\sum_{i=1}^{n} x_i x_i' & 0 \\ \frac{-2n}{\sigma(\nu+1)(\nu+3)} & 0 & \frac{n}{4}\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+5)}{\nu(\nu+1)(\nu+3)}\right] \end{bmatrix} \end{aligned}$$

So we have that $h_1^{\frac{1}{2}} \propto \sigma^{-1}$, $h_2^{\frac{1}{2}} \propto c$ and $h_3^{\frac{1}{2}} \propto \psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+5)}{\nu(\nu+1)(\nu+3)}$.

$$\therefore P_{REF}^6(\sigma,\beta,\nu) \propto \sigma^{-1}\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+5)}{\nu(\nu+1)(\nu+3)}\right]^{\frac{1}{2}}$$

Probability-matching prior. The probability-matching prior, another non-informative prior, provides accurate frequentist intervals and is also used for comparisons in Bayesian analysis. Datta & Ghosh (1995) provide a method for finding probability-matching priors by deriving a differential equation that a prior must satisfy if the posterior probability of a one-sided credibility interval for a parametric function and its frequentist probability agree up to $O\left(n^{-1}\right)$, where n is the sample size.

For the probability-matching prior, $P_M\left(\nu,\beta,\sigma\right)$, we need the inverse of the Fisher information matrix,

$$I^{-1}\left(\theta\right)=I^{-1}\left(\beta,\sigma,\nu\right)=$$
$$\begin{bmatrix} \frac{\sigma^2(\nu+3)}{\nu+1}\left(\sum_{i=1}^{n}x_ix_i'\right)^{-1} & 0 & 0 \\ 0 & \frac{n}{4D}\left[\psi'\left(\frac{\nu}{2}\right)-\psi'\left(\frac{\nu+1}{2}\right)-\frac{2(\nu+5)}{\nu(\nu+1)(\nu+3)}\right] & \frac{2n}{D\sigma(\nu+1)(\nu+3)} \\ 0 & \frac{2n}{D\sigma(\nu+1)(\nu+3)} & \frac{2n\nu}{D\sigma^2(\nu+3)} \end{bmatrix}$$

where

$$\begin{aligned} D &= \frac{n^2\nu}{2\sigma^2\left(\nu+3\right)}\left[\psi'\left(\frac{\nu}{2}\right)-\psi'\left(\frac{\nu+1}{2}\right)-\frac{2\left(\nu+5\right)}{\nu\left(\nu+1\right)\left(\nu+3\right)}\right]-\frac{4n^2}{\sigma^2\left(\nu+1\right)^2\left(\nu+3\right)^2} \\ &= \frac{n^2\nu}{2\sigma^2\left(\nu+3\right)}\left[\psi'\left(\frac{\nu}{2}\right)-\psi'\left(\frac{\nu+1}{2}\right)\right]-\frac{n^2\left(\nu+5\right)}{\sigma^2\left(\nu+1\right)\left(\nu+3\right)}-\frac{4n^2}{\sigma^2\left(\nu+1\right)^2\left(\nu+3\right)^2} \\ &= \frac{n^2\nu}{2\sigma^2\left(\nu+3\right)}\left[\psi'\left(\frac{\nu}{2}\right)-\psi'\left(\frac{\nu+1}{2}\right)\right]-\frac{n^2}{\sigma^2\left(\nu+1\right)\left(\nu+3\right)^2}\left[\nu+5+\frac{4}{\nu+1}\right] \\ &= \frac{n^2\nu}{2\sigma^2\left(\nu+3\right)}\left[\psi'\left(\frac{\nu}{2}\right)-\psi'\left(\frac{\nu+1}{2}\right)\right]-\frac{n^2\left[\left(\nu+1\right)\left(\nu+5\right)+4\right]}{\sigma^2\left(\nu+1\right)^2\left(\nu+3\right)^2} \\ &= \frac{n^2\nu}{2\sigma^2\left(\nu+3\right)}\left[\psi'\left(\frac{\nu}{2}\right)-\psi'\left(\frac{\nu+1}{2}\right)\right]-\frac{n^2\left(\nu^2+6\nu+9\right)}{\sigma^2\left(\nu+1\right)^2\left(\nu+3\right)^2} \\ &= \frac{n^2\nu}{2\sigma^2\left(\nu+3\right)}\left[\psi'\left(\frac{\nu}{2}\right)-\psi'\left(\frac{\nu+1}{2}\right)\right]-\frac{n^2}{\sigma^2\left(\nu+1\right)^2} \\ &= \frac{n^2\nu}{2\sigma^2\left(\nu+3\right)}\left[\psi'\left(\frac{\nu}{2}\right)-\psi'\left(\frac{\nu+1}{2}\right)-\frac{2\left(\nu+3\right)}{\nu\left(\nu+1\right)^2}\right] \end{aligned}$$

Let $t\left(\theta\right)=\nu$, where $t\left(\theta\right)$ is the parameter of interest.

From this it follows that $\frac{\partial t(\theta)}{\partial\nu} = 1$; $\frac{\partial t(\theta)}{\partial\beta} = 0$; $\frac{\partial t(\theta)}{\partial\sigma} = 0$, and,

$$\begin{aligned}
\nabla_t'(\theta) &= \begin{bmatrix} \frac{\partial t(\theta)}{\partial\beta} & \frac{\partial t(\theta)}{\partial\sigma} & \frac{\partial t(\theta)}{\partial\nu} \end{bmatrix} \\
&= \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \\
\therefore \nabla_t'(\theta)\, I^{-1}(\theta) &= \left[\begin{matrix} 0 & \frac{2n}{D\sigma(\nu+1)(\nu+3)} & \frac{2n\nu}{D\sigma^2(\nu+3)} \end{matrix} \right),
\end{aligned}$$

which means that,

$$\left[\nabla_t'(\theta)\, I^{-1}(\theta)\, \nabla_t(\theta)\right]^{\frac{1}{2}} = \left(\frac{2n\nu}{D\sigma^2(\nu+3)}\right)^{\frac{1}{2}}.$$

$$\begin{aligned}
\zeta'(\theta) &= \frac{\nabla_t'(\theta)\, I^{-1}(\theta)}{\left[\nabla_t'(\theta)\, I^{-1}(\theta)\, \nabla_t(\theta)\right]^{\frac{1}{2}}} \\
&= \begin{bmatrix} \zeta_1(\theta) & \zeta_2(\theta) & \zeta_3(\theta) \end{bmatrix} \\
&= \begin{bmatrix} 0 & \frac{(2n)^{\frac{1}{2}}}{D^{\frac{1}{2}}\nu^{\frac{1}{2}}(\nu+1)(\nu+3)^{\frac{1}{2}}} & \frac{(2n\nu)^{\frac{1}{2}}}{D^{\frac{1}{2}}\sigma(\nu+3)^{\frac{1}{2}}} \end{bmatrix}
\end{aligned}$$

This indicates that the probability-matching prior is:

$$\begin{aligned}
p(\theta) = P_M(\nu, \beta, \sigma) &\propto D^{\frac{1}{2}} \frac{(\nu+3)^{\frac{1}{2}}}{\nu^{\frac{1}{2}}} \\
&\propto \sigma^{-1}\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+3)}{\nu(\nu+1)^2}\right]^{\frac{1}{2}}
\end{aligned}$$

because the differential equation $\frac{\partial}{\partial\beta}\left[\zeta_1(\theta)\, p(\theta)\right] + \frac{\partial}{\partial\sigma^2}\left[\zeta_2(\theta)\, p(\theta)\right] + \frac{\partial}{\partial\nu}\left[\zeta_3(\theta)\, p(\theta)\right] = 0$. The probability-matching prior is therefore the same as the reference priors for the orderings $\{\nu, \beta, \sigma\}$, $\{\beta, \nu, \sigma\}$, and $\{\nu, \sigma, \beta\}$.

Appendix B

MATLAB code

The programs run in this thesis were personally written in MATLAB (2012).

B.1 SRMI Programs for Chapters 4 and 6

Any program used in this thesis that calls the SRMI algorithm requires a fundamental set of programs. These are the first listed in this appendix. The following is the function that runs the SRMI.

```
function [ComplD, ConvD, cppp_values]=SRMI(data, idvars, explvars, ...
    models, rounds, iterations, varargin)

%varargin - step 0/1: stepwise regression, or not (default)
%          - [pmm_vector] (0/1/2/3s indicating pmm/lrd/erd, same length...
     as models), donor proportion
%models - a vector specifying which regression to use for each ...
    variable
%of the original matrix: 0:id variable 1:Normal; 1.5: t; 2:Bernoulli; ...
    3:Poisson
%4:Ordinal/categorical
%Must specify a model for each variable - even the filled variables
%step is a 0-1 specifying the use of stepwise regressions
%step is not programmed yet
prior_num = 2; %Ind Jeff prior for t
%data must be in form: [idvars data]

e = length(models);
f = e - idvars - explvars; % the num. of variables to try impute
ConvD = zeros(length(data(:,1)),e-explvars,rounds);
ComplD = zeros(length(data(:,1)),e-explvars,iterations);
if nargin>6
     cppp_test = varargin{1};
else
     cppp_test = 0;
```

```
end

if nargin>7
    step = varargin{2};
else
    step = 0;
end
if nargin>8
    adapt_vec = varargin{3};
    donor_prop = varargin{4};
else
    adapt_vec = zeros(1,e);
    donor_prop = 0;
end

[sdata J fulls] = SRMIsortmis(data(:,(idvars+explvars+1):e));

nmodels = models(:,(idvars+explvars+1):e);
adapt_vec = adapt_vec(:,(idvars+explvars+1):e);
seqreg = nmodels(:,J);
adapt_vec = adapt_vec(:,J);
cppp_values = -1*ones(2,f);


for impute = 1:iterations
    if (models(1) == 1.5) ||  (models(1) == 1.6)
        disp(impute)
    end
    %YR = [];
    if impute ≠ 1
        [sdata J fulls] = SRMIsortmis(data(:,(idvars+explvars+1):e));
    end
    %I = [];
    YR = isnan(sdata);
    %YR = isnan(sdata(:,1:fulls));
    for round = 1:rounds
        for regr = (fulls+1):f
            %fulls also have models and pmm specified
            y = sdata(:,regr);
            if explvars == 0
                if round == 1
                    %YR = [YR isnan(y)];
                    X = sdata(:,1:(regr-1)); % this will work since %...
                       there must be at least 1 full variable
                else
                    if regr == f
                        X = sdata(:,1:(regr-1));
                    else
                        X = sdata(:,[1:(regr-1) (regr+1):f]);
                    end
                end
            else
                if round == 1
                    %YR = [YR isnan(y)];
                    X = [data(:,(idvars+1):(idvars+explvars)) sdata...
                       (:,1:(regr-1))]; % this will work since %there...
                        must be at least 1 full variable
                else
                    if regr == f
                        X = [data(:,(idvars+1):(idvars+explvars)) ...
                            sdata(:,1:(regr-1))];
                    else
                        X = [data(:,(idvars+1):(idvars+explvars)) ...
                            sdata(:,[1:(regr-1) (regr+1):f])];
                    end
                end
            end
            I = find(YR(:,regr) == 1);
            adapt_switch = adapt_vec(1,regr);
            if seqreg(regr) == 1
                [M,¬,¬] = SRMInorm(y, X, I, step, adapt_switch, ...
                    donor_prop);
            elseif seqreg(regr) == 1.1
                %[M,¬,¬,¬,¬] = SRMIt(y, X, I, step, prior_num, ...
                    adapt_switch, donor_prop);
                [M,¬,¬] = SRMIlogit_norm(y, X, I, step);
            elseif seqreg(regr) == 1.5
                %[M,¬,¬,¬,¬] = SRMIt(y, X, I, step, prior_num, ...
                    adapt_switch, donor_prop);
                [M,¬,¬,¬,¬,¬,¬] = SRMItskew(y, X, I, 0, step, ...
                    prior_num, adapt_switch, donor_prop);
            elseif seqreg(regr) == 1.6
                [M,¬,¬,¬,¬,¬,¬] = SRMItskew(y, X, I, 1, step, ...
                    prior_num);
            elseif seqreg(regr) == 2
                [M] = SRMIbern(y, X, I);
            elseif seqreg(regr) == 3
                [M] = SRMIpois(y, X, I);
            elseif seqreg(regr) == 4
                [M] = SRMIcat(y, X, I);
            end
            sdata(I,regr) = M; %Updating the data matrix
        end

        if impute == 1
            unsortedData = SRMIunsort(sdata, J);
            ConvD(:,:,round) = [data(:,1:idvars) unsortedData]; %...
                explvars are dropped; their use is finished
        end
    end
    unsortedData = SRMIunsort(sdata, J);
    ComplD(:,:,impute) = [data(:,1:idvars) unsortedData]; %explvars ...
        are dropped; their use is finished
end

%ComplD is a 3D completed data matrix with the z-dimension showing ...
    filled data from
%each iteration - matrix is in rows(obs), columns(var) and depth(...
    iteration)
% The ComplD matrix does not include the variables listed as 0.5 in ...
    the
% models vector (they are dropped)

%ConvD is a 3D matrix with the z-dimension showing the 1st dataset (...
    the
%dataset during the 1st iteration) after each round of imputation.
```

Within SRMI, sorting and unsorting (by missingness) is required. The next two programs do these procedures.

```
function [sdata I fulls] = SRMIsortmis(data)
% Finding proportions of missing data:
[r c] = size(data);
R = isnan(data);

% Finding the proportion of data missing for each column.
propmiss = (sum(R)./r)'; %a column vector
fulls = length(find(propmiss == 0));
% Sorts the data according to missingness
[Tmp,I] = sortrows(propmiss, 1); %a column vector length models-advars...
    -explvars
sdata = data(:,I');

%I'

%sR = R(:,I');
```

```
function [data] = SRMIunsort(sdata, J)

[Tmp,I] = sortrows(J, 1);
data = sdata(:,I);
```

The following function is the Normal model regression in SRMI.

```
function [M,bs,sig2] = SRMInorm(y, X, I, varargin)
% Output gives a vector of imputed values, and the indices of the rows...
     of
% the data matrix where the imputed values fit into the y vector.
% varargin: step, adapt_switch (nothing=0, pmm=1, lrd=2, erd=3), ...
    adapt_prop (prop donor cases)

warning off all
adapt_switch = 0;
if nargin > 3
    step = varargin{1};
else
    step = 0;
end
if nargin > 4
    adapt_switch = varargin{2};
end
if nargin > 5
    adapt_prop = varargin{3};
end

% The regression fit

switch step
    case 0
```

```
        % if X isempty, then X is ones - this should be implemented
        X = [ones(length(X(:,1)),1) X];
        b = regress(y,X);
    case 1
        [b,se,pval,inmodel,stats,nextstep,history] = stepwisefit(X,y,'...
           display','off');
        b = [stats.intercept ; b((inmodel == 1))];
        X = [ones(length(X(:,1)),1) X(:,(inmodel == 1))];
        clear se pval stats nextstep history
end

% the missing y's, and th corresponding X's
Xmiss = X(I,:);


%the non-missing y's
YR = isnan(y);
J = find(YR == 0);
Xcompl = X(J,:);
ycompl = y(J,:);


% Generating sig
u = chi2rnd(length(ycompl)-length(Xcompl(1,:)));
sig2 = ((ycompl-Xcompl*b)'*(ycompl-Xcompl*b))./u;

% Generating B
cov = sig2*inv(Xcompl'*Xcompl);
T = (chol(cov))';
bs = b + T*randn(length(b),1);

% Now to do Predictive Mean Matching, Local Residual Draw, or Expanded
% Residual Draw if need be, or to just draw a prediction

if adapt_switch == 1 || adapt_switch == 2 || adapt_switch == 3
    M = zeros(length(I),1);
    pred_mean_compl = Xcompl*bs;
    resid_compl = ycompl-pred_mean_compl;
    for i = 1:length(I)
        pred_mean_inc = Xmiss(i,:)*bs;
        if adapt_switch == 1 || adapt_switch == 2
            [donors,D] = sort(abs(pred_mean_compl-pred_mean_inc));
            % D is the index numbers col. vector of the rows of Xcompl...
                that are
            % the closest to furthest predictive means to the
            % predictive mean of the incomplete obs
            cutoff = max(round(adapt_prop*length(donors)),1);
            donors = donors(1:cutoff,1);
            D = D(1:cutoff,1);
            % These previous 3 commands cut the donor vector down to ...
               size
            if adapt_switch == 1 %do PMM
                M(i) = ycompl(D(ceil(rand*length(donors))));
            elseif adapt_switch == 2 %do LRD
                M(i) = pred_mean_inc + resid_compl(D(ceil(rand*length(...
                   donors))));
            end
        elseif adapt_switch == 3 %do ERD
            std_resid_compl = resid_compl./sqrt(sum(resid_compl.^2)./(...
                length(resid_compl)-length(b)));
            M(i) = pred_mean_inc + sig2^(0.5).*std_resid_compl(ceil(...
                rand*length(std_resid_compl)));
        end
    end
else
    % Imputing values from generated Bs
    M = Xmiss*bs + sig2^(0.5).*randn(length(Xmiss(:,1)),1);
end
```

