UNSW

THE UNIVERSITY OF NEW SOUTH WALES
SYDNEY • AUSTRALIA

University of New South Wales
School of Economics

Honours Thesis

A Variational Bayes Approach To Copula Estimation

Author:
Anny Francis
Student ID: 3415655

Supervisors:
Dr. David Gunawan
Scientia Prof. Robert Kohn

Bachelor of Economics (Econometrics & Economics) (Honours)
and
Bachelor of Commerce and Science (Actuarial, Statistics & Mathematics)

22nd November, 2019

Declaration

I declare that this thesis is my own work and that, to the best of my knowledge, it contains no material which has been written by another person or persons, except where acknowledgement is made. This thesis has not been submitted for the award of any degree or diploma at the University of New South Wales Sydney, or at any other institute of higher education.

....

Anny Francis
22nd November, 2019

Acknowledgements

To Dr. David Gunawan and Prof. Robert Kohn, thanks for giving me the opportunity to learn from both of you. You have taught me everything I know about Bayesian Statistics. From only knowing what a posterior distribution is at the start of this year to now knowing variational Bayes, sequential Monte Carlo, how to estimate random effects models, and more, words cannot explain how grateful I am. Not only have I come into the honours program wanting to see what research is like, but I have learnt much more. I have seen what it takes to become a confident, reliable, and dedicated researcher. And importantly, learning the skills, having the knowledge to learn cutting-edge topics, and my own path to contribute to research. Thanks again for all the support, advice, mentorship, encouragement, and for taking me on as a student. I want to continue to learn from both of you!

I am grateful for my managers, department, and team at the Reserve Bank of Australia for supporting my studies and the flexibility to attend classes. It has been a wonderful journey to see the world though an economic lens. I have enjoyed every moment of it!

To Robert Salomone and Tess Stafford thanks for providing me with valuable comments. To Hung, Igor, KD and Yu it was fun hanging out with you guys this year. Thanks to the 2019 honours cohort for being great people. I wish I got to bond with you guys more but from the times we have spent together, it has been joyous!

Lastly, thank you to my family, especially mum, grandma, Connie and Duncan for giving me the endless support to pursue my studies. I would not have had such a great learning experience without your support.

Cheers,
Anny

Contents

List of Figures

List of Tables

Abstract

Demand for models to understand discrete datasets has increased with the availability of survey data. This thesis extends recent literature on modelling the dependence structure of a large set of discrete random variables using a copula. Copulas can be used to flexibly model the joint distribution of a set of random variables. The Gaussian copula is one of the most popular copulas used by practitioners. It is simpler than other copulas and is able to capture the symmetric dependence of high-dimensional datasets. For these reasons, this thesis focuses on the Gaussian copula. Standard Bayesian and Maximum Likelihood methods are typically used to estimate the copula parameters when the data is continuous. However, these methods are infeasible for high-dimensional discrete datasets. The reason is that for N J-dimensional observations, evaluating the likelihood requires $N \times 2^J$ evaluations of the copula function which is expensive when both N and J are large. I use parsimonious representations of the Gaussian copula and an augmented posterior distribution to overcome the computational challenges involved with estimating the copula parameters. Additionally, I use Variational Bayes (VB) to approximate the posterior distribution of the copula parameters with a simpler distribution, which is also easy to simulate from. I show in simulation studies that VB is an efficient alternative to traditional Bayesian methods, such as Markov Chain Monte Carlo (MCMC), for estimating the copula parameters. The VB approach leads to over a forty-fold improvement in computation time with little loss of accuracy compared to MCMC. Extensions to ordinal data are also considered for modelling survey data. In an empirical application to the Household, Income and Labour Dynamics in Australia survey, I model the dependence structure of responses to $J = 36$ health questions. The increased efficiency of the proposed VB method allows for fast inference and prediction. I show this, by constructing a Bayesian multidimensional health index to track health over time and exploiting properties of the copula to understand the relationships between responses to the survey questions. Example MATLAB code is also provided.

Chapter 1
Introduction

Surveys such as the Household, Income and Labour Dynamics in Australia survey (HILDA), British Household Panel survey, and Growing Up in Ireland study serve as useful sources of information to understand health. For example, there are 36 health questions from the self-completion questionnaire in HILDA, all of which have discrete responses. The questions, such as rating one's ability to climb a flight of stairs, walk 100 metres, or lift or carry groceries, aim to gauge physical and mental ability, the impact of health on work completion, and general reflections on health. Hence, models built to understand discrete datasets are important in survey data applications. This thesis provides an efficient and flexible methodology to understand how these health responses relate to each other. For example, whether a person has a positive outlook on life given that they are 'limited a lot' in being able to complete physical activities. In particular, I use this methodology to construct a Bayesian multidimensional health index to track health over time.

The methodology in this thesis uses copulas to model the joint dependence of a set of discrete variables. A copula is a cumulative distribution function which binds together marginal distributions of a set of random variables to determine their joint distribution. The marginal distributions are also referred to as the margins of a copula. For the HILDA example, each health question can be modelled as a discrete random variable and a copula can be used to determine the probability that a person responds negatively to all physical health questions. In related health applications, Stander et al. (2019) use copulas to identify children with unusual eyesight in both eyes, who otherwise would not be detected when using a univariate model. Murteira and Lourenço (2011) use a bivariate copula to model the relationship between the number of doctor visits and self-assessed health given a set of covariates. They use a bivariate copula to avoid the endogeneity problem of the number of doctor visits and self-assessed health, inherent in regression, as they are both driven by past healthcare receipts.

The copula parameters for discrete data can be estimated using a Bayesian framework. Pitt et al. (2006) were one of the first to derive a method to estimate the parameters of a Gaussian copula with discrete margins using Markov Chain Monte

Carlo (MCMC). Smith and Khaled (2012) extend the sampling scheme of Pitt et al. (2006) to copula families where the conditional copula distribution can be evaluated, such as Archimedean copulas. As an alternative to the method of Smith and Khaled (2012), Gunawan et al. (2019) propose using Pseudo-Marginal MCMC to estimate the parameters of Archimedean copulas with discrete margins.

Although MCMC methods are exact, they can be very slow compared to approximate methods such as variational Bayes (VB), which involves approximating the true posterior distribution with a simpler distribution that is also easy to simulate from. Despite the approximate nature of VB, the increase in speed makes prediction, cross-validation, and approximate inference, which would otherwise be difficult, possible. Recently, Gunawan et al. (2019) and Loaiza-Maya and Smith (2019) propose VB methods as an alternative to MCMC methods to estimate the parameters of copulas with discrete margins. Gunawan et al. (2019) propose a Variational Bayes Intractable Likelihood (VBIL) method, which builds on the method of Tran et al. (2019), to estimate the parameters of Archimedean copulas with discrete margins. Although Archimedean copulas can capture asymmetric tail dependence, capturing the dependence structure of a large dataset is difficult as there is only one copula parameter. For example, the dependence structure of the 36 health questions from HILDA is unlikely to be captured by one parameter. A more flexible copula which also captures asymmetric dependence is the D-Vine copula, which is constructed from a sequence of bivariate "pair-copulas". Loaiza-Maya and Smith (2019) propose a Variational Bayes Data Augmentation (VBDA) method, which builds on the MCMC method of Smith and Khaled (2012), for estimating the parameters of a D-Vine copula.

Following recent developments in the VB literature, this thesis makes three contributions. The first and main contribution extends VBIL and VBDA to the Gaussian copula with discrete margins. The Gaussian copula is one of the most popular copulas used by practitioners (Stander et al., 2019; Conlon et al., 2017; Murray et al., 2013). Similarly to D-Vine copulas, Gaussian copulas can capture the dependence of high-dimensional datasets flexibly, as each pair of variables has its own parameter. Additionally, Gaussian copulas are simple and can capture the symmetric dependence of a set of random variables through a correlation structure. The second contribution is to extend VBIL to ordinal data applications. This extension increases the applicability of VBIL to analyse survey data, many of which have ordinal responses. This extension also addresses a remark by Dunson (2010) that categorisation of ordinal and categorical responses into Bernoulli data results in a reduction in power to detect associations. The third contribution is to compare the performance of VBIL, VBDA and MCMC for Gaussian copulas with discrete margins. I find that VB methods provide a good approximation of the posterior distribution of the copula parameters with little loss in accuracy. Of the two VB methods, VBIL performs better than VBDA in terms of time and accuracy, with MCMC assumed to be the most accurate. VBIL is the main focus of the thesis because I found it to be over 40 times faster than MCMC in the simulation studies considered.

I model the 36 health questions in HILDA to illustrate the applicability of VBIL for Gaussian copulas with discrete margins. I use the posterior distribution of the copula parameters to construct a Bayesian multidimensional health index to track the health of Australian women over time and to exploit the properties of copulas to understand joint and conditional dependence across the health dimensions. The Bayesian multidimensional health index is based on the multidimensional poverty measure of Alkire and Foster (2011). Each health question in HILDA forms a dimension in the multidimensional health measure. By definition, a person is deprived in a dimension if their selected response is equal to or below a predefined cutoff for this dimension. For example, a person is deprived in dimension 'able to climb a flight of stairs', if limited a lot is chosen as the dimension cutoff and it is selected as a response. Additionally, a person is classified as unhealthy if their total number of deprivations is equal to or larger than a predefined cut-off k (Alkire and Foster, 2011). Based on these definitions, the multidimensional health measure (M_0) explains the share of deprivations faced by persons that are unhealthy out of the total possible deprivations that can be faced by the population.

Under a Bayesian framework, inference about M_0 is based on its posterior distribution. Figure 1.1 shows estimates of the posterior density, the posterior mean, the posterior median, and the 95 percent credibility intervals of M_0 for year 2004. The credibility interval, similar to the confidence interval, can be used to quantify the uncertainty of an estimate. Quantifying uncertainty is important because research findings often shape government policy and those of private health providers in determining resource allocation. Statistically significant changes can be identified and can suggest when a policy response is required. Similarly to the multidimensional poverty measure of Alkire and Foster (2011), the Bayesian multidimensional health index can be used to monitor reductions in the number of people who are classified as unhealthy, evaluate the effectiveness of government policies to improve health, and identify areas of health which require more government funding (Alkire et al., 2018). See Chapter 7 for more details on M_0.

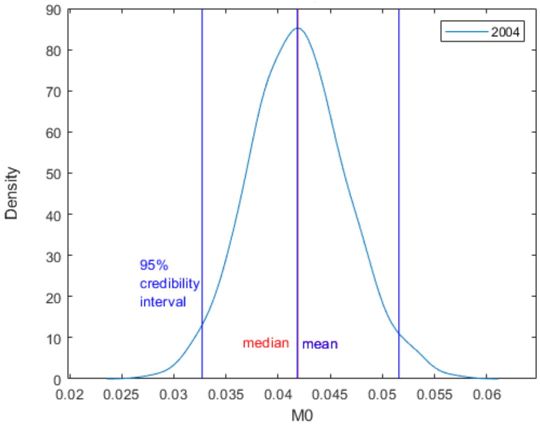


Figure 1.1: Estimates of the posterior density, posterior mean, posterior median, and 95 percent credibility intervals of the multidimensional health measure M_0 for 2004 and $k = 3$

The rest of the thesis is organised as follows: Chapter 2 formally reviews copulas with a focus on the Gaussian copula. Chapter 3 outlines what is required in Bayesian inference to estimate the posterior distribution of parameters of copulas with discrete margins. Chapter 4 reviews exact Bayesian inference, including MCMC methods, used to estimate the copula parameters. Chapter 5 provides a review of approximate Bayesian inference and gives details of the VBIL method for the Gaussian copula. Appendix J gives details of VBDA. Chapter 6 evaluates the efficiency and accuracy of VBIL compared to VBDA and MCMC through simulations studies. Chapter 7 provides example applications of VBIL to the HILDA data to model health over time. Chapter 8 concludes and discusses future research. There are twelve appendices which include proofs, explanations of concepts and figures.