The following function is the t model regression in SRMI.

```
function [M,bs,sig2,v,lambdas] = SRMIt(y, X, I, varargin)
% Output gives a vector of imputed values, and the indices of the rows...
     of
% the data matrix where the imputed values fit into the y vector.
% varargin: step, prior type for v; adapt_switch (nothing=0, pmm=1, ...
    lrd=2, erd=3), adapt_prop (prop donor cases)
% v's prior type: 1: truncated exponential, 2:Independence Jeffrey's, ...
    3:
% Probability Matching or Reference for the orders {v, mu, sig2}, {v, ...
    sig2,
% mu}, {mu, v, sig2}, 4: Reference for the orders {mu, sig2, v}, {sig2...
    , v,
% mu}, {sig2, mu, v}

warning off all

if nargin > 3
    step = varargin{1};
end
if nargin > 4
    prior_num = varargin{2};
end
if nargin > 5
    adapt_switch = varargin{3};
end
if nargin > 6
    adapt_prop = varargin{4};
end

% The regression fit
X = [ones(length(X(:,1)),1) X];
%b = regress(y,X);

% the missing y's, and th corresponding X's
Xmiss = X(I,:);

%the non-missing y's
YR = isnan(y);
J = find(YR == 0);
Xcompl = X(J,:);
ycompl = y(J,:);

% Generating parameters
```

```

[beta sig2 v lambdas] = draw_gibbs_t(ycompl, Xcompl, 1);
bs = beta';

% Now to do Predictive Mean Matching, Local Residual Draw, or Expanded
% Residual Draw if need be, or to just draw a prediction

if adapt_switch == 1 || adapt_switch == 2 || adapt_switch == 3
    M = zeros(length(I),1);
    pred_mean_compl = Xcompl*bs;
    resid_compl = ycompl-pred_mean_compl;
    for i = 1:length(I)
        pred_mean_inc = Xmiss(i,:)*bs;
        if adapt_switch == 1 || adapt_switch == 2
            [donors,D] = sort(abs(pred_mean_compl-pred_mean_inc));
            % D is the index numbers col. vector of the rows of Xcompl...
                that are
            % the closest to furthest predictive means to the
            % predictive mean of the incomplete obs
            cutoff = max(round(adapt_prop*length(donors)),1);
            donors = donors(1:cutoff,1);
            D = D(1:cutoff,1);
            % These previous 3 commands cut the donor vector down to ...
               size
            if adapt_switch == 1 %do PMM
                M(i) = ycompl(D(ceil(rand*length(donors))));
            elseif adapt_switch == 2 %do LRD
                M(i) = pred_mean_inc + resid_compl(D(ceil(rand*length(...
                    donors))));
            end
        elseif adapt_switch == 3 %do ERD
            std_resid_compl = resid_compl./sqrt(sum(resid_compl.^2)./(...
                length(resid_compl)-length(X(1,:))));
            M(i) = pred_mean_inc + std_resid_compl(ceil(rand*length(...
                std_resid_compl)));
        end
    end
else
    % Imputing values from generated Bs
    M = Xmiss*bs + sig2^(0.5).*random('t',v,[length(Xmiss(:,1)),1]);
end
```

The following function is the skew t model regression in SRMI.

```
function [M,bs,tau,v,lambda,Δ,Z] = SRMItskew(y, X, I, skew, varargin)
% Output gives a vector of imputed values, and the indices of the rows...
     of
% the data matrix where the imputed values fit into the y vector.
% varargin: step, prior type for v;
% v's prior type: 1: truncated exponential, 2:Independence Jeffrey's, ...
    3:
% Probability Matching or Reference for the orders {v, mu, sig2}, {v, ...
    sig2,
% mu}, {mu, v, sig2}, 4: Reference for the orders {mu, sig2, v}, {sig2...
    , v,
% mu}, {sig2, mu, v}

```

```
warning off all
prior_num = 1; %Exponential by default.
adapt_switch = 0; %by default, no PMM, LRD, or ERD
if nargin > 4
    step = varargin{1};
end
if nargin > 5
    prior_num = varargin{2};
end
if nargin > 6
    adapt_switch = varargin{3};
end
if nargin > 7
    adapt_prop = varargin{4};
end

% The regression fit
X = [ones(length(X(:,1)),1) X];

% the missing y's, and th corresponding X's
Xmiss = X(I,:);

%the non-missing y's
YR = isnan(y);
J = find(YR == 0);
Xcompl = X(J,:);
ycompl = y(J,:);

% Generating parameters

[beta, tau, v, lambda, ∆, Z] = draw_gibbs_t_skew(ycompl, Xcompl, 1, ...
    skew, prior_num);
bs = beta';

if (adapt_switch == 1 || adapt_switch == 2 || adapt_switch == 3) && (...
    skew == 0)
    M = zeros(length(I),1);
    pred_mean_compl = Xcompl*bs;
    resid_compl = ycompl-pred_mean_compl;
    for i = 1:length(I)
        pred_mean_inc = Xmiss(i,:)*bs;
        if adapt_switch == 1 || adapt_switch == 2
            [donors,D] = sort(abs(pred_mean_compl-pred_mean_inc));
            % D is the index numbers col. vector of the rows of Xcompl...
                that are
            % the closest to furthest predictive means to the
            % predictive mean of the incomplete obs
            cutoff = max(round(adapt_prop*length(donors)),1);
            donors = donors(1:cutoff,1);
            D = D(1:cutoff,1);
            % These previous 3 commands cut the donor vector down to ...
              size
            if adapt_switch == 1 %do PMM
                M(i) = ycompl(D(ceil(rand*length(donors))));
            elseif adapt_switch == 2 %do LRD
                M(i) = pred_mean_inc + resid_compl(D(ceil(rand*length(...
                    donors))));
            end
        elseif adapt_switch == 3 %do ERD
            std_resid_compl = resid_compl./sqrt(sum(resid_compl.^2)./(...
                length(resid_compl)-length(X(1,:))));
            M(i) = pred_mean_inc + std_resid_compl(ceil(rand*length(...
                std_resid_compl)));
        end
    end
else
    M = Xmiss*bs + ∆*abs(randn([length(Xmiss(:,1)),1])) + tau^(-0.5).*...
        random('t',v,[length(Xmiss(:,1)),1]);
end
```

Whenever the t model is fitted, the following program is called.

```
function [beta sig2 v lambdas] = draw_gibbs_t(y, X, draws, varargin)
% y is a column vector
% vs was an additional output originally
if nargin > 3
    prior_num = varargin{1};
else
    prior_num = 1;
end
if nargin > 4
    v_discrete = varargin{2};
else
    v_discrete = 1;
end
if nargin > 5
    burn_in = varargin{3};
else
    burn_in = 200;
end
%% Core program
[n p] = size(X);


sig2 = ones((burn_in+draws),1);
lambdas = ones((burn_in+draws),n);
beta = ones((burn_in+draws),p);
v = ones((burn_in+draws),1)*3;

% initial values

[¬,¬,¬,¬,thestats] = regress(y,X);
sig2(1) = thestats(4);

beta(1,:) = (sqrtm(sig2(1).*(X'*diag(lambdas(1,:))*X))\randn(p,1) + (X...
    '*diag(lambdas(1,:))*X)\X'*diag(lambdas(1,:))*y)';


if v_discrete
    [v_draws] = draw_post_v_discr(1, n, lambdas(1,:), prior_num);
else
    [v_draws] = draw_post_v(1, n, lambdas(1,:), prior_num);
end
v(1) = v_draws(end);
```

```
%vs = [];

for i = 2:(burn_in+draws)
    sig2(i) = (y-X*beta(i-1,:)')' * diag(lambdas(i-1,:)) * (y-X*beta(i...
        -1,:)') / chi2rnd(n);
%     sig2(i) = 1;
    beta(i,:) = (sqrtm(sig2(i)*(X'*diag(lambdas(i-1,:))*X))\randn(p,1)...
        + (X'*diag(lambdas(i-1,:))*X)\X'*diag(lambdas(i-1,:))*y)';

    lambdas(i,:) = chi2rnd(v(i-1)+1,[1,n]) ./ (v(i-1)+((y-X*beta(i,:)...
        ')').^2./sig2(i));
    %%lambdas(i,:) = gamrnd(0.5*(v(i-1)+1),0.5,[1,n]) ./ (v(i-1)+((y-X...
        *beta(i,:)')').^2./sig2(i));
    if v_discrete
         [v(i)] = draw_post_v_discr(1, n, lambdas(i,:), prior_num,i);
    else
         [v(i)] = draw_post_v(1, n, lambdas(i,:), prior_num);
    end
    %vs = [vs ; v_draws];
end

% figure(1)
% plot((1:(burn_in+draws))',v)
% title('v draws')
% figure(2)
% plot((1:(burn_in+draws))',sig2)
% title('sig2 draws')
% plot((1:(burn_in+draws))',lambdas)
% title('lambdas draws')
% figure(3)
% hist(v)
% title('Histogram of v draws')
% figure(4)
% hist(sig2)
% title('Histogram of sig2 draws')
% for i = 1:p
%     figure(4+i)
%     hist(beta(:,i))
%     title(strcat(['Histogram of draws for beta ' num2str(p)]))
% end

beta = beta((burn_in+1):end,:);
sig2 = sig2((burn_in+1):end,1);
lambdas = lambdas((burn_in+1):end,:);
v = v((burn_in+1):end,1);

% beta = beta(1:end,:);
% sig2 = sig2(1:end,1);
% lambdas = lambdas(1:end,:);
% v = v(1:end,1);
end
```

Likewise, if the skew t model is fitted, the following program is called.

```
function [beta tau v lambdas ∆ Z] = draw_gibbs_t_skew(y, X, draws, ...
    varargin)
% y is a column vector
% vs was an additional output originally
% skew is a 0 (for symmetric t) or 1 for skew t.
%X includes a column of ones for the intercept term
if nargin > 3
    skew = varargin{1};
else
    skew = 1;
end
if nargin > 4
    prior_num = varargin{2};
else
    prior_num = 1;
end
if nargin > 5
    v_discrete = varargin{3};
else
    v_discrete = 0;
end
if nargin > 6
    burn_in = varargin{4};
else
    burn_in = 200;
end

%% Core program
[n p] = size(X);

%Priors

mu∆ = 0; %these are the parameters on the proper Normal prior for ∆
sig2∆ = 1000;

mubeta = 0*ones(1,p);
sig2beta = 10000*ones(1,p);

agamma = 0.1;
lgamma = 0.1;

muz = 0;
sig2z = 1;


%Initialise vectors
%Initial values
tau = zeros((burn_in+draws),1);
[¬,¬,¬,¬,thestats] = regress(y,X);
tau(1) = 1/thestats(4);
lambdas = ones((burn_in+draws),n);
beta = zeros((burn_in+draws),p);
beta(1,:)=(sqrtm((tau(1)^(-1)).*(X'*diag(lambdas(1,:))*X))\randn(p,1) ...
    + (X'*diag(lambdas(1,:))*X)\X'*diag(lambdas(1,:))*y)';
Z = repmat(abs(randn(1,n)),burn_in+draws,1).*ones((burn_in+draws),n);
∆ = zeros((burn_in+draws),1); %0 is the prior mean for ∆.
v = 3*ones((burn_in+draws),1);
if v_discrete
    [v_draws] = draw_post_v_discr(1, n, lambdas(1,:), prior_num);
else
```

```
    [v_draws] = draw_post_v(1, n, lambdas(1,:), prior_num);
end
v(1) = v_draws(end);

ystar = zeros(n,p);
ycurl = zeros(n,burn_in+draws);

for i = 2:(burn_in+draws)
    if p>1
        for j = 1:p
            switch j
                case 1
                    ystar(:,j) = y - X(:,(j+1):end)*beta((i-1),(j+1)...
                        :end)' - ∆(i-1)*Z((i-1),:)';
                case p
                    ystar(:,j) = y - X(:,1:(end-1))*beta((i),1:(end-1)...
                        )' - ∆(i-1)*Z((i-1),:)';
                otherwise
                    ystar(:,j) = y - X(:,[1:(j-1) (j+1):end])*[beta((i...
                        ),1:(j-1)) beta((i-1),(j+1):end)]' - ∆(i-1)*Z...
                        ((i-1),:)';
            end
            beta(i,j) = randn.* (tau(i-1)*lambdas((i-1),:)*(X(:,j).^2)...
                +1/sig2beta(j))^(-0.5) + (tau(i-1).*lambdas((i-1),:)*(...
                X(:,j).^2)+1/sig2beta(j))^(-1) * (tau(i-1).*lambdas((i...
                -1),:)*(X(:,j).*ystar(:,j))+mubeta(j)/sig2beta(j));
        end
    else
        ystar(:,1) = y - X*beta(i-1) - ∆(i-1)*Z((i-1),:)';
        %and beta? I added this line
        beta(i,1) = randn.* (tau(i-1)*lambdas((i-1),:)*(X(:,j).^2)+1/...
            sig2beta(j))^(-0.5) + (tau(i-1).*lambdas((i-1),:)*(X(:,j)....
            ^2)+1/sig2beta(j))^(-1) * (tau(i-1).*lambdas((i-1),:)*(X...
            (:,j).*ystar(:,j))+mubeta(j)/sig2beta(j));
    end
    ycurl(:,i) = y - X*beta(i,:)' - ∆(i-1)*Z((i-1),:)';
    tau(i) = gamrnd(n/2 + agamma,(0.5*lambdas((i-1),:)*(ycurl(:,i).^2)...
     + 2*lgamma)^-1);
    if skew == 1
        yhat = y - X*beta(i,:)';
        Z(i,:) = trunc_N(0,Inf,(tau(i).*lambdas(i-1,:).*∆(i-1)^2 +1)....
            ^-1 .* tau(i).*lambdas(i-1,:).*∆(i-1).*yhat',sqrt((tau(i)....
            *lambdas(i-1,:).*∆(i-1)^2  +1      ).^-1),1);
        ∆(i) = randn* (tau(i)*lambdas((i-1),:)*(Z(i,:).^2)'+1/sig2z)...
            ^(-0.5) + (tau(i)*lambdas((i-1),:)*(Z(i,:).^2)'+1/sig2z)...
            ^(-1) * (tau(i)*lambdas((i-1),:)*(Z(i,:)'.*yhat) + muz/...
            sig2z);
        ycurl(:,i) = y - X*beta(i,:)' - ∆(i)*Z((i),:)';
    end
    lambdas(i,:) = chi2rnd(v(i-1)+1,[1,n]) ./ (v(i-1)+tau(i)*(ycurl(:,...
        i)').^2);
    [v(i)] = draw_post_v_discr(1, n, lambdas(i,:), prior_num,i); %...
        discretised v or not?
 end

beta = beta((burn_in+1):end,:);
tau = tau((burn_in+1):end,1);
```

```
lambdas = lambdas((burn_in+1):end,:);
v = v((burn_in+1):end,1);
Z = Z((burn_in+1):end,:);
Δ = Δ((burn_in+1):end,1);
```

The Metropolis-Hastings sampler for the degrees of freedom in the t or skew t is in the following program.

```
function [v_draws] = draw_post_v(draws, n, lambdas, varargin)
%burn-ins are passed through as well
% lambdas are the draws from the gibbs sampler for the other
% parameters
% v's prior type: 1: truncated exponential, 2:Independence Jeffrey's, ...
   3:
% Probability Matching or Reference for the orders {v, mu, sig2}, {v, ...
   sig2,
% mu}, {mu, v, sig2}, 4: Reference for the orders {mu, sig2, v}, {sig2...
   , v,
% mu}, {sig2, mu, v}

if nargin > 3
    prior_num = varargin{1};
else
    prior_num = 1; %truc exp default
end
jumpsize = 1;
burn_in = 50;
v_draws = zeros(1,(burn_in+draws));
v_old = 5; %10*rand+4; %Pre-seed value
logpost_v_old = -Inf;
i = 0;
j = 0;
while i < (burn_in + draws)
    v = v_old + randn*jumpsize;
    while v < 2
        v = v_old + randn*jumpsize;
    end

% v = rand*40;
    j = j + 1;
    switch prior_num
        case 1
            if v > 2 % shouldn't be used, since the previous check ...
               ensures v>2
                logprior_v = -v; % v > 2
            else
                logprior_v = 0;
            end
        case 2
            logprior_v = 0.5*(log(v)-log(v+3)) + 0.5*log(psi(1,v/2)-...
               psi(1,(v+1)/2)-2*(v+3)/(v*(v+1)^2));
        case 3
            logprior_v = 0.5*log(psi(1,v/2)-psi(1,(v+1)/2)-2*(v+3)/(v...
               *(v+1)^2));
        case 4
            logprior_v = 0.5*log(psi(1,v/2)-psi(1,(v+1)/2)-2*(v+5)/(v...
               *(v+1)*(v+3)));
    end
    logpost_v = logprior_v + 0.5*v*n*log(v) - 0.5*v*n*log(2) - n*...
        gammaln(v/2) + (0.5*v-1)*sum(log(lambdas)) - 0.5*v*sum(lambdas...
        );
    if ((logpost_v - logpost_v_old) > log(rand))
        i = i + 1;
        v_draws(i) = v;
        v_old = v;
        logpost_v_old = logpost_v;
%     else
%         v_draws(i) = v_old;
    end
end
v_draws = v_draws(1,(burn_in+1):end); %drop the burn-in v's
%acceptrate = i/j
```