Chapter 2
Copulas

Copulas are applied in many disciplines, including health, economics, finance, medicine, mathematics, and marketing. For example, Patton (2006) uses copulas to show that there is an asymmetric relationship between the mark-dollar and yen-dollar exchange rates prior to the introduction of the Euro, with depreciations of the two currency pairs more likely to occur at the same time than appreciations. Patton remarks that this relationship is likely to reflect the attempts by central banks to maintain a competitive exchange rate. Choroś et al. (2010) use a copula to model a portfolio with a large number of assets to price collaterised debt obligations and find that there is negative tail dependence, implying that portfolio managers should consider tail dependence when selecting assets to diversify their portfolios. In a marketing application, Panagiotou and Stavrakoudis (2015) study the effect of changing market conditions on the prices of different grades of pork cutlets using a copula and conclude that there is no asymmetric price response.

2.1 Overview of Copulas

Copulas are functions which bind together information from univariate margins to model the joint distribution. Consider the bivariate case with random variables Y_1 and Y_2. Let the marginal cumulative distribution functions of Y_1 and Y_2 be $F_1(y_1)$ and $F_2(y_2)$. A copula can be represented as a function of $F_1(y_1)$ and $F_2(y_2)$, with the joint cumulative distribution of Y_1 and Y_2

$$\mathbb{P}(Y_1 \leq y_1, Y_2 \leq y_2) = C(F_1(y_1), F_2(y_2)). \tag{2.1}$$

For continuous Y, the corresponding joint density of Y_1 and Y_2 is

$$f(y_1, y_2) = c(F_1(y_1), F_2(y_2))f(y_1)f(y_2), \tag{2.2}$$

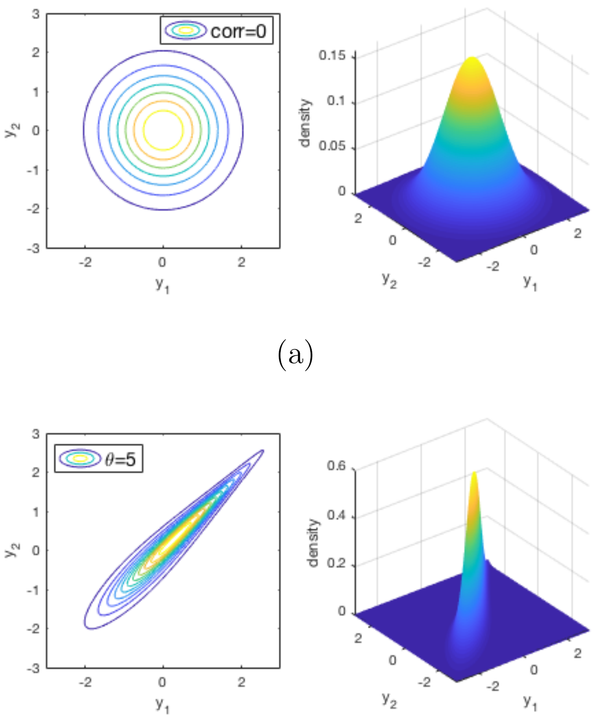
where $f_1(y_1)$ and $f_2(y_2)$ are the marginal probability density functions of Y_1 and Y_2. I refer to C as the copula cumulative distribution function (CDF) and c as the copula density. Sklar's theorem (Nelsen, 2007) states that (1) only the univariate margins and a copula CDF are required to construct the joint distribution; and (2) conversely, given the margins and the joint distribution, there exists a copula CDF which satisfies Equation 2.1. Theoretically, a copula is unique when the margins are all continuous but is not when they are discrete. In practice, the aim of copula modelling is not to find the copula which satisfies Sklar's theorem, but to construct a joint distribution given the margins and a copula (Smith, 2011). This implies that the non-uniqueness of a copula with discrete margins is not of concern. Furthermore, the same copula can be used for both continuous and discrete margins. See Nelsen (2007) for a more rigorous mathematical definition of copulas.

There are two main benefits of using a copula to model the joint distribution. The first is that the joint dependence can be modelled separately from the margins. The parameters of each margin can be estimated individually for each dimension and then the dependence between the margins can be introduced through a copula.[1] This separation makes it easier to obtain the joint distribution in high dimensions. The second benefit is that the marginal distributions implied by the copula construction are consistent with estimates of the marginal distributions when these are separately estimated. Hence, the margins of a copula can follow any distribution, including parametric distributions, such as Poisson, Dirichlet, and inverse gamma; non-parametric distributions such as empirical distributions and rank; time series models such as generalised autoregressive conditional heteroskedasticity; and regression models (Smith, 2011).

I construct a series of contour and density plots to illustrate the joint dependence of Y associated with different copulas. To build intuition, I only consider the case when Y is continuous because it is easier to illustrate than the discrete case. Consider first Figure 2.1 (a), which plots the joint density of Y modelled using to a Gaussian copula density with standard normal margins and zero correlation i.e. Y_1 and Y_2 follow standard normal distributions. The density in Figure 2.1 (a) is equivalent to the joint density of Y derived analytically - a bivariate normal with mean zero and covariance matrix equal to the identity matrix. The joint density is symmetric around zero, implying that both Y_1 and Y_2 are positive and negative with the same probability. Additionally, the bivariate normal has zero correlation which means that no additional information about Y_2 is available when Y_1 is known and vice versa. Asymmetry is introduced into the joint density of Y when the Gaussian copula density is replaced with a Gumbel copula density. With the same standard normal margins, Figure 2.1 (b) plots the joint density of Y constructed using a Gumbel copula density with parameter equal to 5. Unlike the joint density of Y in Figure 2.1 (a), which is symmetric around zero, this density exhibits positive tail

[1] The parameters of a copula and its margins can also be estimated at the same time. Although for most problems it is easier to separate the two.

dependence, which means that if Y_1 is large and positive, then Y_2 will also be large and positive with a positive probability. A consequence of positive tail dependence is that the probability of both Y_1 and Y_2 being greater than zero is larger than 0.25, whereas it is equal to 0.25 for the joint density of Y modelled using a Gaussian copula density with standard normal margins. This example illustrates that the choice of copula affects the joint dependence of Y.

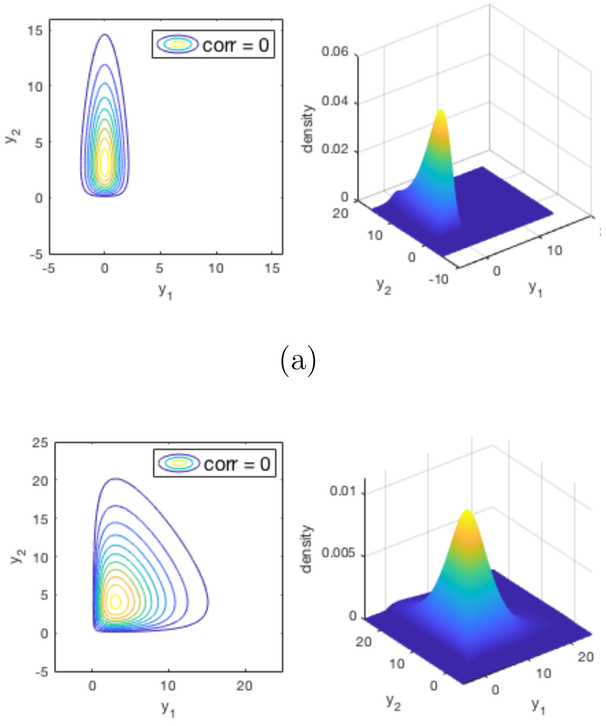


(a)

(b)

Figure 2.1: Contour and density plots of Y modelled using a bivariate (a) Gaussian copula with standard normal margins and zero correlation (b) Gumbel copula with standard normal margins and copula parameter equal to 5

For a given copula density, the marginal distributions and densities of Y control the concentration of mass in the corresponding joint density of Y. Figure 2.2 plots the joint density of Y modelled using a Gaussian copula density with different margins. In Figure 2.2 (a), Y_1 follows a standard normal and Y_2 follows a gamma distribution with shape parameter 2 and scale parameter 3. The domain of Y_2 is strictly greater than zero and the mass is concentrated around 2/3 (the mean), reflecting the properties of the gamma margin. In contrast, in Figure 2.1 (a) most of the mass is around two standard deviations from the mean for both Y_1 and Y_2, reflecting the properties of the standard normal margins. For the same copula density, in Figure 2.2 (b), Y_1 follows a gamma distribution with shape parameter 2 and scale parameter 3 and Y_2 follows a gamma distribution with shape parameter 2 and scale parameter 4. In this figure, the joint density of Y is no longer symmetric because the margins for both Y_1 and Y_2 follow different distributions and neither margins are symmetric.



(a)

(b)

Figure 2.2: Contour and density plots of Y modelled using a bivariate Gaussian copula with a (a) standard normal (y_1) and gamma(2,3) (y_2) margin and zero correlation; (b) gamma(2,3) (y_1) and gamma(2,4) (y_2) margins and zero correlation

Many families of copulas exist to capture the different relationships between variables. The Archimedean copula family is a class of parametric copulas that can capture non-linear dependence between variables. For example, the Gumbel copula captures positive tail dependence and the Clayton copula captures negative tail dependence. Archimedean copulas only have one parameter irrespective of the number of margins. In contrast, the elliptical copula family, which includes the Gaussian copula, and the vine copula family, are more flexible parametric families that capture pairwise dependence between variables. In both these families, the number of parameters increases with the number of dimensions.

2.2 Gaussian Copulas

2.2.1 Definition

In practice, the Gaussian copula is the most popular choice of copula because of its simplicity and flexibility in capturing dependence in high dimensions. Consider the bivariate case with random variables Y_1 and Y_2. The Gaussian copula CDF is defined as

$$\mathbb{P}(Y_1 \leq y_1, Y_2 \leq y_2) = \Phi_\Lambda(\Phi^{-1}(u_1), \Phi^{-1}(u_2)), \tag{2.3}$$

where Φ_Λ is the multivariate normal CDF with mean zero and correlation matrix Λ, and $u_i = F_i(y_i)$, $i = 1, 2$ are uniform random numbers. The symbol Φ without a subscript denotes the univariate standard normal CDF.

2.2.2 Properties of a Gaussian Copula

The Gaussian copula captures the symmetric dependence of high-dimensional datasets through its correlation matrix. The correlation matrix is the (multivariate) parameter of a Gaussian copula. Each element of the matrix captures the correlation between a pair of $\Phi^{-1}(u)$. To build intuition, I explore the effect of different correlation matrices of a Gaussian copula on the joint distribution of Y. Here, Y is also assumed to be continuous as the discrete case is more difficult to illustrate. From Equation (2.3), the joint density of Y is

$$\begin{aligned} f(y_1, y_2) &= c(u_1, u_2) f(y_1) f(y_2) \\ &= \phi_\Lambda(\Phi^{-1}(u_1), \Phi^{-1}(u_2)) \frac{\partial \Phi^{-1}(u_1)}{\partial u_1} \frac{\partial \Phi^{-1}(u_2)}{\partial u_2} f(y_1) f(y_2), \end{aligned} \tag{2.4}$$

where ϕ_Λ is the multivariate normal probability density function with mean zero and correlation matrix Λ. The joint density of Y can be decomposed into a Gaussian copula density and the marginal densities of Y (Equation 2.4).

The Gaussian copula density is symmetric with the sign and magnitude of the correlation coefficient implying the orientation of the density and the extent that this relationship holds. Figure 2.3 (a) shows Gaussian copula densities with correlation coefficient $\rho = 0$, 0.5 and -0.9. For $\rho = 0$, the value of u_1 and u_2 are equally likely on $[0, 1]$ and additionally, u_1 and u_2 are independent of each other. For $\rho = 0.5$, the probability that both u_1 and u_2 are large is the same as the probability that both u_1 and u_2 are small. For $\rho = -0.9$, the probability that u_1 is large and u_2 is small is the same as the probability that u_2 is large and u_1 is small. These examples illustrate the symmetric dependence of the Gaussian copula for different correlation coefficients.

Multiplying the Gaussian copula densities with the marginal densities of Y and plotting the joint density of Y on the y_1 and y_2 axes yield a rescaled and reshaped contour plot. Figure 2.3 (b) plots the Gaussian copula densities with correlation coefficient 0, 0.5 and -0.9. Here, Y_1 follows a standard normal distribution and Y_2 follows a gamma distribution with shape parameter 2 and scale parameter 3. The mass of the joint density of Y is redistributed to reflect the marginal distributions. For example, y_2 ranges from zero to infinity, reflecting the gamma distribution. The joint distributions of Y associated with $\rho = 0.5$ and -0.9 are no longer symmetric as implied from their contour plots. Additionally, Figure (b) also shows that the orientation of the Gaussian copula densities are preserved, with a positive correlation coefficient associated with a positive slope and a negative correlation coefficient associated with a negative slope.

2.2.3 The Copula parameter to be estimated

Direct estimation of the correlation matrix is difficult because elements on the diagonal are restricted to one and elements on the off-diagonal are restricted to $[-1, 1]$. To reduce the number of restrictions, I rewrite the Gaussian copula as a Gaussian distribution with zero mean and a covariance matrix Σ. From Equation 2.3:

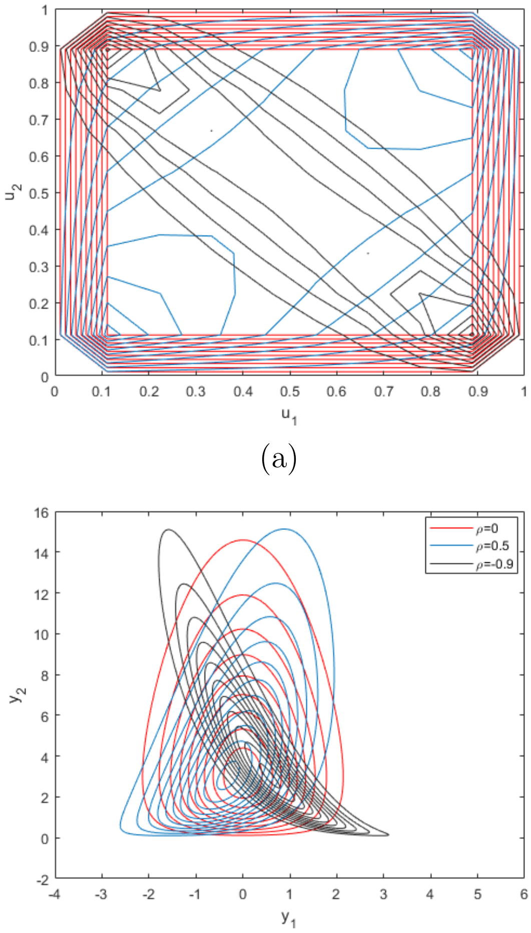
$$\begin{aligned} P(Y_1 \leq y_1, Y_2 \leq y_2) &= \Phi_\Lambda(\Phi^{-1}(u_1), \Phi^{-1}(u_2)) \\ &= \Phi_\Lambda(z_1^*, z_2^*) \\ &= \Phi_\Sigma(\sigma_{11} z_1^*, \sigma_{22} z_2^*) \\ &= \Phi_\Sigma(x_1, x_2), \end{aligned} \tag{2.5}$$

where $z_j^* = \Phi^{-1}(u_j)$, σ_{jj} is the jth diagonal entry of Σ, and Φ_Σ is a multivariate normal CDF with mean zero and covariance Σ satisfying $\Lambda = D\Sigma D$.[2] The diagonal matrix D has entries $1/\sigma_{jj}$ and the latent variable x_j is equal to $\sigma_{jj} z_j^*$.

Modelling the joint distribution of Y using a Gaussian copula does not mean Y is normally distributed. Instead x, which is a function of y, is normally distributed with mean zero and covariance matrix Σ (see Equation 2.5). Each element in Σ captures the bivariate relationship of a pair of x.

Despite fewer restrictions, modelling Σ for high-dimensional data is still difficult

[2] $\Lambda = D\Sigma D$ is a unique decomposition so there are no identification issues.



(a)

(b)

Figure 2.3: Contour plots of (a) a bivariate Gaussian copula density with correlation coefficient 0, 0.5 and -0.9; (b) the density of Y modelled using a Gaussian copula density with a standard normal (y_1) and gamma(2,3) (y_2) margin and correlation coefficient 0, 0.5 and -0.9.

because the number of unknown parameters is large. For example, for a copula with 37 margins it is necessary to estimate the $37 \times (37+1)/2 = 703$ parameters in Σ. Following Murray et al. (2013), I reduce the number of parameters to be estimated by assuming the covariance matrix associated with a Gaussian Copula has a factor structure. For J-dimensions:

$$\begin{aligned}\mathbb{P}(Y_1 \leq y_1, Y_2 \leq y_2, \ldots, Y_J \leq y_J) &= \Phi_{\Sigma}(x_1, x_2, \ldots, x_J) \\ &= \Phi_{\beta\beta^{\intercal}+I}(x_1, x_2, \ldots, x_J),\end{aligned}$$

where $\Sigma = \beta\beta^{\intercal} + I$, I is an identity matrix and β is a lower triangular matrix with positive diagonals and dimension $J \times r$ with $r \ll J$. The lower triangular restriction on β ensures that the matrix is well identified (Geweke and Zhou, 1996). See appendix B for more information on the identification of factor structures. Inference on Σ is through β since I is known.

To incorporate the lower triangular and positive diagonals restrictions, let $\theta = \mathrm{vech}(\beta)$ with log diagonal elements. The operation vech refers to excluding the zeros of a lower triangular matrix and stacking the columns of the matrix into a column vector. For all simulations and empirical applications presented in later chapters, I take the number of factors as $r = 2$ for simplicity. An alternative to fixing the number of factors is to choose r using log predictive Bayes factors (Kastner et al., 2017).

Chapter 3
Bayesian Inference

3.1 Posterior Distribution of the Copula Parameter

In Bayesian inference, interest is on the posterior distribution of the parameters θ. The standard form of the posterior is

$$\underbrace{p(\theta|y)}_{\text{posterior}} \propto \underbrace{p(y|\theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}}, \tag{3.1}$$

where the prior distribution $p(\theta)$ captures previous beliefs of where the parameter θ lies before data is observed and $p(y|\theta)$ is the likelihood. After the data is observed, the belief about θ is updated using Bayes formula.

Adding beliefs of θ before data has been observed allows the user to specify prior subject matter knowledge. This belief can be based on previous experimental outcomes, or used to specify restrictions on the parameter space. For example, if θ is a probability then the prior can specify that it should be between zero and one. When there is no previous knowledge of θ, a non-informative prior can be used.[3]

The contribution of the prior depends on the size of the dataset. When the dataset is large, the likelihood has more influence on the posterior distribution than the prior distribution, and when the dataset is small, the prior distribution may have more influence. Since this thesis explores copula modelling with large datasets, the likelihood will have more influence on the posterior distribution than the prior distribution in the applications considered.

3.2 Computing the Likelihood for discrete y

From Equation 3.1, the likelihood for 2-dimensions and N independent observations is

$$p(Y_1 = y_1, Y_2 = y_2|\theta) = \prod_{n=1}^{N} p(Y_{1,n} = y_1, Y_{2,n} = y_2|\theta).$$

[3] Maximising the posterior of θ when a non-informative prior is used yields the Maximum Likelihood estimate of θ.

The probability at a point when y is discrete is:

$$p(Y_{1,n} = y_1, Y_{2,n} = y_2|\theta) = [C(b_1, b_2|\theta) - C(b_1, a_2|\theta)] - [C(a_1, b_2|\theta) - C(a_1, a_2|\theta)],$$

where $a_j = F(y_j^-)$, $b_j = F(y_j)$ for $j = 1, 2$, and C is a copula CDF. In this 2-dimensional example, four copula evaluations are required to compute the probability at a point. For high-dimensional y, the likelihood becomes computationally expensive to evaluate as the probability at a point for J-dimensions, $p(Y_{1,n} = y_1, Y_{2,n} = y_2, \ldots, Y_{J,n} = y_J|\theta)$ requires 2^J copula evaluations. For example, when $J = 50$, 2^{50} evaluations of the copula CDF are required to compute the probability at a point, and a total of $2^{50} \times N$ evaluations of a copula are required to compute the likelihood. For discrete high-dimensional y, the likelihood of θ is in practice intractable because it is computationally too expensive to evaluate. Similarly, due to the intractability of the likelihood, using Maximum Likelihood to compute an estimate of θ is also infeasible.

Since the computation cost involved with computing the likelihood for high-dimensional y increases exponentially, this means that using the standard form of the posterior distribution should be avoided.[4] Instead, a cheaper alternative which involves using an estimate of the likelihood and inference on an augmented space θ, u can be used (Andrieu et al., 2009). The augmented posterior distribution is

$$\tilde{p}(\theta, u|y) \propto \widehat{p}_M(y|\theta, u)p(\theta, u),$$

where u is a vector of auxiliary variables and the likelihood is replaced by an unbiased estimate $\widehat{p}_M(y|\theta, u)$. The number of random numbers used to estimate the likelihood is denoted as M. Although inference is on the augmented posterior distribution, the posterior distribution of interest is obtained by integrating out u from $\tilde{p}(\theta, u|y)$ (Andrieu et al., 2009).

To estimate the augmented posterior distribution, I set the prior distribution of θ to be $N(0, b^{-1}I)$, where $b > 0$ is a scalar and I is an identity matrix (Murray et al., 2013) and I derive an unbiased estimate of the likelihood in Section 3.2.1. Chapters 4 and 5 detail two Bayesian methods, namely pseudo-marginal MCMC methods and VB methods, which can be used to estimate the parameters of a copula with discrete margins. pseudo-marginal methods yield exact results, but can be slow to converge. In contrast, VB methods provide approximate results to the exact posterior but are fast to implement.

[4] Incorrect results are obtained if the discrete nature of the data is ignored (Genest and Nešlehová, 2007).

3.2.1 An Unbiased Likelihood Estimate

Given the form of the augmented posterior distribution from Section 3.2, I derive an unbiased estimate of the likelihood, $\widehat{p}_M(y|\theta, u)$. Appendix A gives the proof. For J-dimensions and N independent observations,

$$\widehat{p}_M(y|\theta, u) = \prod_{n=1}^{N} \widehat{p}_M(Y_{1,n} = y_1, Y_{2,n} = y_2, \ldots, Y_{J,n} = y_J|\theta), \tag{3.2}$$

where M is the number of uniform random numbers used to estimate the likelihood, and the likelihood term is equal to

$$p(Y_{1,n} = y_1, \ldots, Y_{J,n} = y_J|\theta) = \int_{\sigma_{11}\Phi^{-1}(a_1)}^{\sigma_{11}\Phi^{-1}(b_1)} \cdots \int_{\sigma_{JJ}\Phi^{-1}(a_J)}^{\sigma_{JJ}\Phi^{-1}(b_J)} \phi_\Sigma(x)\, dx_1 \ldots\, dx_J,$$

where σ_{jj}, for $j = 1, \ldots, J$, corresponds to the diagonal elements of the covariance matrix Σ associated with the Gaussian copula, and $a_j = F_j(y_j^-)$ and $b_j = F_j(y_j)$ for $j = 1, \ldots, J$ are functions of the margins. Each likelihood term can be estimated unbiasedly using a modified version of Genz's algorithm (1992) or Botev's algorithm (2017) (see Appendix D for information on a modified version of the Genz (1992) algorithm). Similarly to Gunawan et al. (2019), an unbiased estimate of each likelihood term can be obtained using importance sampling with a uniform density proposal. I find that estimating J integrals to compute each likelihood term is cheaper than evaluating the copula 2^J times.

3.2.2 Constructing the Integral Bounds of an Unbiased Likelihood Term

The integral bounds of each likelihood term are functions of the margins of a copula, $F_j(y_j^-)$ and $F_j(y_j)$. In this thesis, the margins are estimated by their empirical CDFs for simplicity. The empirical CDF for the jth margin can be computed using

$$F_j(l) = p(Y_{\cdot j} \leq l) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}(y_{nj} \leq l),$$

where $Y_{\cdot j}$ is the jth column of the data matrix, l is the discrete value of interest, and $\mathbb{I}$ is an indicator function. The marginal distribution $F_j(y_j^-)$ refers to $F_j(y_j - 1)$.