To save time in the Gibbs sampler, the degrees of freedom draws are discretised in the following program.

```
function [v_draws] = draw_post_v_discr(draws, n, lambdas, varargin)
%burn-ins are passed through as well
% lambdas are the draws from the gibbs sampler for the other
% parameters
% v's prior type: 1: truncated exponential, 2:Independence Jeffrey's, ...
    3:
% Probability Matching or Reference for the orders {v, mu, sig2}, {v, ...
    sig2,
% mu}, {mu, v, sig2}, 4: Reference for the orders {mu, sig2, v}, {sig2...
    , v,
% mu}, {sig2, mu, v}
if nargin > 3
    prior_num = varargin{1};
else
    prior_num = 1; %exponential
end
if nargin > 4
    gibbs_round = varargin{2};
else
    gibbs_round = 0;
end
v_draws = zeros(1,draws);
% v = [2.1:0.1:120 150 250 500 1000]'; %maybe chop up into 0.01 ...
    intervals
v = logspace(0.3223,3,200)';
for i = 1:draws
    switch prior_num
        case 1
            logprior_v = -v; % v > 2
        case 2
            logprior_v = 0.5.*(log(v)-log(v+3)) + 0.5.*log(psi(1,v./2)...
                -psi(1,(v+1)./2)-2.*(v+3)./(v.*(v+1).^2));
        case 3
            logprior_v = 0.5.*log(psi(1,v./2)-psi(1,(v+1)./2)-2.*(v+3)...
                ./(v.*(v+1).^2));
        case 4
            logprior_v = 0.5.*log(psi(1,v./2)-psi(1,(v+1)./2)-2.*(v+5)...
                ./(v.*(v+1).*(v+3)));
    end
    logpost_v =  logprior_v  + 0.5.*v.*n.*log(v/2) - n.*gammaln(v./2) ...
       + (0.5.*v-1).*sum(log(lambdas)) - 0.5.*v.*sum(lambdas);
    post_v = exp(logpost_v - max(logpost_v));
    cum_post_v = cumsum(post_v);
    cum_post_v = cum_post_v./cum_post_v(end);
    cum_post_v = [0  cum_post_v']';
    choice = find(cum_post_v < rand,1,'last');
    v_draws(i) = v(choice);
    if  (gibbs_round > 0) && (gibbs_round < 31);
        if v_draws(i) < 3
            v_draws(i) = 3;
        end
        if v_draws(i) > 30
            v_draws(i) = 30;
        end
    end
end
```

B.2 Programs for Chapter 4

The following program runs the first analysis for Chapter 4.

```
% In this program I will generate 3 n-by-4 datasets.
% Case 1: y1¬N. y2|y1¬N. y3|y1,y2¬N. y4|y1,y2,y3¬N.
% MAR randomness generated using logit with logit_betas as parameters.
% y1 complete, y2M¬logit(y1), y3M¬logit(y1,y2), y4M¬logit(y1,y2,y3),
% but if any of the logit arguments are missing, they are ignored.
% This part of project 1 will do the MSE coverage graphs.
% 100 incomplete data sets from MCAR and from MAR are created, and in ...
    each
% 5 iterations are imputed. The MSE's of each idata, cpart_idata, and ...
    the
% imputation methods are created.
% In v2, I will pool the 5 cdatas for one scenario [cdata(:,:,1);
% cdata(:,:,2) ...], then do mse's of percentiles. So one completed ...
    data
% set for each missingness scenario
clear
clc
start = tic;
n = 1000; %dataset size
draws = 500;
mi_iterations = 5;
mi_rounds = 15;
simulations = 100;
logit_betas = [-0.3 -0.3 -0.3 -0.3];
mcar_prop = 0.2;
data_scenarios = 5;
mi_options = 9;
original_data_cell = cell(data_scenarios,1); % 1 for each ...
    data_scenario
idata_cell = cell(1,2,data_scenarios);
cpart_idata_cell = cell(1,2,data_scenarios);
completed_data_cell = cell(9,2,simulations,data_scenarios);
parameter_text = {'n =', n;
    'draws =', draws;
    'mi_iterations =', mi_iterations;
    'logit_betas =', NaN;
    'mcar_prop =', mcar_prop;
    NaN, NaN;
    'MCAR missing y1', NaN; % line 7
    'MCAR missing y2', NaN;
    'MCAR missing y3', NaN;
    'MCAR missing y4', NaN;
    'MCAR cpart miss', NaN;
    'MAR missing y1', NaN;
    'MAR missing y2', NaN;
    'MAR missing y3', NaN;
    'MAR missing y4', NaN;
    'MAR cpart miss', NaN;
    NaN, NaN;
    'MSE QQ Y2',NaN;
    'MDM MCAR',NaN;
    'incomplete',NaN; % now line 20
    'case deleted',NaN;
    'N',NaN;
    'N_{pmm}',NaN;
    'N_{lrd}',NaN;
    'N_{erd}',NaN;
    't',NaN;
    't_{pmm}',NaN;
    't_{lrd}',NaN;
    't_{erd}',NaN;
    't_{skew}',NaN;
    NaN, NaN;
    'MDM MAR',NaN;
    'incomplete',NaN; % now line 33
    'case deleted',NaN;
    'N',NaN;
    'N_{pmm}',NaN;
    'N_{lrd}',NaN;
    'N_{erd}',NaN;
    't',NaN;
    't_{pmm}',NaN;
    't_{lrd}',NaN;
    't_{erd}',NaN;
    't_{skew}',NaN;
    NaN, NaN;
    'MSE QQ Y3',NaN;
    'MDM MCAR',NaN;
    'incomplete',NaN; % now line 47
    'case deleted',NaN;
    'N',NaN;
    'N_{pmm}',NaN;
    'N_{lrd}',NaN;
    'N_{erd}',NaN;
    't',NaN;
    't_{pmm}',NaN;
    't_{lrd}',NaN;
    't_{erd}',NaN;
    't_{skew}',NaN;
    NaN, NaN;
    'MDM MAR',NaN;
    'incomplete',NaN; % now line 60
    'case deleted',NaN;
    'N',NaN;
    'N_{pmm}',NaN;
    'N_{lrd}',NaN;
    'N_{erd}',NaN;
    't',NaN;
    't_{pmm}',NaN;
    't_{lrd}',NaN;
    't_{erd}',NaN;
    't_{skew}',NaN;
    NaN, NaN;
    'MSE Y4',NaN;
    'MDM MCAR',NaN;
    'incomplete',NaN; % now line 74
    'case deleted',NaN;
    'N',NaN;
    'N_{pmm}',NaN;
    'N_{lrd}',NaN;
```

```
       'N_{erd}',NaN;
       't',NaN;
       't_{pmm}',NaN;
       't_{lrd}',NaN;
       't_{erd}',NaN;
       't_{skew}',NaN;
       NaN,NaN;
       'MDM MAR',NaN;
       'incomplete',NaN; % now line 87
       'case deleted',NaN;
       'N',NaN;
       'N_{pmm}',NaN;
       'N_{lrd}',NaN;
       'N_{erd}',NaN;
       't',NaN;
       't_{pmm}',NaN;
       't_{lrd}',NaN;
       't_{erd}',NaN;
       't_{skew}',NaN;
       NaN, NaN;
       'MDM MCAR',NaN; % line 18, now 99
       'idata_KS_h_2',NaN;
       'cpart_KS_h_2',NaN;
       'cdata_KS_h_2',NaN;
       NaN, NaN;
       'idata_KS_h_3',NaN;
       'cpart_KS_h_3',NaN;
       'cdata_KS_h_3',NaN;
       NaN, NaN;
       'idata_KS_h_4',NaN;
       'cpart_KS_h_4',NaN;
       'cdata_KS_h_4',NaN;
       NaN, NaN;
       'MDM MAR',NaN; % line 31, now 112
       'idata_KS_h_2',NaN;
       'cpart_KS_h_2',NaN;
       'cdata_KS_h_2',NaN;
       NaN, NaN;
       'idata_KS_h_3',NaN;
       'cpart_KS_h_3',NaN;
       'cdata_KS_h_3',NaN;
       NaN, NaN;
       'idata_KS_h_4',NaN;
       'cpart_KS_h_4',NaN;
       'cdata_KS_h_4',NaN;
       NaN, NaN;
       'MSE IMP QQ Y2 Y3 Y4',NaN;
       'MDM MCAR',NaN;
       'N',NaN; % now line 127
       'N_{pmm}',NaN;
       'N_{lrd}',NaN;
       'N_{erd}',NaN;
       't',NaN;
       't_{pmm}',NaN;
       't_{lrd}',NaN;
       't_{erd}',NaN;
       't_{skew}',NaN;
```

```
    NaN, NaN;
    'MDM MAR',NaN;
    'N',NaN; % now line 138
    'N_{pmm}',NaN;
    'N_{lrd}',NaN;
    'N_{erd}',NaN;
    't',NaN;
    't_{pmm}',NaN;
    't_{lrd}',NaN;
    't_{erd}',NaN;
    't_{skew}',NaN};

mi_option_text = cell(14,9);
mi_option_text(1,:) = {'N','N_{pmm}','N_{lrd}','N_{erd}','t','t_{pmm}'...
    ,'t_{lrd}','t_{erd}','t_{skew}'};
mi_option_text(14,:) = {'N','N_{pmm}','N_{lrd}','N_{erd}','t','t_{pmm}...
    ','t_{lrd}','t_{erd}','t_{skew}'};

xlswrite('Project_01.xlsx',parameter_text,'Sheet1','A1');
xlswrite('Project_01.xlsx',logit_betas,'Sheet1','B4');
xlswrite('Project_01.xlsx',mi_option_text,'Sheet1','B99');
xlswrite('Project_01.xlsx',parameter_text,'Sheet2','A1');
xlswrite('Project_01.xlsx',logit_betas,'Sheet2','B4');
xlswrite('Project_01.xlsx',mi_option_text,'Sheet2','B99');
xlswrite('Project_01.xlsx',parameter_text,'Sheet3','A1');
xlswrite('Project_01.xlsx',logit_betas,'Sheet3','B4');
xlswrite('Project_01.xlsx',mi_option_text,'Sheet3','B99');
xlswrite('Project_01.xlsx',parameter_text,'Sheet4','A1');
xlswrite('Project_01.xlsx',logit_betas,'Sheet4','B4');
xlswrite('Project_01.xlsx',mi_option_text,'Sheet4','B99');
xlswrite('Project_01.xlsx',parameter_text,'Sheet5','A1');
xlswrite('Project_01.xlsx',logit_betas,'Sheet5','B4');
xlswrite('Project_01.xlsx',mi_option_text,'Sheet5','B99');

for data_scenario = 1:data_scenarios
    scenario_timer = tic;
    switch data_scenario
        case 1 % Normality
            y1 = randn(n,1);
            y2 = 1 + y1 + randn(n,1);
            y3 = 1 + y1 + y2 + randn(n,1);
            y4 = 1 + y1 + y2 + y3 + randn(n,1);
            Y = [y1 y2 y3 y4]; clear y1 y2 y3 y4
            Original_Data = Y;
        case 2 %
            y1 = randn(n,1);
            y2 = 1 + y1 + trnd(6,n,1);
            y3 = 1 + y1 + y2 + trnd(6,n,1) - randn(n,1);
            y4 = 1 + y1 + y2 + y3 + trnd(6,n,1) - 2*randn(n,1);
            Y = [y1 y2 y3 y4]; clear y1 y2 y3 y4
            Original_Data = Y;
        case 3
            y1 = randn(n,1);
            y2 = 1 + y1 + trnd(3,n,1);
            y3 = 1 + y1 + y2 + trnd(3,n,1) - randn(n,1);
            y4 = 1 + y1 + y2 + y3 + trnd(3,n,1) - 2*randn(n,1);
            Y = [y1 y2 y3 y4]; clear y1 y2 y3 y4
            Original_Data = Y;
        case 4
            y1 = randn(n,1);
            u2 = rand(n,1);
            y2 = 1 + y1 + (exp(u2)-1).*exp(-0.125*u2.^2); % ...
               exponential type
            u3 = rand(n,1);
            y3 = 1 + y1 + y2 + (exp(0.75.*u3)-1)./0.75.*exp(0.125*u3....
               ^2); % skew type
            u4 = rand(n,1);
            y4 = 1 + y1 + y2 + y3 + (exp(u4)-1); % lognormal
            Y = [y1 y2 y3 y4]; clear y1 y2 y3 y4 u2 u3 u4
            Original_Data = Y;
        case 5
            y1 = randn(n,1);
            e2 = 1+exp(1+randn(n,1));
            e2 = e2-mean(e2);
            e2 = e2/std(e2) * sqrt(var(y1)*3);
            y2 = 1 + y1 + e2;
            e3 = trnd(3,n,1);
            e3 = e3 - mean(e3);
            e3 = e3/std(e3) * sqrt(2*(var(y1)+var(y2)));
            y3 = 1 + y1 + y2 + e3;
            e4 = trnd(3,n,1)-2*randn(n,1);
            e4 = e4 -mean(e4);
            e4 = e4/std(e4) * (var(y1)+var(y2)+var(y3));
            y4 = 1 + y1 + y2 + y3 +e4;
            Y = [y1 y2 y3 y4]; clear y1 y2 y3 y4 e2 e3 e4
            Original_Data = Y;
    end
    original_data_cell{data_scenario,1} = Original_Data;
    [n p] = size(Y);
    missingness = NaN*ones(10,simulations);
    % MCAR y1
    % MCAR y2
    % MCAR y3
    % MCAR y4
    % MCAR cpart_idata
    % MAR y1
    % MAR y2
    % MAR y3
    % MAR y4
    % MAR cpart_idata

    for miss_mech = 1:2;

        ks_test_slice = zeros(11,9);
        ks_test = NaN*ones(11,9); %See output file
        ks_test(:,1) = zeros(11,1);
        ks_test(3,:) = zeros(1,9);
        ks_test(7,:) = zeros(1,9);
        ks_test(11,:) = zeros(1,9);

        idata_3D = ones(n,p,simulations);
        cpart_3D = NaN*ones(n,p,simulations);

        for sim = 1:simulations
```

```
% C is the binary matrix with 1 indicating observed and 0
% indicating missing (completeness matrix).

switch miss_mech
    case 1 % MCAR missingness
        idata = Y;
        P_M = [zeros(n,1) rand(n,3)];
        C = ones(n,4);
        for col = 2:4  %variables with missingness
            %U = rand(n,1);
            C((P_M(:,col)<mcar_prop),col) = zeros(sum(P_M...
               (:,col)<mcar_prop),1);
        end
        % Calculate missingness:
        idata(C==0)=NaN; % incomplete data
        cpart_idata = idata(isfinite(sum(idata,2)),:); % ...
           Case deleted data
        missingness(1:4,sim) = (sum(isnan(idata))./n)';
        missingness(5,sim) = 1- length(cpart_idata(:,1))/n...
           ;
    case 2 % MAR missingness
        idata = Y;
        P_M = zeros(n,4); %probability missing
        C = ones(n,4);
        for col = 2:4 %variables with missingness
            a = logit_betas(1) * ones(n,1);
            for i = 1:(col-1)
                a = nansum([a logit_betas(i+1)*(idata(:,i)...
                   -nanmean(idata(:,i)))./nanstd(idata(:,...
                   i))],2);
            end
            P_M(:,col) = 0.4./(1+exp(-a));
            U = rand(n,1);
            C((U<P_M(:,col)),col) = zeros(sum(U<P_M(:,col)...
               ),1);
            idata(C==0)=NaN; % incomplete data
        end
        % Calculate missingness:
        % idata(C==0)=NaN; % incomplete data
        cpart_idata = idata(isfinite(sum(idata,2)),:); % ...
           Case deleted data
        missingness(6:9,sim) = (sum(isnan(idata))./n)';
        missingness(10,sim) = 1- length(cpart_idata(:,1))/...
           n;
end
idata_3D(:,:,sim) = idata;
cpart_3D(1:length(cpart_idata(:,1)),:,sim) = cpart_idata;

% Calculate contribution to kstest reject proportion in
% this simulation, for idata and cpart_idata
ks_test_slice(1,1) = kstest2(Original_Data(:,2),idata(:,2)...
   )/simulations;
ks_test_slice(2,1) = kstest2(Original_Data(:,2),...
   cpart_idata(:,2))/simulations;
ks_test_slice(5,1) = kstest2(Original_Data(:,3),idata(:,3)...
   )/simulations;
ks_test_slice(6,1) = kstest2(Original_Data(:,3),...
   cpart_idata(:,3))/simulations;
ks_test_slice(9,1) = kstest2(Original_Data(:,4),idata(:,4)...
   )/simulations;
ks_test_slice(10,1) = kstest2(Original_Data(:,4),...
   cpart_idata(:,4))/simulations;

M = isnan(idata);
% idata is then used for the MI part
for mi_option = 1:mi_options
    disp(strcat(['Data_Scenario ' num2str(data_scenario) '...
       , MI_option ' num2str(mi_option), ', Miss_Mech ' ...
       num2str(miss_mech)]))
    switch mi_option
        case 1
            [cdata,¬,¬]=SRMI(idata,0,0,[1 1 1 1],mi_rounds...
               ,mi_iterations);
        case 2
            [cdata,¬,¬]=SRMI(idata,0,0,[1 1 1 1],mi_rounds...
               ,mi_iterations,0,0,[1 1 1 1],0.1);
        case 3
            [cdata,¬,¬]=SRMI(idata,0,0,[1 1 1 1],mi_rounds...
               ,mi_iterations,0,0,[2 2 2 2],0.1);
        case 4
            [cdata,¬,¬]=SRMI(idata,0,0,[1 1 1 1],mi_rounds...
               ,mi_iterations,0,0,[3 3 3 3],0.1);
        case 5
            [cdata,¬,¬]=SRMI(idata,0,0,[1.5 1.5 1.5 1.5],...
               mi_rounds,mi_iterations);
        case 6
            [cdata,¬,¬]=SRMI(idata,0,0,[1.5 1.5 1.5 1.5],...
               mi_rounds,mi_iterations,0,0,[1 1 1 1],0.1)...
               ;
        case 7
            [cdata,¬,¬]=SRMI(idata,0,0,[1.5 1.5 1.5 1.5],...
               mi_rounds,mi_iterations,0,0,[2 2 2 2],0.1)...
               ;
        case 8
            [cdata,¬,¬]=SRMI(idata,0,0,[1.5 1.5 1.5 1.5],...
               mi_rounds,mi_iterations,0,0,[3 3 3 3],0.1)...
               ;
        case 9
            [cdata,¬,¬]=SRMI(idata,0,0,[1.6 1.6 1.6 1.6],...
               mi_rounds,mi_iterations);
    end
    completed_data_cell{mi_option,miss_mech,sim,...
       data_scenario} = cdata;
    %kstest
    for j = 2:4
        for i = 1:mi_iterations
            ks_test_slice(4*(j-1)-1,mi_option) = ...
               ks_test_slice(4*(j-1)-1, mi_option) + ...
               kstest2(Original_Data(:,j),cdata(:,j,i))/(...
               mi_iterations*simulations);
        end
    end

end
```

```
            disp(strcat(['Simulation ' num2str(sim) ', finished with ...
               MI option ' num2str(mi_option) ', MDM ' num2str(...
               miss_mech) ' on data scenario ' num2str(data_scenario)...
               ]))
            ks_test(:,:) = ks_test(:,:) + ks_test_slice(:,:);
        end
        idata_cell{1,miss_mech,data_scenario} = idata_3D;
        cpart_idata_cell{1,miss_mech,data_scenario} = cpart_3D;
        xlswrite('Project_01.xlsx',ks_test(:,:),strcat(['Sheet' ...
           num2str(data_scenario)]),strcat(['B' num2str(100+13*(...
           miss_mech-1))]));
    end
    xlswrite('Project_01.xlsx',missingness,strcat(['Sheet' num2str(...
       data_scenario)]),'B7');
    disp(strcat(['Finished with page ' num2str(data_scenario)]))
    toc(scenario_timer)
    save 'D:\Workspace_P1_v2_n1000.mat'
end
%now that I have all the data arrays, I must find the mse's of the qq-...
   plots
%for the columns within each cell (against Original_Data columns)
%Original_Data n x p
%idata_cell 1,2,4 each with n x p
%cpart_idata_cell 1,2,4 each with n x p
%cdata_cell 9,2,4 each with n x p
disp('Starting MSE calculations of distribution QQ plots')
mse_QQ_i = cell(1,2,data_scenarios);
mse_QQ_c = cell(1,2,data_scenarios);
% Preallocation
for data_scenario = 1:data_scenarios
    for miss_mech = 1:2;
        mse_QQ_i{1,miss_mech,data_scenario} = zeros(2,simulations,3);
        mse_QQ_c{1,miss_mech,data_scenario} = zeros(9,simulations,3);
    end
end

for data_scenario = 1:data_scenarios
    target_pctiles = prctile(original_data_cell{data_scenario}(:,2:end...
       ),1:99);
    for miss_mech = 1:2;

        for sim = 1:simulations
            %idata mses
            mse_QQ_i{1,miss_mech,data_scenario}(1,sim,:) = permute(...
               mean((target_pctiles - prctile(idata_cell{1,miss_mech,...
               data_scenario}(:,2:end,sim),1:99)).^2),[1 3 2]);
            %cpart_idata mses
            % Variable 1 might have a different distribution!
            mse_QQ_i{1,miss_mech,data_scenario}(2,sim,:) =  permute(...
               mean((target_pctiles - prctile(cpart_idata_cell{1,...
               miss_mech,data_scenario}(:,2:end,sim),1:99)).^2),[1 3 ...
               2]);
            %here I stack the mi_iterations of cdata on top of one ...
               another,
            %without the complete first variable
            tmp = zeros(mi_iterations*n,p-1);
            for mi_option = 1:mi_options
```

```
                for mi_iteration = 1:mi_iterations
                    tmp(((mi_iteration-1)*n+1):(mi_iteration*n),:) = ...
                        completed_data_cell{mi_option,miss_mech,sim,...
                        data_scenario}(:,2:end,mi_iteration);
                end
                %cdata mses
                mse_QQ_c{1,miss_mech,data_scenario}(mi_option,sim,:) =...
                      permute(mean((target_pctiles - prctile(tmp,1:99)...
                    ).^2),[1 3 2]);
            end

        end
    end
end
% now we have mse matrices (in a 1 x 2 x 4 cell) that are 11 x
% mi_iterations x 3 (columns in depth)
disp('Finished with MSE calculations of distribution QQ plots')

% MSE Data writer
mse_QQ_cell = cell(1,2,data_scenarios);
for data_scenario = 1:data_scenarios
    for miss_mech = 1:2
        mse_QQ = NaN*ones(11,simulations,3);
        mse_QQ(1:2,:,:) = mse_QQ_i{1,miss_mech,data_scenario};
        mse_QQ(3:11,:,:) = mse_QQ_c{1,miss_mech,data_scenario};
        for col = 1:3
            xlswrite('Project_01.xlsx',mse_QQ(:,:,col),strcat(['Sheet'...
                 num2str(data_scenario)]),strcat(['B' num2str(20+(col...
                -1)*27+(miss_mech-1)*13)]));
        end
        mse_QQ_cell{1,miss_mech,data_scenario} = mse_QQ;
    end
end

toc(start)
%Plotting bar graphs for mses
label_array = {'incomplete','case deleted','N','N_{pmm}','N_{lrd}','N_...
   {erd}','t','t_{pmm}','t_{lrd}','t_{erd}','t_{skew}'};
for data_scenario = 1:data_scenarios
    for miss_mech = 1:2
        figure
        title(strcat(['Imputation coverage intervals for data scenario...
             ' num2str(data_scenario) ' and MDM ' num2str(miss_mech)])...
            )
        for i = 1:3 % the number of columns with missingness
            subplot(3,1,i)
            boxplot(mse_QQ_cell{1,miss_mech,data_scenario}(:,:,i)','...
                labels',label_array)
            title(strcat(['Data scenario ' num2str(data_scenario) ', ...
                MDM ' num2str(miss_mech) ', MSE of Y' num2str(i+1)]))
        end
    end
end
```