Chapter 4
Exact Bayesian Inference

4.1 Markov Chain Monte Carlo

MCMC algorithms such as the Metropolis Hastings (MH) method (Chib and Greenberg, 1995) are popular choices to simulate from complex, non-standard distributions. The MH method was introduced by Metropolis et al. (1953) and generalised by Hastings (1970) to solve statistical problems. The aim of MH is to sample from the posterior distribution by constructing a Markov chain, which after convergence, forms a sample from the posterior distribution of interest.

Let θ^p be the proposed value, θ^c the current value, and $p(y|\theta)$ the density of y, $q_\Theta(\cdot;\theta^c)$ the proposal distribution and $p_\Theta(\theta)$ be the prior for θ. The acceptance probability based on the current and proposed value is denoted as $\alpha(\theta^c;\theta^p)$. The MH algorithm is implemented by the following steps:

1. Propose $\theta^p \sim q_\Theta(\theta;\theta^c)$.

2. Calculate the acceptance ratio

$$\alpha(\theta^c;\theta^p) = \min\left\{1, \frac{p(y|\theta^p)p_\Theta(\theta^p)q_\Theta(\theta^c;\theta^p)}{p(y|\theta^c)p_\Theta(\theta^c)q_\Theta(\theta^p;\theta^c)}\right\}. \tag{4.1}$$

3. Accept θ^p with probability α, otherwise take the current value as the next value in the chain.

4. Repeat $1-3$ until some stopping criteria is satisfied.

Prior to constructing a chain under the MH method, an arbitrary initial starting value is necessary. The first step in constructing the chain is to sample a value from a proposal distribution. Based on the current value (or the initial starting value if it is the first value in the chain), the proposed value is either accepted or rejected with probability α. If the proposed value is accepted, it becomes the next value in the chain. Otherwise the next value in the chain is the current value. The steps are repeated with the current value being the most recently accepted value until a stopping criteria is satisfied. It is straight forward to show the θ iterates are Markov.

The burn-in period is the set of iterates of the chain which is removed to exclude the influence of the initial starting value. The length of the burn-in period depends on how close the starting value is to the true posterior sample. If the initial starting value is close to values in the posterior sample then the burn-in period can be short. Determining the length of the burn-in period can be difficult. If the burn-in period is too short, there may be a risk of including values that are not from the posterior which would bias results. On the other hand, a burn-in period that is too long is inefficient and costly.

A related critical decision associated with MCMC is determining when to stop the algorithm. This decision entails using a stopping criteria to evaluate whether the chain has converged and whether the samples are representative of the posterior. After excluding the burn-in period, a chain that has converged will resemble stationary data. Convergence of the Markov chain depends on the proposal and how correlated the samples are. A large literature studying MCMC convergence diagnostics, such as polynomial time convergence bounds (Polson 1994), Markov minorisation conditions (Rosenthal 1993, 1995a, 1995b), 'thick felt-tip pen test' (Gelfand and Smith 1990), batching (Ripley 1987) exists. Most of these diagnostics, however, apply to particular problems (Cowles and Carlin, 1996). Instead, I use trace plots, which are plots of the chain, to show convergence. Trace plots provide graphical evidence of whether samples resemble stationary data, suggesting convergence. Trace plots can also be used to determine the length of the burn-in period. The left panel of Figure 4.1 is an example of a chain that may have converged and the right panel is an example that has not converged, as it has a time varying mean. There is a risk that the chain shown in the left panel of Figure 4.1 looks stationary because the MCMC chain is stuck at a local mode. Running the MCMC algorithm for longer or starting the algorithm at different initial parameter values and seeing if the chain converges to the same posterior distribution can help determine whether this is true.

Since the proposal distribution generates the samples used to construct the chain, it is important to choose a distribution which shares similar properties to the posterior. The closer the proposal is to the posterior, the more efficient the MH algorithm is in generating samples from the full posterior distribution. Properties of the proposal should include, the same or a larger support of the posterior. An example of a bad proposal is a Gamma proposal chosen to estimate a standard normal posterior. The posterior estimated using MH would be incorrect because only the positive half of the normal distribution would be recovered.

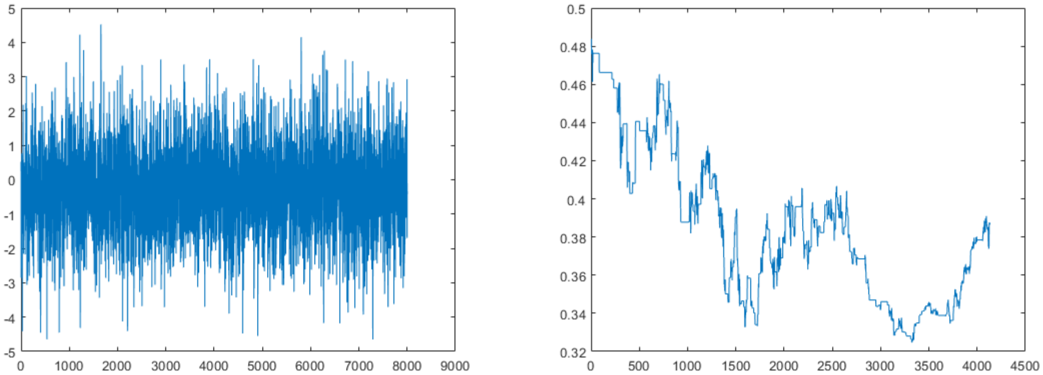


Figure 4.1: The left panel graph is an example of a chain that may have converged, while the right panel graph is an example of a chain that has not converged.

4.2 Pseudo-Marginal Methods

The standard pseudo-marginal (PM) method is an extension of the MH algorithm to tackle problems involving intractable likelihoods which can be estimated unbiasedly (Andrieu et al., 2009; Pitt et al., 2012). Instead of dealing with an intractable likelihood which may be difficult or impossible to evaluate, the likelihood is replaced with an unbiased estimate. An example application includes, generalised linear mixed models with random effects, where integrals are used to integrate out the random effects, which capture the dependence between multiple observations per individual (Tran et al., 2017). Andrieu et al. (2009) show that the sampler still targets the true posterior distribution.

For the standard PM method, let $\widehat{p}_M(y|\theta, u)$ be an unbiased estimate of the likelihood of θ and a set of random numbers, u. Let M be the number of random numbers used to estimate the likelihood. Here, $\widehat{p}_M(y|\theta, u)$ refers to the unbiased estimate of the likelihood derived in Section 3.2.1. Similarly to MH, let $q_\Theta(\cdot;\theta^c)$ be the proposal distribution for θ and $p_\Theta(\theta)$ be the prior specified in Section 3. Let $p_U(u)$ be the distribution of u used to construct the estimate of the likelihood, for example, uniform or standard normal. The following is an implementation of the standard PM method

1. Propose $\theta^p \sim q_\Theta(\theta;\theta^c)$ and $u^p \sim p_U(u)$.

2. Calculate the acceptance ratio

$$\begin{aligned}\alpha(\theta^c, u^c; \theta^p, u^p) &= \min\left\{1, \frac{\widehat{p}_M(y|\theta^p, u^p)p_\Theta(\theta^p)p_U(u^p)q_\Theta(\theta^c; \theta^p)p_U(u^c)}{\widehat{p}_M(y|\theta^c, u^c)p_\Theta(\theta^c)p_U(u^c)q_\Theta(\theta^p; \theta^c)p_U(u^p)}\right\} \\ &= \min\left\{1, \frac{\widehat{p}_M(y|\theta^p, u^p)p_\Theta(\theta^p)q_\Theta(\theta^c; \theta^p)}{\widehat{p}_M(y|\theta^c, u^c)p_\Theta(\theta^c)q_\Theta(\theta^p; \theta^c)}\right\}.\end{aligned} \tag{4.2}$$

3. Accept (θ^p, u^p) with probability α, otherwise take the current value as the next step in the chain.

4. Repeat $1 - 3$ until some stopping criteria is satisfied.

For the standard PM method, the variance of the log of the estimated likelihood increases linearly with the dimension of the data (Tran et al., 2016). A large variance in the log of the estimated likelihood may reduce the rate of acceptance, causing the chain to get 'stuck'. To reduce the chance of the chain getting stuck, the number of random samples used to estimate the likelihood needs to increase in proportion to the dimension of the data. However, for high-dimensional problems, such as those considered in this thesis, the number of random samples to compute the unbiased likelihood estimate will need to be very large.

Deligiannidis et al. (2018) propose the correlated PM method and Tran et al. (2016) propose the block PM method as extensions of the standard PM method to reduce the cost of estimating the log of the estimated likelihood. Both methods involve adding correlation to the random numbers used to estimate the likelihood at the proposed and the current values of θ. A consequence is that the variance of the log of the estimated likelihood can be larger than the standard PM method without the chain becoming stuck. A larger variance in the log of the estimated likelihood implies that a smaller number of random numbers relative to the standard PM method can be used to compute the likelihood.

The correlated PM method is a more general version of the block PM method and relies on generating a full set of random numbers to estimate the likelihood at each iteration. Rather than updating the full set of random numbers, in the block PM method the random numbers, u, are split into G blocks and one of the G blocks is updated at random. By only updating a block of u and leaving $G-1$ blocks fixed, the current and proposed values of the likelihood are correlated, reducing the variance of $\log\widehat{p}_M(y|\theta^p, u^p) - \log\widehat{p}_M(y|\theta^c, u^c)$. Hence, the proposed value is more likely to be accepted. In addition, as only a block is updated at a time, the block PM method requires less CPU time than the standard PM and correlated PM methods at each iteration.

The algorithm of block PM is the same as the standard PM but with step 2 in Equation 4.2 replaced with

$$\alpha(\theta^c, u^c; \theta^p, u^p) = \min\left\{1, \frac{\widehat{p}_M(y|\theta^p, u^c_{(1)}, u^c_{(2)}, ..., u^c_{(k-1)}, u^p_{(k)}, u^c_{(k+1)}, ..., u^c_{(G)})p_\Theta(\theta^p)q_\Theta(\theta^c;\theta^p)}{\widehat{p}_M(y|\theta^c, u^c_{(1)}, u^c_{(2)}, ..., u^c_{(k-1)}, u^c_{(k)}, u^c_{(k+1)}, ..., u^c_{(G)})p_\Theta(\theta^c)q_\Theta(\theta^p;\theta^c)}\right\}, \tag{4.3}$$

where $u^p_{(k)}$ denotes the updated block of random numbers.

The smaller the G the larger the number of random numbers that are required to estimate the likelihood to reduce the chance of the MCMC chain getting stuck. Overall, the number of blocks, G, should be selected such that the chance of each block being updated is not too small given the total number of MCMC iterations, to ensure that the parameter space is explored (Tran et al., 2016).

Gunawan et al. (2019) show that the block PM method performs better than the correlated PM method in all simulations for VBIL applied to Archimedean copulas in terms of computation time and integrated autocorrelation time (IACT). IACT indicates the number of samples to be drawn until an independent sample is obtained. Therefore, the block PM method will be used as a benchmark to evaluate the accuracy and efficiency of VB methods in this thesis.

4.2.1 Choosing a Proposal Distribution for θ

From Equation 4.2, let $q_\Theta(\theta; \theta^c)$ be a random walk Gaussian distribution with an adaptive covariance matrix (Garthwaite et al., 2016). I use the random walk proposal

$$\theta \sim N(\theta^c, a^2 A),$$

where θ^c is the previously accepted value of θ (or in the first iteration, an arbitrary starting point), a is an adaptive scale parameter, and A is an adaptive covariance matrix. At the tth iteration, $A_t = \widehat{\Sigma}_t + \frac{a_t^2}{t} I$, where $\widehat{\Sigma}_t$ is the sample covariance matrix constructed using the past iterations of the MCMC chain and I is an identity matrix, which ensures the adaptive covariance matrix is positive definite.

The proposal $q_\Theta(\theta; \theta^c)$ is useful when properties such as the mode[5], shape, and domain of the posterior distribution are unknown, as the domain on the real line ensures that the support of the posterior is well covered. Additionally, the

[5] Temperature annealing, a Sequential Monte Carlo method, can be used to determine the number of modes at a high level (Gill and Casella, 2004).

adaptive mean and covariance matrix allow the proposal distribution to develop with the constructed chain (Garthwaite et al., 2016). The scale parameter α is tuned based on the Robbins-Monro process such that when the proposed value of θ is rejected, the scale parameter of the covariance matrix decreases and the mean of the covariance matrix remains at the current value. With a smaller variance and a mean centred at the previously accepted value, this increases the chance that the proposal distribution will sample a value that is accepted in the next iteration. On the other hand, when the proposed value is accepted, the variance of the proposal distribution increases, allowing the algorithm to explore the support of the posterior.