The following program box-plotted the ranks for the simulation study in Chapter 4.

```
l_string = {'INC','CC','N','N_{PMM}','N_{LRD}','N_{ERD}','t','t_{PMM}'...
```

```
   ,'t_{LRD}','t_{ERD}','t_{skew}'};
data = [10  11  2   7   5   1   3   8   6   9   4
9   11  1   7   6   3   2   8   5   10  4
10  11  2   8   6   1   4   7   5   9   3
10  11  3   7   5   1   2   8   6   9   4
10  11  1   8   5   3   4   9   6   7   2
10  11  5   9   2   4   6   8   1   7   3
10  11  3   8   6   1   4   9   5   7   2
10  11  5   8   2   3   6   9   1   7   4
10  11  2   9   6   4   5   8   7   1   3
8   11  1   10  6   4   2   9   7   5   3
10  11  1   8   7   2   4   9   6   3   5
10  11  3   8   4   2   5   9   7   6   1
10  11  7   9   3   6   1   8   4   5   2
10  11  1   9   5   2   4   8   6   7   3
10  11  3   9   4   7   1   8   5   6   2
10  11  4   9   6   3   2   8   5   7   1
10  11  3   9   7   1   4   8   5   2   6
10  11  2   8   3   7   4   9   5   1   6
10  11  5   9   6   1   2   8   7   3   4
10  11  5   8   4   3   7   9   2   1   6
10  11  2   9   7   1   4   8   6   5   3
10  11  1   9   5   4   3   8   6   7   2
10  11  3   9   6   1   2   8   7   5   4
10  11  4   8   6   2   3   9   5   7   1
10  11  2   9   5   1   4   8   7   3   6
6   11  10  8   1   9   4   7   2   5   3
10  11  4   9   6   1   3   8   7   2   5
8   11  10  7   1   9   3   6   4   5   2
9   11  5   8   6   1   2   10  7   3   4
8   11  4   10  5   3   2   9   6   7   1
10  11  2   8   4   1   5   9   6   7   3
10  11  4   9   5   3   1   8   6   7   2
10  11  2   8   5   3   1   9   6   4   7
8   11  4   9   5   1   3   10  6   7   2
10  11  4   9   6   1   2   8   7   5   3
10  11  2   9   5   4   3   8   6   7   1
10  11  2   8   5   1   3   7   6   9   4
7   11  1   9   5   2   4   8   6   10  3
10  11  1   7   5   2   3   8   6   9   4
9   11  4   8   5   3   1   7   6   10  2
10  11  1   8   5   2   4   9   6   7   3
9   11  4   10  6   1   2   8   5   7   3
10  11  2   8   6   3   4   9   5   7   1
10  11  4   8   6   2   3   9   5   7   1
10  11  4   8   5   1   3   9   6   7   2
10  11  4   8   5   2   1   9   6   7   3
10  11  3   9   6   1   2   8   5   7   4
10  11  2   8   5   3   1   9   6   7   4
1   11  8   4   2   3   9   7   6   5   10
1   9   8   5   3   6   11  2   4   7   10
1   11  10  2   3   6   9   7   4   5   8
1   11  10  6   4   5   9   3   2   7   8
4   11  6   8   5   3   1   7   9   10  2
2   11  9   8   1   3   7   6   4   10  5
9   11  5   7   4   3   2   8   6   10  1
9   11  8   7   2   5   3   6   1   10  4
```

```
1   11  5   8   6   2   4   9   7   10  3
2   10  9   7   6   1   4   8   5   11  3
1   11  5   6   8   3   2   7   9   10  4
9   10  8   4   5   1   3   6   7   11  2];


subplot(1,3,3)
[s_mean_data, I] = sort(mean(data));
s_data = data(:,I);
s_labels = l_string(I);
boxplot(s_data,'labels',s_labels,'plotstyle','compact')
title('Combined')

subplot(1,3,1)
data200 = data(1:2:59,:);
[s_mean_data, I] = sort(mean(data200));
s_data = data200(:,I);
s_labels = l_string(I);
boxplot(s_data,'labels',s_labels,'plotstyle','compact')
title('n = 200')

subplot(1,3,2)
data1000 = data(2:2:60,:);
[s_mean_data, I] = sort(mean(data1000));
s_data = data1000(:,I);
s_labels = l_string(I);
boxplot(s_data,'labels',s_labels,'plotstyle','compact')
title('n = 1000')

l_string = {'N','N_{PMM}','N_{LRD}','N_{ERD}','t','t_{PMM}','t_{LRD}',...
    't_{ERD}','t_{skew}'};
data =[1    8   6   3   2   7   5   9   4
8   2   6   7   4   1   9   3   5
2   3   9   7   5   6   8   1   4
8   1   3   6   7   2   4   9   5
8   3   6   5   2   7   1   9   4
1   8   7   2   6   9   5   4   3
8   6   4   2   3   7   5   9   1
6   2   3   4   7   5   9   1   8
4   8   5   3   6   7   1   9   2
2   5   4   6   9   1   8   3   7
1   7   6   5   4   8   3   9   2
5   1   4   7   6   2   3   9   8
3   1   7   5   6   2   8   9   4
2   7   6   4   1   5   8   9   3
1   3   8   4   5   2   7   9   6
3   9   5   4   6   8   7   1   2
8   5   3   7   6   2   1   4   9
7   8   5   2   1   6   4   9   3
7   4   5   8   6   3   2   9   1
6   4   5   1   2   8   3   9   7
4   8   5   1   3   7   2   9   6
7   5   6   8   2   4   3   9   1
8   1   7   5   6   3   2   9   4
6   3   1   5   7   4   2   9   8
8   1   2   6   5   7   4   9   3
8   7   5   6   4   3   1   9   2
```

```
8   4   7   2   1   5   6   9   3
8   6   3   4   2   5   7   9   1
8   4   6   1   3   7   5   9   2
3   4   8   5   6   1   2   9   7
4   3   5   2   8   7   6   9   1
6   3   7   5   4   1   8   9   2
5   8   7   6   1   9   4   3   2
7   8   4   2   3   6   1   9   5
2   5   7   1   6   3   8   9   4
2   7   5   6   1   8   3   9   4
8   7   2   6   4   3   1   9   5
7   3   6   5   1   8   4   9   2
1   8   6   5   7   4   2   9   3
8   6   2   7   1   4   3   9   5
5   6   2   4   3   7   1   9   8
6   5   7   4   1   3   8   9   2
9   1   7   3   4   2   6   5   8
8   1   7   5   4   2   3   9   6
1   3   8   6   4   2   7   9   5
9   1   2   7   5   6   4   3   8
1   6   5   4   3   7   8   9   2
1   7   8   4   3   5   6   9   2
6   3   5   2   7   4   1   9   8
6   1   8   3   4   2   7   9   5
7   2   5   4   6   1   3   9   8
7   1   5   4   6   2   3   9   8
6   8   2   3   1   7   5   9   4
8   2   5   4   7   1   3   9   6
8   2   6   1   5   3   7   9   4
8   1   2   7   6   4   5   9   3
1   8   4   7   6   5   2   9   3
8   3   1   5   6   2   4   9   7
8   4   7   5   1   3   6   9   2
1   6   2   5   3   8   4   9   7];

figure
subplot(1,3,3)
[s_mean_data, I] = sort(mean(data));
s_data = data(:,I);
s_labels = l_string(I);
boxplot(s_data,'labels',s_labels,'plotstyle','compact')
title('Combined')

subplot(1,3,1)
data200 = data(1:30,:);
[s_mean_data, I] = sort(mean(data200));
s_data = data200(:,I);
s_labels = l_string(I);
boxplot(s_data,'labels',s_labels,'plotstyle','compact')
title('n = 200')

subplot(1,3,2)
data1000 = data(31:60,:);
[s_mean_data, I] = sort(mean(data1000));
s_data = data1000(:,I);
s_labels = l_string(I);
boxplot(s_data,'labels',s_labels,'plotstyle','compact')
```

```
title('n = 1000')
```

The following program performed the second analysis study in Chapter 4.

```
% In this program I will generate 3 n-by-4 datasets.
% Case 1: y1¬N. y2|y1¬N. y3|y1,y2¬N. y4|y1,y2,y3¬N.
% MAR randomness generated using logit with logit_betas as parameters.
% y1 complete, y2M¬logit(y1), y3M¬logit(y1,y2), y4M¬logit(y1,y2,y3),
% but if any of the logit arguments are missing, they are ignored.

% This part of project 1 will do the imputation coverage graphs.
% One incomplete data set from MCAR and one from MCAR are imputed, and...
     100
% iterations are imputed. The KS test accept/reject fraction is ...
    calculated
% for each iteration, and a coverage graph for T2, Y3 and Y4 is ...
    created.
clear
clc
start = tic;
n = 1000; %dataset size
draws = 500;
mi_rounds = 15;
mi_iterations = 200;
logit_betas = [-0.3 -0.3 -0.3 -0.3];
mcar_prop = 0.2;
data_scenarios = 5;
mi_options = 9;
imp_cov_cdata = cell(1,2,data_scenarios);
mse_imp_cov_cdata = cell(1,2,data_scenarios);
for i = 1:2 % The MDMs
    for j = 1:4 % The different data scenarios
        imp_cov_cdata{1,i,j} = zeros(mi_options,99,3); %mi_opt x 1% ...
            percentile x cols
        mse_imp_cov_cdata{1,i,j} = zeros(mi_options,3); %mi_opt x cols
    end
end
original_data_cell = cell(data_scenarios,1); % 1 for each ...
    data_scenario
idata_cell = cell(1,2,data_scenarios);
cpart_idata_cell = cell(1,2,data_scenarios);
completed_data_cell = cell(9,2,data_scenarios);
parameter_text = {'n =', n;
    'draws =', draws;
    'mi_iterations =', mi_iterations;
    'logit_betas =', NaN;
    'mcar_prop =', mcar_prop;
    NaN, NaN;
    'MCAR missing y1', NaN; % line 7
    'MCAR missing y2', NaN;
    'MCAR missing y3', NaN;
    'MCAR missing y4', NaN;
    'MCAR cpart miss', NaN;
    'MAR missing y1', NaN;
    'MAR missing y2', NaN;
    'MAR missing y3', NaN;
    'MAR missing y4', NaN;
    'MAR cpart miss', NaN;
    NaN, NaN;
    'MSE IMP QQ Y2 Y3 Y4',NaN;
    'MDM MCAR',NaN;
    'N',NaN; %line 20
    'N_{pmm}',NaN;
    'N_{lrd}',NaN;
    'N_{erd}',NaN;
    't',NaN;
    't_{pmm}',NaN;
    't_{lrd}',NaN;
    't_{erd}',NaN;
    't_{skew}',NaN;
    NaN, NaN;
    'MDM MAR',NaN;
    'N',NaN; %line 31
    'N_{pmm}',NaN;
    'N_{lrd}',NaN;
    'N_{erd}',NaN;
    't',NaN;
    't_{pmm}',NaN;
    't_{lrd}',NaN;
    't_{erd}',NaN;
    't_{skew}',NaN};

xlswrite('Project_01_Part_2.xlsx',parameter_text,'Sheet1','A1');
xlswrite('Project_01_Part_2.xlsx',logit_betas,'Sheet1','B4');
xlswrite('Project_01_Part_2.xlsx',parameter_text,'Sheet2','A1');
xlswrite('Project_01_Part_2.xlsx',logit_betas,'Sheet2','B4');
xlswrite('Project_01_Part_2.xlsx',parameter_text,'Sheet3','A1');
xlswrite('Project_01_Part_2.xlsx',logit_betas,'Sheet3','B4');
xlswrite('Project_01_Part_2.xlsx',parameter_text,'Sheet4','A1');
xlswrite('Project_01_Part_2.xlsx',logit_betas,'Sheet4','B4');
xlswrite('Project_01_Part_2.xlsx',parameter_text,'Sheet5','A1');
xlswrite('Project_01_Part_2.xlsx',logit_betas,'Sheet5','B4');
for data_scenario = 1:data_scenarios
    scenario_timer = tic;
    switch data_scenario
        case 1 % Normality
            y1 = randn(n,1);
            y2 = 1 + y1 + randn(n,1);
            y3 = 1 + y1 + y2 + randn(n,1);
            y4 = 1 + y1 + y2 + y3 + randn(n,1);
            Y = [y1 y2 y3 y4]; clear y1 y2 y3 y4
            Original_Data = Y;
        case 2 %
            y1 = randn(n,1);
            y2 = 1 + y1 + trnd(6,n,1);
            y3 = 1 + y1 + y2 + trnd(6,n,1) - randn(n,1);
            y4 = 1 + y1 + y2 + y3 + trnd(6,n,1) - 2*randn(n,1);
            Y = [y1 y2 y3 y4]; clear y1 y2 y3 y4
            Original_Data = Y;
        case 3
            y1 = randn(n,1);
            y2 = 1 + y1 + trnd(3,n,1);
            y3 = 1 + y1 + y2 + trnd(3,n,1) - randn(n,1);
            y4 = 1 + y1 + y2 + y3 + trnd(3,n,1) - 2*randn(n,1);
            Y = [y1 y2 y3 y4]; clear y1 y2 y3 y4
            Original_Data = Y;
        case 4
            y1 = randn(n,1);
            u2 = rand(n,1);
            y2 = 1 + y1 + (exp(u2)-1).*exp(-0.125*u2.^2); % ...
               exponential type
            u3 = rand(n,1);
            y3 = 1 + y1 + y2 + (exp(0.75.*u3)-1)./0.75.*exp(0.125*u3....
               ^2); % skew type
            u4 = rand(n,1);
            y4 = 1 + y1 + y2 + y3 + (exp(u4)-1); % lognormal
            Y = [y1 y2 y3 y4]; clear y1 y2 y3 y4 u2 u3 u4
            Original_Data = Y;
        case 5
            y1 = randn(n,1);
            e2 = 1+exp(1+randn(n,1));
            e2 = e2-mean(e2);
            e2 = e2/std(e2) * sqrt(var(y1)*3);
            y2 = 1 + y1 + e2;
            e3 = trnd(3,n,1);
            e3 = e3 - mean(e3);
            e3 = e3/std(e3) * sqrt(2*(var(y1)+var(y2)));
            y3 = 1 + y1 + y2 + e3;
            e4 = trnd(3,n,1)-2*randn(n,1);
            e4 = e4 -mean(e4);
            e4 = e4/std(e4) * (var(y1)+var(y2)+var(y3));
            y4 = 1 + y1 + y2 + y3 +e4;
            Y = [y1 y2 y3 y4]; clear y1 y2 y3 y4 e2 e3 e4
            Original_Data = Y;
    end
    original_data_cell{data_scenario,1} = Original_Data;
    [n p] = size(Y);
    missingness = zeros(10,1);
    % MCAR y1
    % MCAR y2
    % MCAR y3
    % MCAR y4
    % MCAR cpart_idata
    % MAR y1
    % MAR y2
    % MAR y3
    % MAR y4
    % MAR cpart_idata
    for miss_mech = 1:2;
        % C is the binary matrix with 1 indicating observed and 0
        % indicating missing (completeness matrix).
        switch miss_mech
            case 1 % MCAR missingness
                idata = Y;
                P_M = [zeros(n,1) rand(n,3)];
                C = ones(n,4);
                for col = 2:4  %variables with missingness
                    %U = rand(n,1);
                    C((P_M(:,col)<mcar_prop),col) = zeros(sum(P_M(:,...
                       col)<mcar_prop),1);
                end
                % Calculate missingness:
                idata(C==0)=NaN; % incomplete data
                cpart_idata = idata(isfinite(sum(idata,2)),:); % Case ...
                   deleted data
                missingness(1:4,1) = (sum(isnan(idata))./n)';
                missingness(5,1) = 1- length(cpart_idata(:,1))/n;
            case 2 % MAR missingness
                idata = Y;
                P_M = zeros(n,4); %probability missing
                C = ones(n,4);
                for col = 2:4 %variables with missingness
                    a = logit_betas(1) * ones(n,1);
                    for i = 1:(col-1)
                        a = nansum([a logit_betas(i+1)*(idata(:,i)-...
                           nanmean(idata(:,i)))./nanstd(idata(:,i))...
                           ],2);
                    end
                    P_M(:,col) = 0.4./(1+exp(-a));
                    U = rand(n,1);
                    C((U<P_M(:,col)),col) = zeros(sum(U<P_M(:,col)),1)...
                       ;
                    idata(C==0)=NaN; % incomplete data
                end
                % Calculate missingness:
                % idata(C==0)=NaN; % incomplete data
                cpart_idata = idata(isfinite(sum(idata,2)),:); % Case ...
                   deleted data
                missingness(6:9,1) = (sum(isnan(idata))./n)';
                missingness(10,1) = 1- length(cpart_idata(:,1))/n;
        end
        M = isnan(idata);
        for mi_option = 1:9
            mi_start = tic; % Timer for the SRMI procedure
            disp(strcat(['Data_Scenario ' num2str(data_scenario) ', ...
               MI_option ' num2str(mi_option), ', Miss_Mech ' num2str...
               (miss_mech)]))
            switch mi_option
                case 1
                    [cdata,¬,¬]=SRMI(idata,0,0,[1 1 1 1],20,...
                       mi_iterations);
                    e_col = 'B';
                case 2
                    [cdata,¬,¬]=SRMI(idata,0,0,[1 1 1 1],20,...
                       mi_iterations,0,0,[1 1 1 1],0.1);
                    e_col = 'C';
                case 3
                    [cdata,¬,¬]=SRMI(idata,0,0,[1 1 1 1],20,...
                       mi_iterations,0,0,[2 2 2 2],0.1);
                    e_col = 'D';
                case 4
                    [cdata,¬,¬]=SRMI(idata,0,0,[1 1 1 1],20,...
                       mi_iterations,0,0,[3 3 3 3],0.1);
                    e_col = 'E';
                case 5
                    [cdata,¬,¬]=SRMI(idata,0,0,[1.5 1.5 1.5 1.5],20,...
                       mi_iterations);
                    e_col = 'F';
                case 6
                    [cdata,¬,¬]=SRMI(idata,0,0,[1.5 1.5 1.5 1.5],20,...
                       mi_iterations,0,0,[1 1 1 1],0.1);
                    e_col = 'G';
                case 7
                    [cdata,¬,¬]=SRMI(idata,0,0,[1.5 1.5 1.5 1.5],20,...
                       mi_iterations,0,0,[2 2 2 2],0.1);
                    e_col = 'H';
                case 8
                    [cdata,¬,¬]=SRMI(idata,0,0,[1.5 1.5 1.5 1.5],20,...
                       mi_iterations,0,0,[3 3 3 3],0.1);
                    e_col = 'I';
                case 9
                    [cdata,¬,¬]=SRMI(idata,0,0,[1.6 1.6 1.6 1.6],20,...
                       mi_iterations);
                    e_col = 'J';
            end
            completed_data_cell{mi_option,miss_mech,data_scenario} = ...
               cdata;
            toc(mi_start)

            %counting coverage
            for col = 2:p %we know var 1 is full
                X = repmat(Original_Data(M(:,col)==1,col),1,99);
                tmp = cdata(M(:,col)==1,col,:);
                tmp = reshape(tmp,length(tmp(:,1,1)),mi_iterations);
                P = prctile(tmp,0.5:0.5:99.5,2);
                P1 = P(:,1:99);
                P2 = P(:,199:-1:101);
                col_coverage = mean((P1<X) & (P2>X));
                col_coverage = col_coverage(:,end:-1:1);
                %Save the coverages, and MSEs of coverages:
                imp_cov_cdata{1,miss_mech,data_scenario}(mi_option,:,(...
                   col-1)) = col_coverage;
                mse_imp_cov_cdata{1,miss_mech,data_scenario}(mi_option...
                   ,(col-1)) = mean(((1:99)-col_coverage*100).^2);
            end
        end
        disp(strcat(['Finished with MI option ' num2str(mi_option) ' ...
           on data scenario ' num2str(data_scenario)]))
    end
    disp(strcat(['Finished with page ' num2str(data_scenario)]))
    toc(scenario_timer)
end

%now that I have all the data arrays, I must fins the mse's of the qq-...
   plots
%for the columns within each cell (against Original_Data columns)
%Original_Data n x p
%idata_cell 1,2,4 each with n x p
%cpart_idata_cell 1,2,4 each with n x p
%cdata_cell 9,2,4 each with n x p

% MSE imputations Data writer
for data_scenario = 1:data_scenarios
    for miss_mech = 1:2
            xlswrite('Project_01_Part_2.xlsx',mse_imp_cov_cdata{1,...
                   miss_mech,data_scenario}(:,:),strcat(['Sheet' num2str(...
                   data_scenario)]),strcat(['B' num2str(20+(miss_mech-1)...
                   *11)]));
    end
end
toc(start)
xlswrite('Project_01_Part_2.xlsx',missingness,strcat(['Sheet' num2str(...
    data_scenario)]),'B7');
save D:\Workspace_P1b_n1000

% Plotting coverage for imputations
for data_scenario = 1:data_scenarios
    for miss_mech = 1:2
        figure
        title(strcat(['Imputation coverage intervals for data scenario...
            ' num2str(data_scenario) ' and MDM ' num2str(miss_mech)])...
           )
        for i = 1:3 % the number of columns with missingness
            subplot(3,1,i)
            hold on
            plot([0.01 0.99] ,[0.01 0.99],'k')
            plot(0.01:0.01:0.99, imp_cov_cdata{1,miss_mech,...
                data_scenario}(1,:,i)','b')
            plot(0.01:0.01:0.99, imp_cov_cdata{1,miss_mech,...
                data_scenario}(2,:,i)','.b')
            plot(0.01:0.01:0.99, imp_cov_cdata{1,miss_mech,...
                data_scenario}(3,:,i)','--b')
            plot(0.01:0.01:0.99, imp_cov_cdata{1,miss_mech,...
                data_scenario}(4,:,i)','-.b')
            plot(0.01:0.01:0.99, imp_cov_cdata{1,miss_mech,...
                data_scenario}(5,:,i)','r')
            plot(0.01:0.01:0.99, imp_cov_cdata{1,miss_mech,...
                data_scenario}(6,:,i)','.r')
            plot(0.01:0.01:0.99, imp_cov_cdata{1,miss_mech,...
                data_scenario}(7,:,i)','--r')
            plot(0.01:0.01:0.99, imp_cov_cdata{1,miss_mech,...
                data_scenario}(8,:,i)','-.r')
            plot(0.01:0.01:0.99, imp_cov_cdata{1,miss_mech,...
                data_scenario}(9,:,i)','g')
            title(strcat(['Data page ' num2str(data_scenario) ', MDM '...
                 num2str(miss_mech) ', coverage of Y' num2str(i+1)]))
            legend('45 deg','N','N_{pmm}','N_{lrd}','N_{erd}','t','t_{...
                pmm}','t_{lrd}','t_{erd}','t_{skew}','Location','...
                NorthEastOutside')
            hold off
        end
    end
end
```

B.3 Programs for Chapter 5

The following program performs the basic *cppp* example from Chapter 5.

```
sims_layer1 = 10000;
sims_layer2 = 10000;
n = 16;
mu_0 = 5;
sig2_0 = 9;
sig2 = 4;
cppp0_simulations = zeros(sims_layer1,1);
cppp1_simulations = zeros(sims_layer1,1);

for cppp_sim = 1:sims_layer1
    % one data set
    mu_draw = randn*sig2_0^0.5+mu_0;
    y0 =  mu_draw + sig2^0.5*randn(n,1);
    ybar = mean(y0);
    lRSS = 1/sig2 * sum((y0-ybar).^2);
    Pn = (n*sig2_0)/(sig2+n*sig2_0);
    cppp = 0;
    z1 = randn(sims_layer2,1);
    z2 = randn(sims_layer2,1);
    w = chi2rnd(n-1,sims_layer2,1);
    rhs = lRSS + Pn*(z1-(n^0.5)*(1-Pn)/(Pn^0.5*sig2^0.5)*(ybar-mu_0))....
        ^2 - Pn*(z1 - (n^0.5)*(1-Pn)*sig2_0^0.5/(Pn*sig2^0.5)*z2).^2;
    cppp0_simulations(cppp_sim) = mean(w>rhs);
    if mod(cppp_sim,100) == 0
        disp(cppp_sim)
    end
end
hist(cppp0_simulations)

for cppp_sim = 1:sims_layer1
    % one data set
    mu_draw = randn*sig2_0^0.5+mu_0;
    [truncNs] = trunc_N(0,Inf,0,1,n);
    y1 =  mu_draw + truncNs + sig2^0.5*trnd(5,n,1);
    ybar = mean(y1);
    lRSS = 1/sig2 * sum((y1-ybar).^2);
    Pn = (n*sig2_0)/(sig2+n*sig2_0);
    cppp = 0;
    z1 = randn(sims_layer2,1);
    z2 = randn(sims_layer2,1);
    w = chi2rnd(n-1,sims_layer2,1);
    rhs = lRSS + Pn*(z1-(n^0.5)*(1-Pn)/(Pn^0.5*sig2^0.5)*(ybar-mu_0))....
        ^2 - Pn*(z1 - (n^0.5)*(1-Pn)*sig2_0^0.5/(Pn*sig2^0.5)*z2).^2;
    cppp1_simulations(cppp_sim) = mean(w>rhs);
    if mod(cppp_sim,100) == 0
        disp(cppp_sim)
    end
end
figure(1)
set(1,'units','normalized','position',[0.05 0.05 0.25 0.5]);
subplot(2,1,1)
hist(cppp0_simulations)
title('Histogram of cppp values under the null model')
```

```
subplot(2,1,2)
hist(cppp1_simulations)
title('Histogram of cppp values under the a skew t alternative model')
export_fig cppp_example.pdf -nocrop
```