Chapter 5
Approximate Bayesian Inference

5.1 Background of variational Bayes

Direct inference from the exact posterior distribution is often costly in both time and computational effort. Variational Bayes (VB) is a method which approximates the posterior distribution with a simpler, tractable distribution and uses optimisation to find the parameters of the approximating distribution.

The first step in VB is to choose a class of variational distributions to approximate the posterior. The chosen class of variational distributions should reflect properties of the true posterior distribution if this information is available. For example, a Gamma distribution or a Gaussian distribution with log transformed variables may be good variational distributions (approximating distributions) if the domain of the true posterior is positive.

After choosing a class of variational distributions, optimisation can be used to select the optimal variational distribution in that class. Suppose the class of variational distributions is the Gaussian distribution and let it be represented by the purple set in Figure 5.1. Then each element '$q_\lambda(\theta)$' in the purple set represents a Gaussian distribution with a different parameter λ (mean and variance). The divergence between the variational distribution $q_\lambda(\theta)$ and the true posterior $p(\theta|y)$

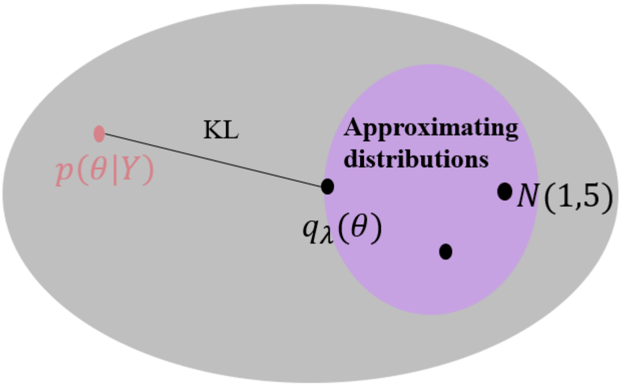


Figure 5.1: Graphical representation of VB approximation.

can be summarised using the Kullback-Leiber (KL) divergence. The KL divergence measures the quality of the variational distribution as an approximation of the true posterior and is defined as

$$\mathrm{KL}(q_\lambda(\theta)||p(\theta|y)) := \int \log\bigg(\frac{q_\lambda(\theta)}{p(\theta|y)}\bigg) q_\lambda(\theta)\, d\theta. \tag{5.1}$$

This equation can be rewritten as

$$\log(p(y)) = \mathrm{KL}(q_\lambda(\theta)||p(\theta|y)) + \int \log\bigg(\frac{p(y|\theta)p(\theta)}{q_\lambda(\theta)}\bigg) q_\lambda(\theta)\, d\theta. \tag{5.2}$$

Therefore, the log of the marginal likelihood $p(y)$ is the sum of the KL divergence and the lower bound. When the variational distribution is a good approximation to the true posterior, the KL divergence is small and the lower bound is close to the log of the marginal likelihood. For example, when $q_\lambda(\theta)$ is equal to the true posterior the KL divergence is equal to zero and the lower bound is equal to the log of the marginal likelihood. Therefore, the optimal variational distribution is the distribution that minimises the KL divergence with respect to the variational parameter λ. Since $p(y)$ is not a function of λ, minimising the KL divergence is equivalent to maximising the lower bound with respect to λ.

Stochastic gradient ascent (SGA) is a method which is often used to find the variational parameter that maximises the lower bound. It involves iteratively updating the variational parameter using gradient information.[6] At the tth iteration,

$$\lambda_{t+1} = \lambda_t + \rho_t \nabla_\lambda \mathrm{LB}(\lambda_t), \tag{5.3}$$

where $\nabla_\lambda \mathrm{LB}(\lambda) = \mathbb{E}_{q_\lambda}[\big(\log(p(y|\theta)p(\theta)) - \log q_\lambda(\theta)\big)\nabla_\lambda \log q_\lambda(\theta)]$ is the gradient of the lower bound with respect to λ, and $\mathbb{E}_{q_\lambda}[.]$ denotes the expectation with respect to $q_\lambda(\theta)$. The current value of the variational parameter is represented as λ_t, whereas the most recently updated is represented as λ_{t+1}. The learning rate ρ_t controls how fast the algorithm proceeds to find the optimal choice of λ and the gradient of the lower bound $\nabla_\lambda \mathrm{LB}(\lambda_t)$ captures the direction of movement. Chapter 6 discusses the learning rate and the gradient of the lower bound.

[6] SGA does not guarantee convergence to a global optimum. One method to increase the chance of obtaining a global optimum is to start the algorithm at different points to see if the algorithm converges to the same value. However, this is very costly as it involves re-running the VB algorithm many times.

5.2 Variational Bayes methods for Copulas with discrete Margins

Two VB methods in the literature have been proposed to estimate copulas with discrete margins. Loaiza-Maya and Smith (2019) propose a VB method for the D-Vine copula and Gunawan et al. (2019) for Archimedean copulas. The main contribution of this thesis is to extend both of these methods to the Gaussian copula and determine whether VB methods provide a good approximation to the true posterior of the copula parameters. The rest of this chapter provides the intuition of VBIL and the specifications of the Gaussian variational distribution chosen to approximate the posterior. I also present the VBIL algorithm. Appendix J details the derivations and specifications of VBDA for the Gaussian copula with discrete margins. Section 5.4 introduces an extension of the variational distribution to a non-Gaussian posterior approximation and Appendix I provides examples. Chapter 6 evaluates by simulation the accuracy and efficiency of VBIL and VBDA.

5.3 Variational Bayes Intractable Likelihood (VBIL)

Similarly to PM methods, VBIL is a VB method to tackle problems involving intractable likelihoods which can be estimated unbiasedly. Following Pitt et al. (2012), Tran et al. (2017) define u as a J-dimensional set of random variables used to construct an unbiased estimate of the likelihood $p(y|\theta)$. An unbiased estimate of the likelihood, $\widehat{p}_M(y|\theta, u)$, can be obtained using importance sampling with M uniform random numbers. Let the error of the log of the unbiased likelihood estimate be a scalar, z, where

$$\begin{aligned} z &= \log \widehat{p}_M(y|\theta, u) - \log p(y|\theta) \\ \text{so that } \widehat{p}_M(y|\theta, u) &= e^z p(y|\theta). \end{aligned} \tag{5.4}$$

Also let z be a function of u and $g_M(z|\theta)$ be the density of z. Then,
$\int e^z g_M(z|\theta)\, dz = 1$ due to unbiasedness.

Using Equation 5.4 and the unbiased estimate of the likelihood derived in Section 3.2, the augmented posterior can be written as:

$$\begin{aligned} \tilde{p}(\theta, u|y) &= \widehat{p}_M(y|\theta, u) g_M(z|\theta) p(\theta)/p(y) \\ &= e^z p(y|\theta) g_M(z|\theta) p(\theta)/p(y) \\ &= p(\theta|y) g_M(z|\theta) e^z. \end{aligned} \tag{5.5}$$

Finally, Equation 5.5 can be used to show that the posterior distribution of interest can be obtained by integrating out z from the augmented posterior:

$$\begin{aligned}\int \tilde{p}(\theta, u|y)\, dz &= \int p(\theta|y) g_M(z|\theta) e^z\, dz \\ &= p(\theta|y),\end{aligned}$$

as $\int e^z g_M(z|\theta)\, dz = 1$. See Pitt et al. (2012) for the proof.

Based on the form of $\tilde{p}(\theta, u|y)$ in Equation 5.5, Tran et al. (2017) and Gunawan et al. (2019) define the variational distribution as $q_\lambda(\theta, z) = q_\lambda(\theta) g_M(z|\theta)$. The variational distribution of $z|\theta$ is set equal to the true distribution $g_M(z|\theta)$.

5.3.1 Choosing the Variational Distribution for the Posterior of θ

Let $q_\lambda(\theta)$ be a multivariate Gaussian distribution. I take the variational distribution for θ as

$$q_\lambda(\theta) = N(\eta, \Gamma\Gamma^\intercal + D^2),$$

where Γ is an unrestricted lower triangular matrix with dimensions $R \times \tilde{r}$ with $\tilde{r} \ll R$, D is a diagonal matrix, and $\lambda = \{\eta^\intercal, \text{vech}(\Gamma)^\intercal, \text{diag}(D)^\intercal\}^\intercal$. For a copula with J margins, I assumed that the factor structure covariance matrix associated with the Gaussian copula has two factors (Section 2.2.3). This means that the dimension of θ is $R \times 1$, where $R = J + (J - 1)$, after accounting for the vech structure.

The Gaussian variational distribution is chosen because it is easy to simulate from and is tractable. However, similarly to the covariance matrix associated with the Gaussian copula, the dimension of covariance matrix of the variational distribution is large. Following Ong et al. (2018b) and Tran et al. (2019), I impose a factor structure on the covariance matrix to reduce the number of parameters. Without a factor structure, the number of parameters in the variational distribution when $J = 37$ is 73 for the mean and $73 \times (73 + 1)/2 = 2,701$ distinct parameters for the covariance matrix. In total, it is necessary to estimate $2,774$ parameters for the variational distribution given an unrestricted covariance matrix. In contrast, when a factor structure with one factor is assumed, the number of distinct parameters in the covariance matrix of the variational distribution reduces to 146. Thus, a total of 219 parameters need to be estimated for a one factor model.

For simplicity, I let the number of factors $\tilde{r} = 1$ for all simulations and empirical applications in this thesis. Results for the block PM method show that $\tilde{r} = 1$ is sufficient for estimating the posterior distribution of the copula parameter. However, instead of assuming a fixed value of $\tilde{r}$, the number of factors can be chosen by running the VBIL algorithm on increasing values of $\tilde{r}$ and stopping when the results stabilise (Ong et al., 2018b).

As an alternative to assuming a factor structure, I model the Cholesky factor of the inverse covariance matrix of the variational distribution. Here, the proposal is given by $q_{\lambda_1}(\theta) = N(\eta_1, \delta^{-\intercal}\delta^{-1})$, where $\lambda_1 = (\eta_1^\intercal, \text{vech}(\delta)^\intercal)^\intercal$, η_1 is the mean, and δ is the Cholesky factor of the inverse covariance matrix with dimensions $R \times R$. Similarly to the factor structure covariance matrix associated with the Gaussian copula, the positive lower triangular restriction is imposed for identification. The Cholesky factor of the inverse covariance matrix has more parameters then the factor structure and can be used as a robustness check to see if more parameters will lead to increased flexibility and an improved fit of the posterior. See Appendix K for more information on the choice of factorisation of Σ.

5.3.2 KL Divergence for VBIL

Given $q_\lambda(\theta, z)$, the variational parameter, λ, can be found by maximising the lower bound using SGA. Let $\tilde{p}(\theta, u|y) \propto h(\theta, u)g_M(z|\theta)$, where $h(\theta, u) = \widehat{p}_M(y|\theta, u)p(\theta)$. The lower bound is given by

$$\text{LB}(\lambda) = \iint \log\biggl(\frac{h(\theta, u)g_M(z|\theta)}{q_\lambda(\theta, z)}\biggr) q_\lambda(\theta, z)\, dz\, d\theta$$

and the corresponding gradient with respect to λ is:

$$\nabla_\lambda \text{LB}(\lambda) = \mathbb{E}_{\theta,z|\theta}\bigl[\bigl(\log h(\theta, u) - \log q_\lambda(\theta)\bigr)\nabla_\lambda \log q_\lambda(\theta)\bigr].$$

See Appendix E for the proof.

5.3.3 VBIL Algorithm

The VBIL algorithm is implemented in two main steps. Step one calculates the control variate $c^{(t)}$ to reduce the variance of the gradient of the lower bound (Tran et al. 2017). Without the control variate, the gradient becomes noisy and the variational parameter λ may not converge. Step two calculates the gradient estimate of the lower bound and updates λ with the new gradient information. The learning rate is denoted as ρ_t and is discussed in Chapter 6.