The following program performs the MCMC approximation of the *cppp* from Chapter 5.

```
clear
clc
n = 200; %dataset size
pppsims = 200;
sims = 500;
ppp0 = zeros(1,sims);
cppp0 = zeros(1,sims);
pppt0 = zeros(pppsims,sims);
ppp1 = zeros(1,sims);
cppp1 = zeros(1,sims);
pppt1 = zeros(pppsims,sims);

model = 1;

for sim = 1:sims
    tic
    sim
    x0 = rand(n,1);
    y0 = x0 + 1 + randn(n,1); %null model
    [cppp0(sim), ppp0(sim), pppt0(:,sim)] = cppp_val_complete(y0, x0, ...
       model, pppsims, 2,0,0);
    xlswrite('CabrasOutput_N_full.xlsx',[ppp0(sim) cppp0(sim)],'Sheet1...
       ', strcat('A', num2str(sim)));
    toc
end

for sim = 1:sims
    tic
    sim
    x1 = rand(n,1);
    y1 = x1 + 1 + 0.5*trunc_N(0,Inf,0,1,n) + trnd(3,n,1);
    [cppp1(sim), ppp1(sim), pppt1(:,sim)] = cppp_val_complete(y1, x1, ...
       model, pppsims, 2,0,0);
    xlswrite('CabrasOutput_N_full.xlsx',[ppp1(sim) cppp1(sim)],'Sheet2...
       ', strcat('A', num2str(sim)));
    toc
end
toc
figure(1)
set(1,'units','normalized','position',[0.05 0.05 0.5 0.8]);
subplot(2,2,1)
hold on
plot(sort(ppp0),'--')
plot(sort(cppp0),'r-.')
plot([0 sims],[0 1],'k')
hold off
legend('ppp','cppp','Location','NorthWest')
title('Distribution of ppp and cppp under the null')
subplot(2,2,3)
hist(cppp0)
```

```
title('Density of the cppp under the null')
subplot(2,2,2)
hold on
plot(sort(ppp1),'--')
plot(sort(cppp1),'r-.')
plot([0 sims],[0 1],'k')
hold off
legend('ppp','cppp','Location','NorthWest')
title('Distribution of ppp and cppp under the alternative')
subplot(2,2,4)
hist(cppp1)
title('Density of the cppp under the alternative')
export_fig cppp_cabras_example2.pdf -nocrop
```

The following program calculates the discrepancy measure used in the previous two programs in Chapter 5.

```
clear
clc
n = 200; %dataset size
pppsims = 200;
sims = 500;
ppp0 = zeros(1,sims);
cppp0 = zeros(1,sims);
pppt0 = zeros(pppsims,sims);
ppp1 = zeros(1,sims);
cppp1 = zeros(1,sims);
pppt1 = zeros(pppsims,sims);

model = 1;

for sim = 1:sims
    tic
    sim
    x0 = rand(n,1);
    y0 = x0 + 1 + randn(n,1); %null model
    [cppp0(sim), ppp0(sim), pppt0(:,sim)] = cppp_val_complete(y0, x0, ...
        model, pppsims, 2,0,0);
    xlswrite('CabrasOutput_N_full.xlsx',[ppp0(sim) cppp0(sim)],'Sheet1...
        ', strcat('A', num2str(sim)));
    toc
end

for sim = 1:sims
    tic
    sim
    x1 = rand(n,1);
    y1 = x1 + 1 + 0.5*trunc_N(0,Inf,0,1,n) + trnd(3,n,1);
    [cppp1(sim), ppp1(sim), pppt1(:,sim)] = cppp_val_complete(y1, x1, ...
        model, pppsims, 2,0,0);
    xlswrite('CabrasOutput_N_full.xlsx',[ppp1(sim) cppp1(sim)],'Sheet2...
        ', strcat('A', num2str(sim)));
    toc
end
toc
figure(1)
set(1,'units','normalized','position',[0.05 0.05 0.5 0.8]);
```

```
subplot(2,2,1)
hold on
plot(sort(ppp0),'--')
plot(sort(cppp0),'r-.')
plot([0 sims],[0 1],'k')
hold off
legend('ppp','cppp','Location','NorthWest')
title('Distribution of ppp and cppp under the null')
subplot(2,2,3)
hist(cppp0)
title('Density of the cppp under the null')
subplot(2,2,2)
hold on
plot(sort(ppp1),'--')
plot(sort(cppp1),'r-.')
plot([0 sims],[0 1],'k')
hold off
legend('ppp','cppp','Location','NorthWest')
title('Distribution of ppp and cppp under the alternative')
subplot(2,2,4)
hist(cppp1)
title('Density of the cppp under the alternative')
export_fig cppp_cabras_example2.pdf -nocrop
```

B.4 Programs for Chapter 6

The following program runs the first (single simulation) analysis for Chapter 6.

```
function [errors,latent_data]=Project_02_Part1(n,categories,draws,...
    burn_in)

clc

%categories = 2;
%n= 1000;
%burn_in = 500;
%draws = 400;
latent_data = zeros(n,3,4);
data_scenarios = 4;
errors = zeros(data_scenarios,2);

for data_model = 1:data_scenarios

    % Setting the real data
    x = 0.5+randn(n,1);
    if categories == 2
        cutoff = 2;
        [latent,skewness,y,X] = new_sample(data_model,n,categories,...
            cutoff);
    elseif categories == 3
        cutoff = [-2 2];
        [latent,skewness,y,X] = new_sample(data_model,n,categories,...
            cutoff);
    end
    latent_data(:,:,data_model) = [latent y x];
    figure(1)
    set(1,'units','normalized','position',[0.05 0.05 0.9 0.5]);
    grapher_1(data_model,categories,latent,skewness) %grapher for ...
        latent and skewness

    tic
    % Fitting the probit
    [beta,Z,Gamma]= gibbs_oprobit_fitter(y,X,draws,burn_in);
    G_norm = median(Gamma,2);
    b_norm = median(beta,2);
    z_norm = median(Z,2);

    figure(2)
    set(2,'units','normalized','position',[0.05 0.05 0.9 0.5]);
    grapher_2(data_model, categories, beta, Gamma) %grapher for probit...
         draws

    disp('Probit finished, starting skew robit')

    tic
    % Fitting skew robit
    [beta,tau,V,¬,∆,¬,W,Gamma] = gibbs_osrobit_fitter3(y,X,draws,...
        burn_in);
    G_t = median(Gamma,2); % mean or median. Must choose.
    b_t = median(beta,2)';
    w = median(W,2);
    d = median(∆);
    v = median(V);
    t = median(tau);


    figure(3)
    set(3,'units','normalized','position',[0.05 0.05 0.9 0.5]);
    grapher_3(data_model,categories,beta,V,∆,tau,Gamma) %grapher for ...
       ostrobit draws
    figure(4)
    set(4,'units','normalized','position',[0.05 0.05 0.4 0.8]);
    grapher_4(data_model,categories,y,z_norm,w)  %grapher for forced-...
       to-fit latent data

    %Predicted augmented data for errors
    if categories == 2
        [latent,¬,y,X] = new_sample(data_model,n,categories,cutoff);
    elseif categories == 3
        [latent,¬,y,X] = new_sample(data_model,n,categories,cutoff);
    end
    z_rep = X*b_norm + randn(n,1);
    [z_draws] = trunc_N(0,Inf,0,1,n);
    w_rep = X*b_t' + d*z_draws + t^(-0.5)*trnd(v);
    y_rep_N = ones(n,1);
    y_rep_t = ones(n,1);
    for i = 1:(categories-1)
        y_rep_N(z_rep≥cutoff(i))=(i+1);
        y_rep_t(w_rep≥cutoff(i))=(i+1);
    end
    normal_shifts = y-y_rep_N;
    t_shifts = y-y_rep_t;
    normal_error = mean(abs(y-y_rep_N));
    t_error = mean(abs(y-y_rep_t));
    errors(data_model,:) = [normal_error t_error];
    if categories == 3
        figure(6)
        set(6,'units','normalized','position',[0.1 0.1 0.4 0.6])
        grapher_6(data_model,normal_shifts,t_shifts)
    end

    figure(5)
    set(5,'units','normalized','position',[0.05 0.05 0.4 0.8]);
    grapher_5(data_model,categories,y,z_rep,w_rep)  %grapher for ...
       forced-to-fit latent data


    toc
end
end


function [latent,skewness,y,X] = new_sample(data_model,n,categories,...
   cutoff)
x = 0.5+randn(n,1);
B_true = [-1 4]';
if data_model == 1;
    %Normal
    latent = B_true(1)+B_true(2)*x+randn(n,1);
elseif data_model == 2;
    %skew t data
    [z_draws] = trunc_N(0,Inf,0,1,n);
    latent = B_true(1)+B_true(2)*x-2*z_draws+0.5*trnd(5,n,1);
elseif data_model == 3;
    %exponential data
    latent = B_true(1)+B_true(2)*x-log(1-rand(n,1));
elseif data_model == 4;
    %uniform data
    latent = B_true(1)+B_true(2)*x+randi(6,n,1)-6;
end
skewness = latent - B_true(1)-B_true(2)*x;

X = [ones(n,1) x];
y = ones(n,1);

for i = 1:(categories-1)
    y(latent≥cutoff(i))=(i+1);
end
end
function grapher_1(data_model,categories,latent,skewness)%grapher for ...
   latent and skewness

figure(1)
if data_model == 1;
    %Normal
    figure(1)
    subplot(2,4,1)
    hist(latent,15)
    h = findobj(gca,'Type','patch');
    set(h,'FaceColor',[0.15 0.15 0.25],'EdgeColor','k')
    title('True Normal latent data')
    subplot(2,4,5)
    hist(skewness,15)
    h = findobj(gca,'Type','patch');
    set(h,'FaceColor',[0.4 0.4 0.7],'EdgeColor','k')
    title('Error added to X\beta for Normal latent')
elseif data_model == 2
    %skew t data
    figure(1)
    subplot(2,4,2)
    hist(latent,15)
    h = findobj(gca,'Type','patch');
    set(h,'FaceColor',[0.15 0.15 0.25],'EdgeColor','k')
    title('True skew t latent data')
    subplot(2,4,6)
    hist(skewness,15)
    h = findobj(gca,'Type','patch');
    set(h,'FaceColor',[0.4 0.4 0.7],'EdgeColor','k')
    title('Error added to X\beta for skew t latent')
elseif data_model == 3
    %exponential data
    figure(1)
    subplot(2,4,3)
    hist(latent,15)
    h = findobj(gca,'Type','patch');
    set(h,'FaceColor',[0.15 0.15 0.25],'EdgeColor','k')
    title('True Exponential latent data')
    subplot(2,4,7)
    hist(skewness,15)
    h = findobj(gca,'Type','patch');
    set(h,'FaceColor',[0.4 0.4 0.7],'EdgeColor','k')
    title('Error added to X\beta for Exponential latent')
elseif data_model == 4
    %uniform data
    figure(1)
    subplot(2,4,4)
    hist(latent,15)
    h = findobj(gca,'Type','patch');
    set(h,'FaceColor',[0.15 0.15 0.25],'EdgeColor','k')
    title('True Uniform latent data')
    subplot(2,4,8)
    hist(skewness,6)
    h = findobj(gca,'Type','patch');
    set(h,'FaceColor',[0.4 0.4 0.7],'EdgeColor','k')
    title('Error added to X\beta for Uniform latent')
    [¬,h] = suplabel('Latent data scenarios' ,'t');
    set(h,'FontSize',14);

    if categories == 2
        export_fig fig_data2.pdf -nocrop
    else
        export_fig fig_data3.pdf -nocrop
    end
end
end
function grapher_2(data_model,categories,beta, Gamma)%grapher for ...
      oprobit draws
figure(2)
subplot(1,4,data_model)
hold on
plot(beta(1,:)','r')
plot(beta(2,:)','b')
if categories == 3
    plot(Gamma(3,:)','k')
end
grid on
hold off
if data_model==1
    title('Unknown latent data is Normal')
    if categories == 2
        legend('B_0','B_1','location','East')
    else
        legend('B_0','B_1','Gamma_2','location','East')
    end
elseif data_model==2
    title('Unknown latent data is skew t')
    if categories == 2
        legend('B_0','B_1','location','East')
    else
        legend('B_0','B_1','Gamma_2','location','East')
    end
elseif data_model==3
```

```
    title('Unknown latent data is Exponential')
    if categories == 2
        legend('B_0','B_1','location','East')
    else
        legend('B_0','B_1','Gamma_2','location','East')
    end
elseif data_model==4
    title('Unknown latent data is Uniform')
    if categories == 2
        legend('B_0','B_1','location','East')
        [¬,h] = suplabel('probit2 parameter draws after burn-in' ,'t')...
            ;
        set(h,'FontSize',14);
        export_fig fig_probit2_draws.pdf -nocrop
    else
        legend('B_0','B_1','Gamma_2','location','East')
        [¬,h] = suplabel('probit3 parameter draws after burn-in' ,'t')...
            ;
        set(h,'FontSize',14);
        export_fig fig_probit3_draws.pdf -nocrop
    end
end
end
function grapher_3(data_model,categories,beta,V,∆,tau,Gamma) %grapher ...
    for ostrobit draws

figure(3)
subplot(1,4,data_model)
hold on
plot(beta(1,:),'r')
plot(beta(2,:),'g')
if categories == 3
    plot(Gamma(3,:)','k')
end
plot(V,'b')
plot(∆,'c')
plot(tau,'m')

grid on
hold off
if data_model==1
    title('Unknown latent data is Normal')
    if categories == 2
        legend('B_0','B_1','v','d','tau','location','North')
    else
        legend('B_0','B_1','Gamma_2','v','d','tau','location','North')
    end
elseif data_model == 2
    title('Unknown latent data is skew t')
    if categories == 2
        legend('B_0','B_1','v','d','tau','location','North')
    else
        legend('B_0','B_1','Gamma_2','v','d','tau','location','North')
    end
elseif data_model == 3
    title('Unknown latent data is Exponential')
    if categories == 2
```

```
        legend('B_0','B_1','v','d','tau','location','North')
    else
        legend('B_0','B_1','Gamma_2','v','d','tau','location','North')
    end
elseif data_model == 4
    title('Unknown latent data is Uniform')
    if categories == 2
        legend('B_0','B_1','v','d','tau','location','North')
        [¬,h] = suplabel('strobit2 parameter draws after burn-in' ,'t'...
            );
        set(h,'FontSize',14);
        export_fig fig_strobit2_draws.pdf -nocrop
    else
        legend('B_0','B_1','Gamma_2','v','d','tau','location','North')
        [¬,h] = suplabel('strobit3 parameter draws after burn-in' ,'t'...
            );
        set(h,'FontSize',14);
        export_fig fig_strobit3_draws.pdf -nocrop
    end
end

end
function grapher_4(data_model,categories,y,z_norm,w) %grapher for ...
    forced-to-fit latent data

%Forced-to-fit augmented data
a = floor(min(z_norm)); b = ceil(max(z_norm));
z_norm1 = z_norm(y==1);
z_norm2 = z_norm(y==2);
if categories == 3
    z_norm3 = z_norm(y==3);
end
if ((b-a) > 29) && (mod(b-a,2) == 0)
    edges1 = a:2:b;
    step1 = 2;
elseif (b-a) > 29
    edges1 = (a-1):2:b;
    step1 = 2;
else
    edges1 = a:b;
    step1 = 1;
end
n1 = histc(z_norm1,edges1);
n2 = histc(z_norm2,edges1);
if categories == 3
    n3 = histc(z_norm3,edges1);
end
xlabels1 = num2cell(edges1);
if length(edges1)>15
    if mod(a,2) == 0;
        xlabels1(2:2:end) = {[]};
    else
        xlabels1(1:2:end) ={[]};
    end
end
a = floor(min(w)); b = ceil(max(w));
w1 = w(y==1)';
```

```
w2 = w(y==2)';
if categories == 3
    w3 = w(y==3)';
end
if ((b-a) > 29) && (mod(b-a,2) == 0)
    edges2 = a:2:b;
    step2 = 2;
elseif (b-a) > 29
    edges2 = (a-1):2:b;
    step2 = 2;
else
    edges2 = a:b;
    step2 = 1;
end

n4 = histc(w1,edges2);
n5 = histc(w2,edges2);
if categories == 3
    n6 = histc(w3,edges2);
end
xlabels2 = num2cell(edges2);
if length(edges2)>15
    if mod(a,2) == 0;
        xlabels2(2:2:end) = {[]};
    else
        xlabels2(1:2:end) ={[]};
    end
end

figure(4)
subplot(4,2,(data_model-1)*2+1)
if categories == 2
    h = bar(edges1,[n1 n2],1,'stacked');
    set(gca,'XTick',edges1-step1/2,'XTickLabel',xlabels1,'FontSize',9)...
        ;
    set(h(1),'FaceColor',[0.4 0.4 0.7],'EdgeColor','k');
    set(h(2),'FaceColor',[0.15 0.15 0.25],'EdgeColor','k')
else
    h = bar(edges1,[n1 n2 n3],1,'stacked');
    set(h(1),'FaceColor',[0.4 0.4 0.7],'EdgeColor','k');
    set(h(2),'FaceColor',[0.15 0.15 0.25],'EdgeColor','k')
    set(h(3),'FaceColor',[0.4 0.4 0.7],'EdgeColor','k');
end
if data_model==1
    title('PROBIT')
    xlabel('Unknown latent data is Normal')
elseif data_model==2
    xlabel('Unknown latent data is skew t')
elseif data_model==3
    xlabel('Unknown latent data is Exponential')
elseif data_model==4
    xlabel('Unknown latent data is Uniform')
end
subplot(4,2,(data_model-1)*2+2)
if categories == 2
    h = bar(edges2,[n4' n5'],1,'stacked');
    set(gca,'XTick',edges2-step2/2,'XTickLabel',xlabels2,'FontSize',9)...
```

```
        ;
    set(h(1),'FaceColor',[0.4 0.4 0.7],'EdgeColor','k');
    set(h(2),'FaceColor',[0.15 0.15 0.25],'EdgeColor','k')
else
    h = bar(edges2,[n4' n5' n6'],1,'stacked');
    set(gca,'XTick',edges2-step2/2,'XTickLabel',xlabels2,'FontSize',9)...
        ;
    set(h(1),'FaceColor',[0.4 0.4 0.7],'EdgeColor','k');
    set(h(2),'FaceColor',[0.15 0.15 0.25],'EdgeColor','k')
    set(h(3),'FaceColor',[0.4 0.4 0.7],'EdgeColor','k');
end

if data_model==1
    title('STROBIT')
    xlabel('Unknown latent data is Normal')
elseif data_model==2
    xlabel('Unknown latent data is skew t')
elseif data_model==3
    xlabel('Unknown latent data is Exponential')
elseif data_model==4
    xlabel('Unknown latent data is Uniform')
    [¬,h] = suplabel('Fitted latent data by observed category' ,'t');
    set(h,'FontSize',14);
    if categories == 2
        export_fig fig_latent2.pdf -nocrop
    else
        export_fig fig_latent3.pdf -nocrop
    end
end
end
function grapher_5(data_model,categories,y,z_rep,w_rep)%grapher for ...
    predicted latent data

a = floor(min(z_rep)); b = ceil(max(z_rep));
z_rep1 = z_rep(y==1);
z_rep2 = z_rep(y==2);
if categories == 3
    z_rep3 = z_rep(y==3);
end
if ((b-a) > 29) && (mod(b-a,2) == 0)
    edges1 = a:2:b;
    step1 = 2;
elseif (b-a) > 29
    edges1 = (a-1):2:b;
    step1 = 2;
else
    edges1 = a:b;
    step1 = 1;
end

n1 = histc(z_rep1,edges1);
n2 = histc(z_rep2,edges1);
if categories == 3
    n3 = histc(z_rep3,edges1);
end
xlabels1 = num2cell(edges1);
if length(edges1)>15
```

```
    if mod(a,2) == 0;
        xlabels1(2:2:end) = {[]};
    else
        xlabels1(1:2:end) ={[]};
    end
end
a = floor(min(w_rep)); b = ceil(max(w_rep));
w_rep1 = w_rep(y==1);
w_rep2 = w_rep(y==2);
if categories == 3
    w_rep3 = w_rep(y==3);
end
if ((b-a) > 29) && (mod(b-a,2) == 0)
    edges2 = a:2:b;
    step2 = 2;
elseif (b-a) > 29
    edges2 = (a-1):2:b;
    step2 = 2;
else
    edges2 = a:b;
    step2 = 1;
end
n4 = histc(w_rep1,edges2);
n5 = histc(w_rep2,edges2);
if categories == 3
    n6 = histc(w_rep3,edges2);
end
xlabels2 = num2cell(edges2);
if length(edges2)>15
    if mod(a,2) == 0;
        xlabels2(2:2:end) = {[]};
    else
        xlabels2(1:2:end) ={[]};
    end
end
figure(5)
subplot(4,2,(data_model-1)*2+1)
if categories == 2
    h = bar(edges1,[n1 n2],1,'stacked');
    set(gca,'XTick',edges1-step1/2,'XTickLabel',xlabels1,'FontSize',9)...
        ;
    set(h(1),'FaceColor',[0.4 0.4 0.7],'EdgeColor','k');
    set(h(2),'FaceColor',[0.15 0.15 0.25],'EdgeColor','k')
else
    h = bar(edges1,[n1 n2 n3],1,'stacked');
    set(h(1),'FaceColor',[0.4 0.4 0.7],'EdgeColor','k');
    set(h(2),'FaceColor',[0.15 0.15 0.25],'EdgeColor','k')
    set(h(3),'FaceColor',[0.4 0.4 0.7],'EdgeColor','k');
end
if data_model==1
    title('PROBIT')
    xlabel('Unknown latent data is Normal')
elseif data_model==2
    xlabel('Unknown latent data is skew t')
elseif data_model==3
    xlabel('Unknown latent data is Exponential')
elseif data_model==4
```

```
    xlabel('Unknown latent data is Uniform')
end
subplot(4,2,(data_model-1)*2+2)
if categories == 2
    h = bar(edges2,[n4 n5],1,'stacked');
    set(gca,'XTick',edges2-step2/2,'XTickLabel',xlabels2,'FontSize',9)...
       ;
    set(h(1),'FaceColor',[0.4 0.4 0.7],'EdgeColor','k');
    set(h(2),'FaceColor',[0.15 0.15 0.25],'EdgeColor','k')
else
    h = bar(edges2,[n4 n5 n6],1,'stacked');
    set(gca,'XTick',edges2-step2/2,'XTickLabel',xlabels2,'FontSize',9)...
       ;
    set(h(1),'FaceColor',[0.4 0.4 0.7],'EdgeColor','k');
    set(h(2),'FaceColor',[0.15 0.15 0.25],'EdgeColor','k')
    set(h(3),'FaceColor',[0.4 0.4 0.7],'EdgeColor','k');
end

if data_model==1
    title('STROBIT')
    xlabel('Unknown latent data is Normal')
elseif data_model==2
    xlabel('Unknown latent data is skew t')
elseif data_model==3
    xlabel('Unknown latent data is Exponential')
elseif data_model==4
    xlabel('Unknown latent data is Uniform')
    [¬,h] = suplabel('Predicted latent data by observed category' ,'t'...
       );
    set(h,'FontSize',14);
    if categories == 2
        export_fig fig_pred_latent2.pdf -nocrop
    else
        export_fig fig_pred_latent3.pdf -nocrop
    end
end

end
function grapher_6(data_model,normal_shifts,t_shifts)
edges = -2:2;
n1 = histc(normal_shifts,edges);
n2 = histc(t_shifts,edges);
figure(6)
subplot(2,2,data_model)
h = bar(edges,[n1 n2],1,'hist');
set(gca,'XTick',edges,'XTickLabel',edges,'FontSize',9);
set(h(1),'facecolor',[0.4 0.4 0.7],'edgecolor','k');
set(h(2),'FaceColor',[0.15 0.15 0.25],'EdgeColor','k')
legend('probit','srobit')
if data_model==1
    xlabel('Unknown latent data is Normal')
elseif data_model==2
    xlabel('Unknown latent data is skew t')
elseif data_model==3
    xlabel('Unknown latent data is Exponential')
elseif data_model==4
    xlabel('Unknown latent data is Uniform')
    [¬,h] = suplabel('Category shifts (classifications) by model' ,'t'...
        );
    set(h,'FontSize',14);
    export_fig fig_classification_err3.pdf -nocrop
end
end
```

The following program runs the second (multi-simulation) analysis for Chapter 6.

```
function [errors,data_cell,classif_cell,cutoff,results]=...
    Project_02_Part2(n,categories,simulations,draws,burn_in)
clc
data_scenarios = 4;
errors = zeros(simulations,2,data_scenarios);
data_cell = cell(simulations,data_scenarios);
cutoff = ones(simulations,categories-1,data_scenarios);
classif_cell = cell(simulations,data_scenarios,2);
zeros(data_scenarios*categories,2*categories,simulations);
headings = cell(1,5);
headings{1} = 'n'; headings{2} = 'cats'; headings{3} = 'sims';
headings{4} = 'draws'; headings{5} = 'burn_in';
xlswrite(strcat(['project2-2_classif' num2str(categories) '-' num2str(...
    n) '-' num2str(simulations) '.xlsx']),headings,'Sheet1','A1');
xlswrite(strcat(['project2-2_classif' num2str(categories) '-' num2str(...
    n) '-' num2str(simulations) '.xlsx']),[n categories simulations ...
    draws burn_in],'Sheet1','A2');
for sim = 1:simulations
     tic
     disp(strcat(['Simulation number ' num2str(sim)]))
     for data_model = 1:data_scenarios
         [latent,y,X,cutoff] = data_creator(data_model,n,categories,...
             cutoff,sim);
         data_cell{sim,data_model} = [latent y X];

         [beta,tau,V,¬,∆,¬,¬,Gamma_t] = gibbs_osrobit_fitter3(y,X,draws...
             ,burn_in);
         [beta_N,¬,Gamma_N]= gibbs_oprobit_fitter(y,X,draws,burn_in);
         G_N = median(Gamma_N,2);
         b_norm = median(beta_N,2);
         G_t = median(Gamma_t,2);
         b_t = median(beta,2)';
         d = median(∆);
         v = median(V);
         t = median(tau);

         [y,X] = new_sample(data_model,n,categories,cutoff(sim,:,...
             data_model));
         z_rep = X*b_norm + randn(n,1);
         [z_draws] = trunc_N(0,Inf,0,1,n);
         w_rep = X*b_t' + d*z_draws + t^(-0.5)*trnd(v);
         y_rep_N = ones(n,1);
         y_rep_t = ones(n,1);

         for i = 1:(categories-1)
             y_rep_N(z_rep≥cutoff(sim,i,data_model))=(i+1);
             y_rep_t(w_rep≥cutoff(sim,i,data_model))=(i+1);
         end
```

```
        normal_error = mean(abs(y-y_rep_N));
        t_error = mean(abs(y-y_rep_t));
        errors(sim,:,data_model) = [normal_error t_error];

        class_matrix_N = zeros(categories,categories);
        for i = 1:n
            tmp = zeros(categories,categories);
            tmp(y(i),y_rep_N(i)) = 1;
            class_matrix_N = class_matrix_N+tmp;
        end
        classif_cell{sim,data_model,1} = class_matrix_N;

        class_matrix_t = zeros(categories,categories);
        for i = 1:n
            tmp = zeros(categories,categories);
            tmp(y(i),y_rep_t(i)) = 1;
            class_matrix_t = class_matrix_t+tmp;
        end
        classif_cell{sim,data_model,2} = class_matrix_t;
        writer(data_model,n,categories,simulations,sim,normal_error,...
           t_error,class_matrix_N,class_matrix_t)
    end
    toc
end
    [results,dif_error_results,¬] = in_analyser(n,simulations,...
       categories,data_scenarios);
    xlswrite(strcat(['project2-2_classif' num2str(categories) '-' ...
       num2str(n) '-' num2str(simulations) '.xlsx']),results,'Sheet1'...
       ,'A4');
    xlswrite(strcat(['project2-2_classif' num2str(categories) '-' ...
       num2str(n) '-' num2str(simulations) '.xlsx']),...
       dif_error_results,'Sheet1','A8');
end

function [latent,y,X,cutoff] = data_creator(data_model,n,categories,...
   cutoff,sim)
x = 0.5+randn(n,1);
B_true = [-1 4]';
if data_model == 1;
    %Normal
    latent = B_true(1)+B_true(2)*x+randn(n,1);
elseif data_model == 2;
    %skew t data
    [z_draws] = trunc_N(0,Inf,0,1,n);
    latent = B_true(1)+B_true(2)*x-2*z_draws+0.5*trnd(5,n,1);
elseif data_model == 3;
    %exponential data
    latent = B_true(1)+B_true(2)*x-log(1-rand(n,1));
elseif data_model == 4;
    %uniform data
    latent = B_true(1)+B_true(2)*x+randi(6,n,1)-6;
end
X = [ones(n,1) x];
y = ones(n,1);
keep_splitting = 1;
attempt_counter = 0;
while keep_splitting == 1
```

```
    keep_splitting = 0;
    for i = 1:(categories-1)
        if i == 1
            a = ceil(min(latent));
        else
            a = cutoff(sim,(i-1),data_model);
        end
        b = floor(max(latent));
        cutoff(sim,i,data_model) = randi((b-a+1),1)+a-1;
        if i == 1
            counter1 = sum(latent<cutoff(sim,i,data_model));
        else
            counter1 = sum(latent<cutoff(sim,i,data_model) & latent>...
               cutoff(sim,(i-1),data_model));
        end
        %if i == (categories-1)
            counter2 = sum(latent>cutoff(sim,i,data_model));
        %end
        if min(counter1,counter2) < 0.02*n
            keep_splitting = 1;
            attempt_counter = attempt_counter + 1;
            disp(strcat(['CatSplit Attempt ' num2str(attempt_counter) ...
               ' not good, trying again.']))
            break
        end
    end
end
for i = 1:(categories-1)
    y(latent≥cutoff(sim,i,data_model))=(i+1);
end
end
function [y,X] = new_sample(data_model,n,categories,cutoffs)
x = 0.5+randn(n,1);
B_true = [-1 4]';
if data_model == 1;
    %Normal
    latent = B_true(1)+B_true(2)*x+randn(n,1);
elseif data_model == 2;
    %skew t data
    [z_draws] = trunc_N(0,Inf,0,1,n);
    latent = B_true(1)+B_true(2)*x-2*z_draws+0.5*trnd(5,n,1);
elseif data_model == 3;
    %exponential data
    latent = B_true(1)+B_true(2)*x-log(1-rand(n,1));
elseif data_model == 4;
    %uniform data
    latent = B_true(1)+B_true(2)*x+randi(6,n,1)-6;
end
X = [ones(n,1) x];
y = ones(n,1);
for i = 1:(categories-1)
    y(latent≥cutoffs(i))=(i+1);
end
end
function writer(data_model,n,categories,simulations,sim,normal_error,...
   t_error,class_matrix_N,class_matrix_t)
xlswrite(strcat(['project2-2_classif' num2str(categories) '-' num2str(...
   n) '-' num2str(simulations) '.xlsx']),[normal_error t_error],...
   strcat(['errors_dm' num2str(data_model)]),strcat(['A' num2str(sim)...
   ]));
xlswrite(strcat(['project2-2_classif' num2str(categories) '-' num2str(...
   n) '-' num2str(simulations) '.xlsx']),class_matrix_N,strcat(['...
   probit_dm' num2str(data_model)]),strcat(['A' num2str((sim-1)*...
   categories+1)]));
xlswrite(strcat(['project2-2_classif' num2str(categories) '-' num2str(...
   n) '-' num2str(simulations) '.xlsx']),class_matrix_t,strcat(['...
   strobit_dm' num2str(data_model)]),strcat(['A' num2str((sim-1)*...
   categories+1)]));
end
function [results,dif_error_results,extreme_index] = in_analyser(n,...
   simulations,categories,data_scenarios)
errors = zeros(simulations,2*data_scenarios);
classif_N = zeros(categories*simulations,categories,data_scenarios);
classif_t = zeros(categories*simulations,categories,data_scenarios);

for data_model = 1:data_scenarios
    errors(:,((2*(data_model-1)+1):(2*(data_model)))) = xlsread(strcat...
        (['project2-2_classif' num2str(categories) '-' num2str(n) '-' ...
        num2str(simulations) '.xlsx']),strcat(['errors_dm' num2str(...
        data_model)]));
    classif_N(:,:,data_model) = xlsread(strcat(['project2-2_classif' ...
        num2str(categories) '-' num2str(n) '-' num2str(simulations) '...
        .xlsx']),strcat(['probit_dm' num2str(data_model)]));
    classif_t(:,:,data_model) = xlsread(strcat(['project2-2_classif' ...
        num2str(categories) '-' num2str(n) '-' num2str(simulations) '...
        .xlsx']),strcat(['strobit_dm' num2str(data_model)]));
end
observed_prop = permute((sum(classif_N,2)/n),[1 3 2]);


extreme_index = zeros(simulations,data_scenarios);
for j = 1:data_scenarios
    for i = 1:simulations
        working = observed_prop((categories*(i-1)+1):(categories*i),j)...
            ;
        extreme_index(i,j) =  -log(working(1)) -log(working(end));
    end
end
N_errors = errors(:,(1:2:(data_scenarios*2)-1));
t_errors = errors(:,(2:2:(data_scenarios*2)));
dif_errors = N_errors - t_errors;

fig_name = strcat(['fig_MAD_scarcity' num2str(categories) '_' num2str(...
   n) '_' num2str(simulations) '.pdf']);
figure(1)
set(1,'units','normalized','position',[0.05 0.05 0.4 0.4]);
hold on
scatter(extreme_index(:,1),dif_errors(:,1),'ob')
scatter(extreme_index(:,2),dif_errors(:,2),'+r')
scatter(extreme_index(:,3),dif_errors(:,3),'xk')
scatter(extreme_index(:,4),dif_errors(:,4),'^g')
title(strcat(['probit minus strobit MAD error, by scarcity within tail...
    categories, n = ' num2str(n)]))
legend('N data','skew t data','Exp data','Unf data','location','...
    NorthEastOutside')
grid on
hold off
export_fig figure.pdf -nocrop
movefile('figure.pdf',fig_name)

results = [mean(N_errors<t_errors);mean(N_errors==t_errors);mean(...
    N_errors>t_errors)];


%Plotting error bar graphs
q_errors = quantile(errors,[0.975 0.025]);
m_errors = mean(errors);
q_dist = abs(q_errors-repmat(m_errors,2,1));


figure(2)
set(2,'units','normalized','position',[0.05 0.05 0.3 0.3]);
errorbar(m_errors,ones(1,length(m_errors)),'.')
h=errorbar(m_errors,ones(1,length(m_errors)),'.');
set(h,'UData',q_dist(1,:)')
set(h,'LData',q_dist(2,:)')
set(gca,'YGrid','on')

q_dif_errors = quantile(dif_errors,[0.975 0.025]);
m_dif_errors = mean(dif_errors);
q_dif_dist = abs(q_dif_errors-repmat(m_dif_errors,2,1));
dif_error_results = [q_dif_errors(2,:)' m_dif_errors' q_dif_errors...
    (1,:)'];

fig_name = strcat(['fig_MAD_scenario' num2str(categories) '_' num2str(...
    n) '_' num2str(simulations) '.pdf']);
figure(3)
set(3,'units','normalized','position',[0.05 0.05 0.25 0.25]);

errorbar(m_dif_errors,ones(1,length(m_dif_errors)),'x')
h=errorbar(m_dif_errors,ones(1,length(m_dif_errors)),'x');
title(strcat(['probit minus strobit MAD error, with 95% empirical ...
    interval, n = ' num2str(n)]))
set(h,'UData',q_dif_dist(1,:)')
set(h,'LData',q_dif_dist(2,:)')
set(gca,'YGrid','on')
set(gca,'XTick',[1 2 3 4])
set(gca,'XTickLabel',{'N data', 'skew t data', 'Exp data', 'Unf data'...
    })
export_fig figure.pdf -nocrop
movefile('figure.pdf',fig_name)
end
```