I now outline the algorithm to implement VBIL with the definition of $q_\lambda(\theta)$ and $\widehat{p}_M(y|\theta, u)$ in Section 5.3. Chapter 3 specifies the prior and an unbiased estimate of the likelihood and Appendix C gives the derivatives of $\nabla_\lambda \text{log} q_\lambda(\theta)$.

Initialise $\lambda^{(0)}$ and let S be the number of samples used to estimate the gradient.

1. Initialisation: Set $t = 0$
   (a) Generate $\theta_s^{(t)} \sim q_\lambda(\theta)$ and $z_s^{(t)} \sim g(z|\theta)$, for $s = 1, 2, ..., S$.
   (b) Denote $h(\theta, u) = \widehat{p}_M(y|\theta, u)p(\theta)$ and set

$$c^{(t)} = \frac{\text{Cov}\big((\text{log}h(\theta, u) - \text{log}q_\lambda(\theta))\nabla_\lambda \text{log}q_\lambda(\theta), \nabla_\lambda \text{log}q_\lambda(\theta)\big)}{\text{Var}\big(\nabla_\lambda \text{log}q_\lambda(\theta)\big)}$$

   where Cov(.) and Var(.) are sample estimates of the covariance and variance based on the S samples from step (a).

2. Repeat the following until a stopping criterion is satisfied.
   (a) Set $t = t+1$ and generate $\theta_s^{(t)} \sim q_\lambda(\theta)$ and $z_s^{(t)} \sim g(z|\theta)$, for $s = 1, 2, ..., S$.
   (b) Estimate the gradient

$$\nabla_\lambda \widehat{\text{LB}}(\lambda)^{(t)} = \frac{1}{S}\sum_{s=1}^{S}(\text{log}h(\theta_s^{(t)}, u_s^{(t)}) - \text{log}q_\lambda(\theta_s^{(t)}) - c^{(t-1)})\nabla_\lambda \text{log}q_\lambda(\theta_s^{(t)})$$

   (c) Estimate the control variate $c^{(t)}$ as in step 1(b)
   (d) Update the variational parameter λ by

$$\lambda^{(t+1)} = \lambda^{(t)} + \rho_t \nabla_\lambda \widehat{\text{LB}}(\lambda)^{(t)}.$$

5.4 An Extension: Transformation to a Non-Gaussian Posterior Approximation

A Gaussian distribution provides a good variational approximation to the posterior when (1) the marginal posterior distributions of θ follow a Gaussian distribution - have zero skewness and kurtosis equal to three, and (2) the dependence of θ_i for $i = 1, \ldots, R$ are linearly related. Conversely, a Gaussian distribution may not be the most optimal approximation when either the marginal posterior distributions are not all univariate Gaussian even though they are linearly related, or, when the marginal posterior distributions are all individually Gaussian but their dependence is non-linear.

Smith et al. (2019) show how to relax these normality assumptions when using a Gaussian variational distribution. Instead of modelling the copula parameter θ directly using a Gaussian variational distribution, the Gaussian variational distribution is used to model ψ, a one-to-one transformation of the copula parameters. Each parameter can be transformed as $\psi_i = t_{\gamma_i}(\theta_i)$, for $i = 1, ..., R$, where ψ follows a multivariate Gaussian distribution. The corresponding variational distribution of the copula parameters is

$$q_\lambda(\theta) = p(\psi; \pi) \prod_{i=1}^{R} t'_{\gamma_i}(\theta_i),$$

where $p(\psi; \pi)$ is the joint density function of the transformed parameters and is Gaussian with parameter π. As a result of the transformations, the marginal posterior distributions of the copula parameters are not necessarily univariate Gaussian.

Smith et al. (2019) propose two transformations t_γ: the Yeo and Johnson-transformation (YJ) (Yeo and Johnson, 2000) and the G&H-transformation (Tukey, 1977). In this thesis, I consider the YJ-transformation, which adds skewness into the Gaussian variational distribution. For $0 < \gamma < 2$, the YJ-transformation is

$$t_\gamma(\theta) = \begin{cases} -\frac{(-\theta+1)^{2-\gamma}-1}{2-\gamma} \text{ if } \theta < 0 \\ \frac{(\theta+1)^{\gamma}-1}{\gamma} \text{ if } \theta \geq 0. \end{cases}$$

Appendix I discusses examples of the YJ-transformation used to estimate the parameters of a Gaussian copula with discrete margins.

Chapter 6
Simulation Studies of VBIL, VBDA and Block PM

This chapter evaluates the performance of VBIL and VBDA relative to the block PM method to estimate the parameters of a Gaussian copula with discrete margins for high-dimensional datasets. Section 6.1 shows how to simulate discrete datasets from a Gaussian copula. Section 6.2 evaluates the accuracy and efficiency of VBIL against the block PM method through ordinal and Bernoulli simulated datasets of increasing dimensions up to 37 dimensions and 1200 observations, to evaluate the performance of VBIL in high dimensions. Application to ordinal datasets is an extension allowing the use of VBIL to analyse datasets such as HILDA. Section 6.3 compares the performance of VBIL and VBDA against the block PM method for Gaussian copulas with discrete margins. Section 6.4 and 6.5 discusses learning rates and the natural gradient, which are two of the most popular approaches to speed up convergence, in an application of VBIL. In all simulations, block PM is regarded as the gold standard and is used to compute the exact posterior.

6.1 Construction of the Simulated Dataset

Let Y be simulated data with dimension $N \times J$, where N is the number of observations and J is the number of dimensions.

The data is generated using steps $1 - 3$

1. Construct a lower triangular matrix with positive diagonal elements and dimension $J \times r$, $r \ll J$. The covariance matrix associated with the Gaussian copula is equal to $\Sigma = \beta\beta^{\intercal} + I$, where I is an identity matrix.

2. Simulate the N rows of Y^*, a latent matrix, where $Y^*_{n\cdot} \sim N(0, \Sigma)$ is the nth row of the latent matrix. The latent matrix Y^* is used to construct the data matrix Y.

3. For each column of Y^*, define a set of latent cut-offs. Use the cut-offs to convert the continuous variables Y^* into discrete variables, corresponding to the columns of Y. For example, in the Bernoulli case, if values in $Y^*_{\cdot j}$ are greater than 0.5, the corresponding values of $Y_{\cdot j}$ are set equal to 1 and are

set to 0 otherwise. For the ordinal case, if values in $Y^*_{\cdot j}$ are less than or equal to 0.2, the corresponding values of $Y_{\cdot j}$ are set equal to 1. If values in $Y^*_{\cdot j}$ are between 0.2 and 0.7, the corresponding values of $Y_{\cdot j}$ are set to 2, and are set to 3 otherwise.

By construction, the rows of the data matrix are independent while the columns are dependent. This reflects most cross-sectional datasets. The simulated dataset described above assumes that the number of factors in the covariance matrix associated with the Gaussian copula is known. I also consider simulating from a full covariance matrix. Here, I replace the covariance matrix in step 1 with a matrix simulated using a Wishart distribution. The results are similar to those involving a simulated dataset using a factor structure and therefore, are not included in this thesis. Figure 6.1 shows plots of example Bernoulli and ordinal datasets.

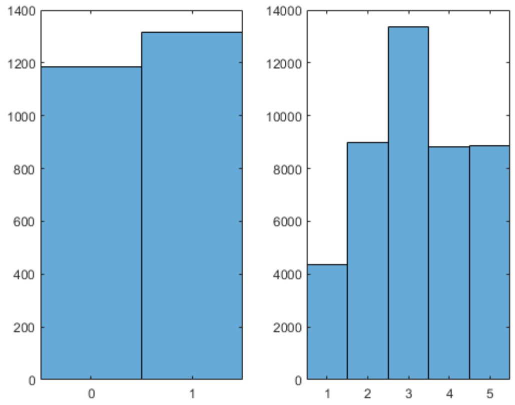


Figure 6.1: Bernoulli and ordinal Data Example

6.2 Simulation Studies: Ordinal and Bernoulli Data

This section studies the efficiency and accuracy of VBIL for the Gaussian copula against block PM across varying dimensions for both ordinal and Bernoulli data. I ran each of the VBIL simulations on a 20-core high performance computer cluster operated by the National Computational Infrastructure (NCI) using MATLAB code.

In the ordinal simulations, I consider the dimensions $J = 10$ and $N = 250$, $J = 25$ and $N = 250$, and $J = 37$ and $N = 1200$. The last simulation provides supporting evidence that VBIL for Gaussian copulas can handle the dimensions of the HILDA survey questions considered in Chapter 7. This combination of data and parameters are much larger than that considered by both Gunawan et al. (2019) and Loaiza-Maya and Smith (2019).

In the Bernoulli simulations, I consider the dimensions $J = 5$ and $N = 250$, $J = 5$ and $N = 1000$, and $J = 10$ and $N = 250$. Since Bernoulli data contain less information than ordinal or categorical data, an estimation method that works well for Bernoulli data should also work well for ordinal or categorical data (Gunawan et al., 2019; Loaiza-Maya and Smith, 2019). For this reason, I also expect that the Bernoulli simulations will perform worse than the ordinal simulations.

Chapter 3 specifies the prior and an estimate of the likelihood of θ for both VBIL and block PM. For VBIL, I set the initial parameter values of η to be a R-dimensional vector of 0.5, Γ to be a R-dimensional vector of 0.1, and $D = 0.2I_{R\times R}$. For simplicity, the marginal distributions of the copula are computed empirically (as discussed in Section 3.2.2).

In the VB algorithm, S specifies the number of samples used to estimate the gradient of the lower bound at each iteration (Section 5.3.3). A high value of S reduces the variance of the estimates of the gradient yielding faster convergence. A high value of S is also associated with increased computation time. In all simulations, I set S to either 50 or 100. The chosen S is relatively small compared to those used by Gunawan et al. (2019) because Genz's algorithm (1992) provides an accurate estimate of the likelihood even with a small S. In Genz's algorithm (1992), M is an input specifying the number of uniform random variables used to estimate the likelihood at each iteration. Figure 6.2 shows the effect of changing M on the log of the estimated likelihood (evaluated at the initial values of λ) for dimensions $J = 25$ and $N = 250$. The figure indicates that only a small number of random samples is required as the variance is relatively small after 20 samples. I set M to 20 for most simulations. Using the same specification as Genz's algorithm (1992), I also use Botev's algorithm (2017) to estimate the likelihood. I find that this algorithm produces similar results but is much slower as it first needs to solve an optimisation problem to calibrate the importance sampler. For this reason, I do not include the results from Botev's algorithm (2017).

Both VBIL and block PM required more clock time as the dimensions increased. In some of the block PM ordinal and Bernoulli applications, the chain did not converge within 48 hours. See Appendix H.1 for example trace plots of block PM starting at the same initial parameter values as VBIL. NCI caps the time a job is run to 48 hours making it infeasible to run the chain for longer. One alternative is to run several chains and combine the chains together to get a better estimate. A more efficient alternative, is to change the starting point of the block PM algorithm to be closer to the posterior mean. I chose the latter option.

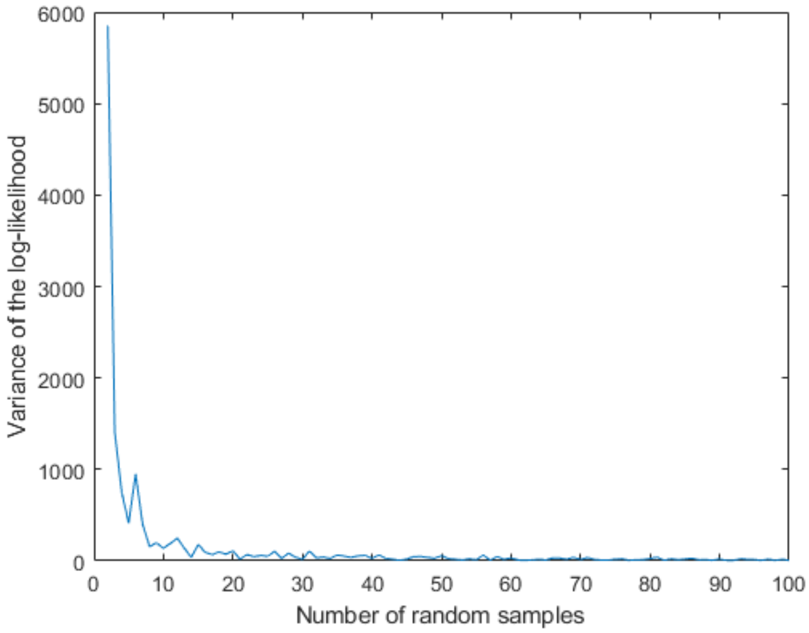


Figure 6.2: Effect of the number of random samples, M, in Genz's algorithm (1992) for dimensions $J = 25$ and $N = 250$.