In Chapter 6, if the (ordered) probit is fitted, the following program is used.

```
function [beta,Z,Gamma]= gibbs_oprobit_fitter(y,X,sims,burn_in)
%y is the categorical respopnse vector (with obs in each category 1:J)
%X is n x p+1, an intercept and p slopes
%beta is the (p+1) x sims final set of beta draws (after burn-in)
%Z is the n x sims final set of underlying latent draws (after burn-in...
    )
```

```
%Gamma is the (J-1) x sims set of bounds for the categories

categories = length(unique(y));
[n p1] = size(X);
beta = zeros(p1,sims);

Z = zeros(n,sims);
Gamma = zeros(categories+1,sims);

%Initialisation
b = regress(y,X);
cov = inv(X'*X);
Bcomp = cov*(X');
bs = mvnrnd(b,cov)';
z = X*bs +randn(n,1); %creating a non-grouped z just to make initial ...
    bounds

if categories == 2
    G = [-Inf 0 Inf];
else
    tmp = levelcounts(ordinal(y));
    tmp = cumsum(tmp)./n;
    tmp = tmp(1:(end-1));
    G = [-Inf quantile(z,tmp') Inf]; %J+1 endpoint, J-1 limits, J cats
end
for j = 1:categories
    z(y==j) = trunc_N(G(j),G(j+1),X(y==j,:)*bs,ones(sum(y==j),1),1);
end
bs = mvnrnd(Bcomp*z,cov)';

%Initialisation complete
for simulation = 1:(burn_in+sims)
    %drawing bounds
    if categories == 2
        G = [-Inf 0 Inf];
    else
        [G,¬] = gibbs_probit_bounds(y,z,G);
    end
    %drawing new z's
    for j = 1:categories
        z(y==j) = trunc_N(G(j),G(j+1),X(y==j,:)*bs,ones(sum(y==j),1)...
            ,1);
    end
    %drawing new betas
    bs = mvnrnd(Bcomp*z,cov)';
    if simulation > burn_in
        beta(:,simulation-burn_in) = bs;
        Z(:,simulation-burn_in) = z;
        Gamma(:,simulation-burn_in) = G;
    end
end
```

Similarly, in Chapter 6, if the (ordered) strobit is fitted, the following program is used.

```
function [beta,tau,V,lambdas,∆,Z,W,Gamma]= gibbs_osrobit_fitter3(y,X,...
    draws,burn_in)
%y is the categorical respopnse vector (with obs in each category 1:J)
```

```
%X is n x p+1, an intercept and p slopes
%beta is the (p+1) x sims final set of beta draws (after burn-in)
%W is the n x sims final set of underlying latent draws (after burn-in...
    )
%Gamma is the (J-1) x sims set of bounds for the categories

%Let's fix the draws across the columns
categories = length(unique(y));
[n,p1] = size(X);
beta = zeros(p1,draws+burn_in);
W = zeros(n,draws+burn_in);
V = 3*ones(1,draws+burn_in);
∆ = zeros(1,draws+burn_in);
Gamma = zeros(categories+1,draws+burn_in);
tau = ones(1,draws+burn_in);
lambdas = ones(n,draws+burn_in);
Z = ones(n,draws+burn_in);



%The fitter must restart if the truncation drawer fails.
[b,t,v,¬,d,z] = draw_gibbs_t_skew1(y, X, 1);
b = b'; z = z';
counter = 0;
while sum(isnan(b))>0
    counter = counter+1;
    if counter > 5
        restarter = 1;
    end
    if restarter == 1; break; end %%
    disp(strcat(['Initialisation draws counter ' num2str(counter)]))
    [b,t,v,¬,d,z] = draw_gibbs_t_skew1(y, X, 1);
    b = b'; z = z';
end

w = X*b + z.*d + t^(-0.5)*trnd(v,n,1); %PROBLEM HERE?

if categories == 2
    G = [-Inf 0 Inf];
else
    tmp = levelcounts(ordinal(y));
    tmp = cumsum(tmp)./n;
    tmp = tmp(1:(end-1));
    G = [-Inf quantile(w,tmp') Inf]; %J+1 endpoint, J-1 limits, J cats
end

for j = 1:categories
    w(y==j) = trunc_nct(G(j),G(j+1),X(y==j,:)*b + z(y==j,1).*d,t^-0.5*...
        ones(length(y(y==j)),1),v.*ones(length(y(y==j)),1),1);
end
W(:,1) = w;
[b,t,v,¬,d,z] = draw_gibbs_t_skew1(w, X, 1);
beta(:,1) = b';
V(1) = v;
∆(1) = d;
tau(1) = t;
Z(:,1) = z';
```

```

round = 1;
while round < (burn_in+draws)
    round = round+1;
    if categories == 2
        G = [-Inf 0 Inf];
    else
        [G,¬] = gibbs_probit_bounds(y,w,G);
    end

    w = zeros(n,1);
    for j = 1:categories
        w(y==j) = trunc_nct(G(j),G(j+1),X(y==j,:)*beta(:,(round-1)) + ...
           Z(y==j,(round-1)).*Δ(round-1),tau(round-1)^-0.5*ones(...
           length(y(y==j)),1),V(round-1).*ones(length(y(y==j)),1),1);
    end
    W(:,round) = w;

    [b,t,v,l,d,z] = draw_gibbs_t_skew1(W(:,round), X, 1, beta(:,(round...
       -1))',tau(round-1),V(round-1),lambdas(:,(round-1))',Δ(round-1)...
       ,Z(:,(round-1))');
    b = b'; z = z';
    counter = 0;
    if sum(isnan(b))>0
        counter = counter+1;
        disp(strcat(['Beta NaNs after trunced w, counter ' num2str(...
           counter) ', round ' num2str(round) ', going back 1 round.'...
           ]))
        round = round-2;
    else
        beta(:,round) = b;
        W(:,round) = w;
        V(1,round) = v;
        Δ(1,round) = d;
        Gamma(:,round) = G';
        tau(1,round) = t;
        lambdas(:,round) = l';
        Z(:,round) = z;
    end

end
beta = beta(:,(burn_in+1):end);
W = W(:,(burn_in+1):end);
V = V(:,(burn_in+1):end);
Δ = Δ(:,(burn_in+1):end);
Gamma = Gamma(:,(burn_in+1):end);
tau = tau(:,(burn_in+1):end);
lambdas = lambdas(:,(burn_in+1):end);
Z = Z(:,(burn_in+1):end);


function [beta,tau,v,lambdas,Δ,Z] = draw_gibbs_t_skew1(y, X, draws, ...
   varargin)
%variable arguments:
%var arg 1: burn_in
%var arg 2: prior (1: trunc exp; 2: Ind Jeff; 3: Ref 1; 4: Ref 2)
%var arg 3: discretised (1) or MH (0) [not implemented yet]
```

```

% y is a column vector
% X includes a column of ones for the intercept term

% if nargin > 3
%     burn_in = varargin{1};
% else
    burn_in = 50;
% end
% if nargin > 4
%     prior_num = varargin{2};
% else
    prior_num = 2;
% end
% if nargin > 5
%     v_discrete = varargin{3};
% else
%     v_discrete = 1;
% end

[n p] = size(X);

%Priors

mu∆ = 0; %these are the parameters on the proper Normal prior for ∆
sig2∆ = 1000;
mubeta = 0*ones(1,p);
sig2beta = 10000*ones(1,p);
agamma = 0.1;
lgamma = 0.1;
muz = 0;
sig2z = 1;

tau = ones((burn_in+draws),1);
lambdas = ones((burn_in+draws),n);
beta = zeros((burn_in+draws),p);
Z = repmat(abs(randn(1,n)),burn_in+draws,1).*ones((burn_in+draws),n);
∆ = zeros((burn_in+draws),1);
v = 3*ones((burn_in+draws),1);

if length(varargin) == 6
    beta(1,:) = varargin{1};
    tau(1) = varargin{2};
    v(1) = varargin{3};
    lambdas(1,:) = varargin{4};
    ∆(1) = varargin{5};
    Z(1,:) = varargin{6};
else
    b = regress(y,X);
    beta(1,:) = b';
    [v_draws] = draw_post_v_discr(1, n, lambdas(1,:), prior_num);
    v(1) = v_draws(end);
end
ystar = zeros(n,p);
ycurl = zeros(n,burn_in+draws);

for i = 2:(burn_in+draws)
```

```
    if p>1
        for j = 1:p
            switch j
                case 1
                    ystar(:,j) = y - X(:,(j+1):end)*beta((i-1),(j+1)...
                        :end)' - Δ(i-1)*Z((i-1),:)';
                case p
                    ystar(:,j) = y - X(:,1:(end-1))*beta((i),1:(end-1)...
                        )' - Δ(i-1)*Z((i-1),:)';
                otherwise
                    ystar(:,j) = y - X(:,[1:(j-1) (j+1):end])*[beta((i...
                        ),1:(j-1)) beta((i-1),(j+1):end)]' - Δ(i-1)*Z...
                        ((i-1),:)';
            end
            beta(i,j) = randn.* (tau(i-1)*lambdas((i-1),:)*(X(:,j).^2)...
                +1/sig2beta(j))^(-0.5) + (tau(i-1).*lambdas((i-1),:)*(...
                X(:,j).^2)+1/sig2beta(j))^(-1) * (tau(i-1).*lambdas((i...
                -1),:)*(X(:,j).*ystar(:,j))+mubeta(j)/sig2beta(j));
        end
    else
        ystar(:,1) = y - X*beta(i-1) - Δ(i-1)*Z((i-1),:)';
        beta(i,1) = randn.* (tau(i-1)*lambdas((i-1),:)*(X(:,1).^2)+1/...
            sig2beta(1))^(-0.5) + (tau(i-1).*lambdas((i-1),:)*(X(:,1)....
            ^2)+1/sig2beta(1))^(-1) * (tau(i-1).*lambdas((i-1),:)*(X...
            (:,1).*ystar(:,1))+mubeta(1)/sig2beta(1));
    end
    ycurl(:,i) = y - X*beta(i,:)' - Δ(i-1)*Z((i-1),:)';
    %tau(i) = gamrnd(n/2 + agamma,(0.5*lambdas((i-1),:)*(ycurl(:,i)....
        ^2) + 2*lgamma)^-1);
    yhat = y - X*beta(i,:)';
    Z(i,:) = trunc_N(0,Inf,(tau(i).*lambdas(i-1,:).*Δ(i-1)^2 +1).^-1 ....
        * tau(i).*lambdas(i-1,:).*Δ(i-1).*yhat',sqrt((tau(i).*lambdas(...
        i-1,:).*Δ(i-1)^2  +1      ).^-1),1);
    Δ(i) = randn* (tau(i)*lambdas((i-1),:)*(Z(i,:).^2)'+1/sig2z)^(-0.5...
        ) + (tau(i)*lambdas((i-1),:)*(Z(i,:).^2)'+1/sig2z)^(-1) * (tau...
        (i)*lambdas((i-1),:)*(Z(i,:)'.*yhat) + muz/sig2z);
    ycurl(:,i) = y - X*beta(i,:)' - Δ(i)*Z((i),:)';
    lambdas(i,:) = chi2rnd(v(i-1)+1,[1,n]) ./ (v(i-1)+tau(i)*(ycurl(:,...
        i)').^2);
    [v(i)] = draw_post_v_discr(1, n, lambdas(i,:), prior_num,i); %...
        discretised v or not?
end

beta = beta((burn_in+1):end,:);
tau = tau((burn_in+1):end,1);
lambdas = lambdas((burn_in+1):end,:);
v = v((burn_in+1):end,1);
Z = Z((burn_in+1):end,:);
Δ = Δ((burn_in+1):end,1);
```

In both the preceding programs, the following function is used.

```
function [G,bounds] = gibbs_probit_bounds(y,z,G)
%G_initial is a vector size J of the initial bounds (also giving ...
    length)
%G_initial includes -Inf and +Inf
%G is a vector size J of the drawn bounds
```

```
%z is a vector size n of the observed underlying latent draws
%y is a vector size n of the actual categories observed

J = length(G); %4
bounds = zeros(J,2);
bounds(1,:) = -Inf * ones(1,2);
bounds(2,:) = zeros(1,2);
bounds(J,:) = Inf * ones(1,2);
G(2) = 0;
if J > 3
for j = 2:(J-2)
    bounds(j+1,1) = max(max(z(y==j)),G(j));
    bounds(j+1,2) = min(min(z(y==(j+1))),G(j+2));
    G(j+1) = rand*(bounds(j+1,2)-bounds(j+1,1))+bounds(j+1,1);
end
end
%G = G(2:(end-1));
%s_bounds = sort(bounds,2);
%G = rand(J,1).*(bounds(:,2)-bounds(:,1))+bounds(:,1);
```

B.5 General Programs

If ever truncated Normal, t, or non-central t values are required, the following functions are called.

```
function [draws] = trunc_N(lower,upper,mu,sigma,n)
%mu and sigma are vectors. Multiple draws create a matrix. If mu and ...
    sigma
%are matrices, a new dimension is created with the draws for each ...
    pair. Only
%2 dimensions for mu and sigma are supported.

[rows cols] = size(mu);
l = repmat(lower,[rows cols n]);
u = repmat(upper,[rows cols n]);
mu = repmat(mu,[1 1 n]);
sigma = repmat(sigma,[1 1 n]);
Z = rand(rows,cols,n);
draws = norminv((normcdf(u,mu,sigma)-normcdf(l,mu,sigma)).*Z + normcdf...
    (l,mu,sigma),mu,sigma);
% if sum(draws) == Inf
%     [Z draws u l mu sigma normcdf(u,mu,sigma) normcdf(l,mu,sigma)]
% end
if rows == 1 %draws down columns
    draws = permute(draws,[3 2 1]);
elseif cols == 1; %draws across rows
    draws = permute(draws,[1 3 2]);
end
```

```
function [draws] = trunc_t(lower,upper,mu,sigma,v,n)
%with this function, centred truncated t-values are given.
%v is a vector of d.f.
%Multiple draws create a matrix. If v is
%a matrices, a new dimension is created with the draws for v. Only
%2 dimensions for v are supported.

[rows cols] = size(v);
l = repmat(lower,[rows cols n]);
u = repmat(upper,[rows cols n]);
Z = rand(rows,cols,n);
draws = tinv((tcdf(u,v)-tcdf(l,v)).*Z+tcdf(l,v),v);
if rows == 1 %draws down columns
    draws = permute(draws,[3 2 1]);
elseif cols == 1; %draws across rows
    draws = permute(draws,[1 3 2]);
end
```

```
function [draws] = trunc_nct(lower,upper,mu,sigma,v,n)
%with this function, noncentred truncated t-values are given.
%v is a vector of d.f.
%mu is the vector of non-centrality parameters, same size as v.
```

```
[rows,cols] = size(v);
l = (repmat(lower,[rows,1,n])-mu)./sigma;
u = (repmat(upper,[rows,1,n])-mu)./sigma;
Z = rand(rows,1,n);
if lower == -Inf
    a = zeros(rows,1);
else
    a = tcdf(l,v);
end
if upper == Inf
    b = ones(rows,1);
else
    b = tcdf(u,v);
end
draws = sigma.*tinv((b-a).*Z+a,v)+mu;
if rows == 1 %draws down columns
    draws = permute(draws,[3 2 1]);
elseif cols == 1; %draws across rows
    draws = permute(draws,[1 3 2]);
end
```