Here, I chose the number of MCMC iterations for each simulation based on the dimension of the data and on trace plots. For $J < 25$, I found that running 200,000 MCMC iterations was sufficient. For $J \geq 25$, I set the initial parameter values of block PM equal to the converged VB estimates and ran 220,000 MCMC iterations. In all simulations, the first 20,000 MCMC iterations formed the burn-in period.

In both the ordinal and Bernoulli case, VBIL led to significant reductions in computation time compared to block PM for all simulations. Reductions of around 40 to 70 times were observed. For example, for ordinal data with dimensions $J = 37$ and $N = 1200$, VBIL took 51.80 minutes, while block PM took 2078.20 minutes. Convergence of VBIL in the Bernoulli case was a little slower than the ordinal case. In the Bernoulli case, convergence of VBIL for dimensions $N = 10$ and $J = 250$ was quicker than for $N = 5$ and $J = 250$. The increased time to convergence is likely to reflect the randomly sampled data.

I determine the accuracy of VBIL by comparing the marginal β posteriors estimated using VBIL against the marginal β posteriors estimated using block PM. If VBIL estimates the posterior distributions of the copula parameters perfectly, the kernel densities of VBIL should overlap with those of block PM. Figure 6.3 (a) shows that the kernel densities of block PM and VBIL overlap almost perfectly, implying that

Table 6.1: Ordinal Data Simulations

Dimension	Method	Specifications	Clock Time (mins)
J=10	block PM	-	252.00
N=250	VBIL	S = 50 M=20	3.60
J=25	block PM	-	429.60
N=250	VBIL	S = 50 M=20	7.80
J=37	block PM	-	2078.20
N=1200	VBIL	S = 100 M=30	51.80

Table 6.2: Bernoulli Data Simulations

Dimension	Method	Specifications	Clock Time (mins)
J=5	block PM	-	168.28
N=250	VBIL	S = 100 M=20	4.20
J=5	block PM	-	385.35
N=1000	VBIL	S = 100 M=20	17.25
J=10	block PM	-	250.24
N=250	VBIL	S = 50 M=30	3.71

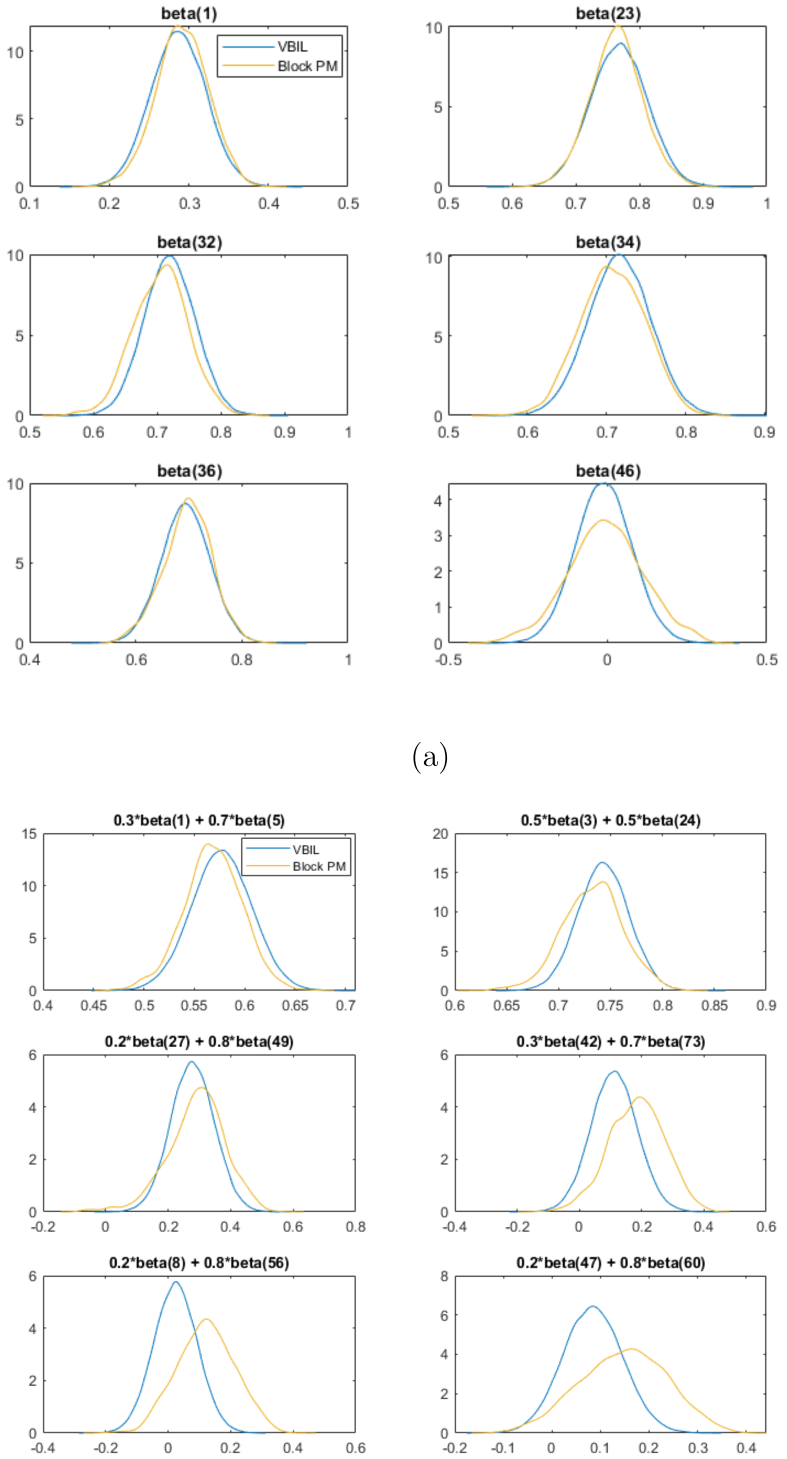
VBIL estimates the copula parameters well.[7] I find that VBIL is more accurate in the ordinal case than in the Bernoulli case. One reason is that convergence improves as the data becomes more continuous.[8] I also determine the accuracy of the marginal posteriors through a mean and standard deviation plot of β. If VBIL estimates the posterior distributions of the copula parameters perfectly, all β estimates should lie on the 45 degree line. Figure 6.6 (a) shows that almost all the means of β lie on the 45 degree line. Comparing the standard deviations show that some of the βs obtained using VBIL have a smaller posterior variance than block PM. A number of authors including Blei et al. (2017), Tan and Nott (2018) and Loaiza-Maya and Smith (2019) report similar underestimation of the posterior variance when using VB methods. The choice of variational distribution and the number of factors chosen for the factor structure of the variational distribution may be reasons for underestimating the posterior variance. All the examples in the thesis are representative of other dimensions.

The large dimension of the data makes it impossible to graph the joint distribution of the βs. Instead, I use linear combinations of the βs to test whether the variational distribution is able to capture joint dependence. Even though the elements of

[7] In the Bernoulli case, assuming a lower triangular structure and an unrestricted β results in better posterior estimation than when a positive lower triangular restriction on β is assumed. Despite this, the lower bounds of both sign restricted and unrestricted β converge to the same value. See Appendix I for more information.

[8] Professor Gael Martin suggested this during the 12th International Conference on Monte Carlo Methods and Applications on the paper by Loaiza-Maya and Smith (2019).

β follow Gaussian distributions, their linear combinations may not. Figure 6.3 (b) shows six plots of the linear combinations of β. In the first three plots, the VBIL linear combinations fit the MCMC linear combinations almost perfectly. For the other three plots, the worst fitted marginals were chosen. Even in these plots, the VBIL kernel densities largely overlap the densities corresponding to block PM, implying that the Gaussian variational distribution is a good choice for these simulations. Overall, the results of both the marginal distributions and the linear combinations show that the joint distribution of the βs resembles a multivariate Gaussian distribution. This provides evidence that VBIL for Gaussian copulas with discrete margins is an efficient alternative to MCMC and does not require much compromise over accuracy.



(a)

(b)

Figure 6.3: Plots of the Gaussian copula parameters based on Bernoulli data simulations with dimensions $J = 37$ and $N = 1200$: (a) Kernel densities of VBIL and block PM; (b) Kernel densities of VBIL and block PM for linear combinations of the parameters.

6.3 Comparing VBIL and VBDA

Using the same specifications for VBIL as in the simulations above, I set the prior of θ and the parameters of VBDA where possible equal to those in VBIL. As VBIL does not require estimating $q_{\lambda_\beta}(u) = N(\eta_\beta, (L_\beta L_\beta^\intercal)^{-1})$, I let η_β be a vector of 0.5 and the Cholesky factor, L_β, be a one banded matrix with non-zero elements set to 1.[9] See Appendix J for more information on VBDA.

Ordinal and Bernoulli data simulations with dimensions $J = 10$ and $N = 250$ were run to compare VBDA, VBIL and block PM. In both the simulations considered, VBDA requires more time to converge than block PM and VBIL. Although, after convergence, both VB methods provide good approximations to the true posterior as the kernel densities of β are largely overlapping. For ordinal data with dimensions $J = 10$ and $N = 250$, VBIL took 3.60 minutes, VBDA took 717.17 minutes, while block PM took 252.00 minutes. For Bernoulli data with dimensions $J = 10$ and $N = 250$, VBIL took 3.71 minutes, VBDA took 802.50 minutes, while block PM took 250.24 minutes. For VBDA, the time taken also reflects a smaller number of cores used. Following Loaiza-Maya and Smith (2019), I used 8 cores. A larger number of cores did not lead to a reduction in computation time, as unlike VBIL, calculation of the likelihood does not require as much parallel computing.

VBDA takes longer to run compared to VBIL because the number of unknown parameters is larger. The variational distribution $q_{\lambda_\beta}(u)$ is estimated in VBDA but not in VBIL because it is integrated out. For dimensions $J = 10$ and $N = 250$, there are 7,307 unknown parameters in VBDA compared to 57 parameters in VBIL. The extra parameters in VBDA relate to η_β, a vector of dimension $(J \times N)$, and L_β, a matrix of dimension $(J \times N) \times (J \times N)$, with $(2J - 1) \times N$ non-zero elements (corresponding to N band one lower triangular Cholesky factors).

Figure 6.4 shows that both the VBIL and VBDA kernel densities of β are largely centred on the true posterior mean, although the VBDA kernel densities underestimate the variance of the posterior by more than the VBIL kernel densities. These figures also show the Bernoulli simulation with dimensions $J = 10$ and $N = 250$, the variance of the β posteriors is underestimated for most βs in VBDA, whereas underestimation is less severe in VBIL. In contrast, Figure F.2 shows little underestimation is observed in the ordinal data simulation for both VBIL and VBDA. For VBDA, underestimation of the variance is partly driven by

[9] The code of VBDA follows that of the authors which use matrix form to store the S iterations used to construct the gradient of the lower bound.

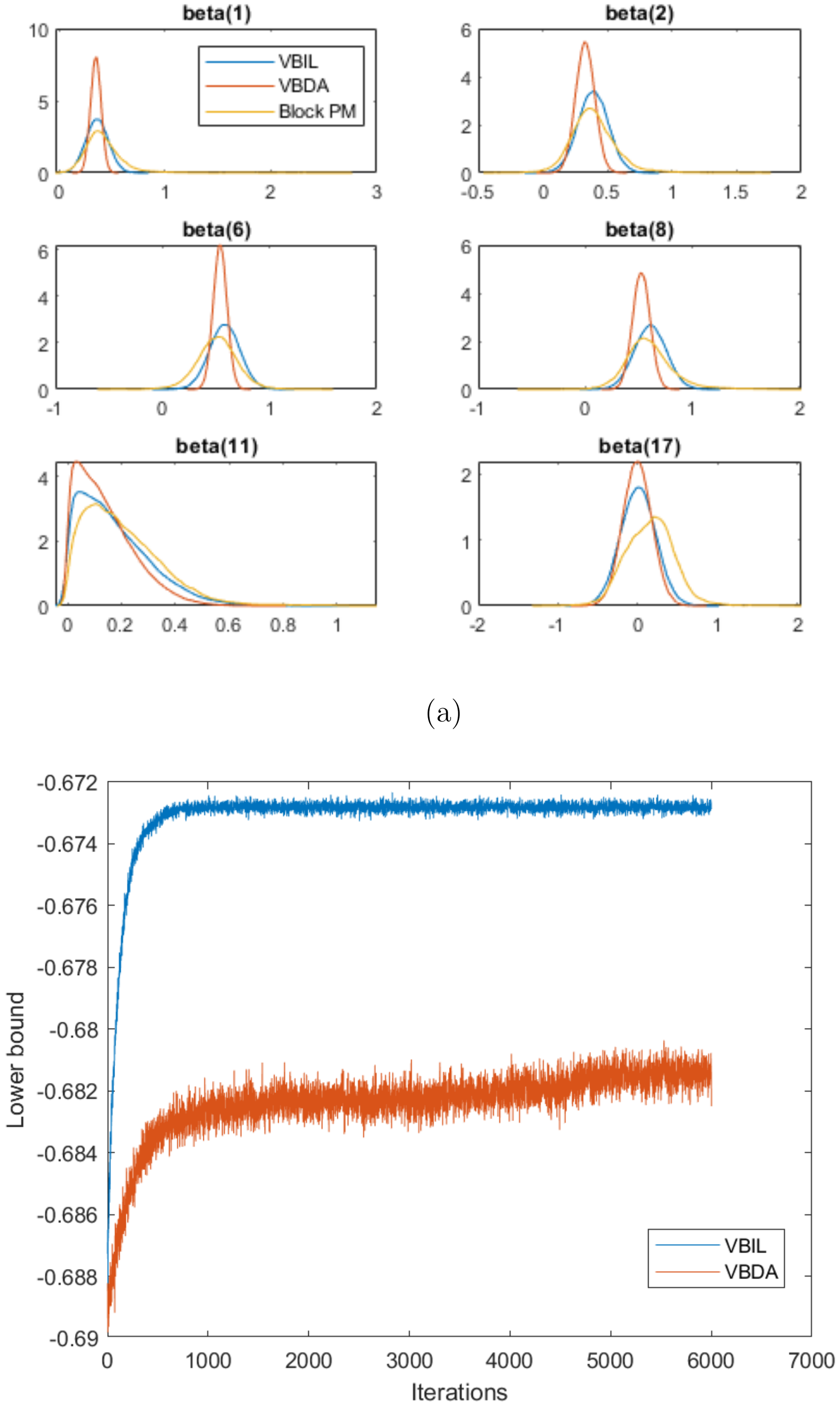
Table 6.3: VBIL and VBDA ordinal and Bernoulli Data Simulations

Data Type	Dimension	Method	No. of Parameters	Specification	Clock Time (mins)
Ordinal	J=10	block PM	-	-	252.00
	N=250	VBIL	57	$S = 50$	3.60
		VBDA	7,307	$S = 50$	717.17
Bernoulli	J=10	block PM	-	-	250.24
	N=250	VBIL	57	$S = 50$	3.71
		VBDA	7,307	$S = 50$	802.50

the independence assumption between $q_{\lambda_\alpha}(\theta)$ and $q_{\lambda_\beta}(u)$ (Blei et al., 2017). The independence assumption allows the variational distribution to capture the marginal distributions of the parameters but not the dependence between them (see Figure F.1 for a description of this phenomenon).

As an additional measure, I use the lower bounds to identify the VB method with the most optimal convergence properties. Since VB aims to find the λ that maximises the lower bound, the method with the highest lower bound will be the optimal VB method. To illustrate the relative magnitude before and after convergence and to exclude the influence of the initial parameter values of $q_{\lambda_\beta}(u)$, Figure 6.4 (b) reports the lower bound of VBDA from the 4000th iteration. Even after excluding 4000 iterations, VBDA takes around another 5000 steps to converge, while VBIL takes less than 1000 steps. The width of the VBDA lower bound implies that there is more variance in the estimates. For all iterations, the lower bound of VBIL is also higher than VBDA. Comparison of the lower bounds suggest that VBIL should yield better posterior estimates than VBDA.

Overall, I find that VBIL performs better than VBDA in speed, variance estimation and maximisation of the lower bound. While VBDA may be more efficient than MCMC for lower dimensions, as shown in Loaiza-Maya and Smith (2019), it is computationally expensive in high dimensions. Therefore, I restrict the analysis to VBIL for the rest of this thesis.



(a)

(b)

Figure 6.4: Plots of the Gaussian copula parameters based on Bernoulli data simulations with dimensions $J = 10$ and $N = 250$: (a) Kernel densities of VBIL and VBDA against block PM; (b) Lower bounds of VBIL and VBDA against the number of iterations.

6.4 Learning Rates

The choice of learning rate is important for convergence as it determines the contribution of new information to the existing estimate of λ. A learning rate that is too small might lead to slow convergence and an algorithm getting stuck at a local maximum, while a large learning rate might cause an algorithm to over-step and diverge.

Learning rates can be 'fixed' or 'adaptive'. Fixed learning rates are manually chosen by the user and satisfy the Robbins-Monro conditions to ensure convergence. Fixed learning rates of the form $1/(t+a)$, where a is an integer and t is t-th iteration are commonly used in practice.

Adaptive learning rates allow each element of λ to have its own learning rate. For example, an adaptive learning rate would be important when there is a parameter on a log scale as it may converge at a different speed to those that are not. Since learning rates are likely to vary across parameters, I expect that high-dimensional problems like the ones considered in this paper will benefit from using adaptive learning rates.

I consider two choices of adaptive learning rates: ADADELTA (Zeiler, 2012) and Ranganath's adaptive learning rate (Ranganath et al., 2013) (see Appendix G for more information). Ranganath's learning rate is used to estimate the Cholesky factor of the inverse covariance matrix described in Section 5.3.1. Both the ADADELTA and Ranganath learning rates incorporate first order and approximations of second order derivative information. ADADELTA does not require specifying the starting values for the learning rates but Ranganath requires initial starting values for the hyperparameters, which are then tuned adaptively. The starting values of Ranganath's adaptive learning rate are calculated based on means at user-specified initial parameter values (see Ranganath et al. (2013) for more information). For Ranganath's adaptive learning rate, Ong et al. (2018a) find that the algorithm is unstable in the early iterations and recommend that a maximum step size of $\phi_t \leq \sqrt{d/c_t}$ is set in the early iterations. They do not comment on when the maximum step size restriction should be removed. In my examples, I restrict ϕ_t until ϕ satisfies $\phi_t - \sqrt{d/c_t} < 0.01$ for one iteration.

6.5 Natural Gradient vs Ordinary Gradient

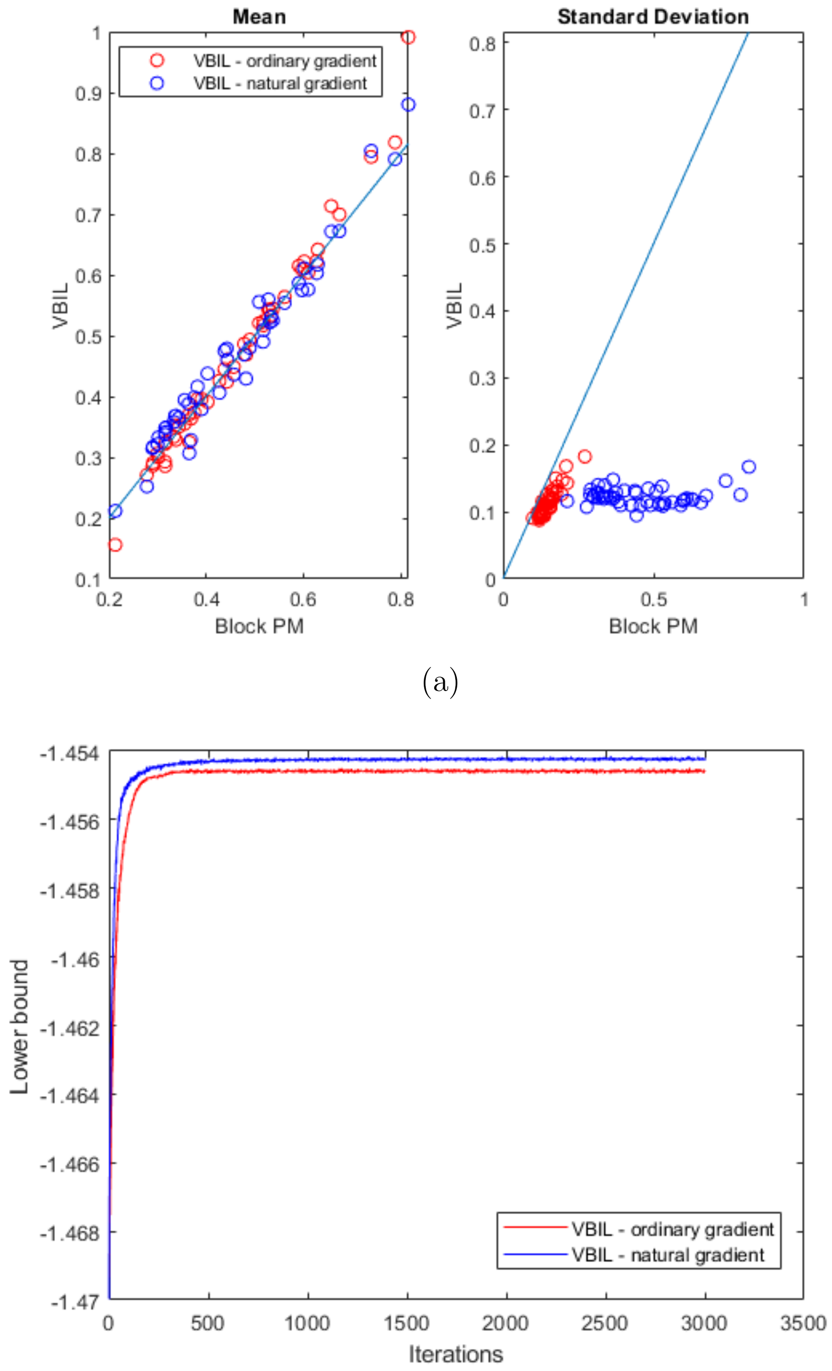
Another option to speed up convergence, is to add second order derivative information into the gradient. Adding second order derivative information increases the speed of convergence by providing additional information on the curvature of the lower bound. Tran et al. (2019) further explains that using an ordinary gradient may insufficiently capture the geometry of the optimisation surface, and therefore reduce the efficiency of convergence. Amari (1998) defines the natural gradient as

$$\nabla_\lambda \text{LB}(\lambda)^{nat} = I_F(\lambda)^{-1}\nabla_\lambda \text{LB}(\lambda), \tag{6.1}$$

where $I_F(\lambda)$ is the fisher information matrix and $\nabla_\lambda \text{LB}(\lambda)$ is the ordinary gradient of the lower bound with respect to λ. Tran et al. (2019), Honkela et al. (2010), and Salimans et al. (2013) also use the natural gradient.

Figure 6.5 (a) and Figure 6.6 (a) show the mean and standard deviation plot of ADADELTA with the ordinary gradient and Ranganath's adaptive learning rate with the natural gradient (see Ong et al. (2018a) for more information) for VBIL applied to the Gaussian Copula. In both $J = 25$ and $N = 250$ and $J = 37$ and $N = 1200$ ordinal data simulations, the natural gradient converges within 250 to 300 iterations, and with a slightly higher lower bound than the ordinary gradient. Despite the faster convergence and higher lower bound, the natural gradient does not perform as well as the ordinary gradient when compared to block PM.

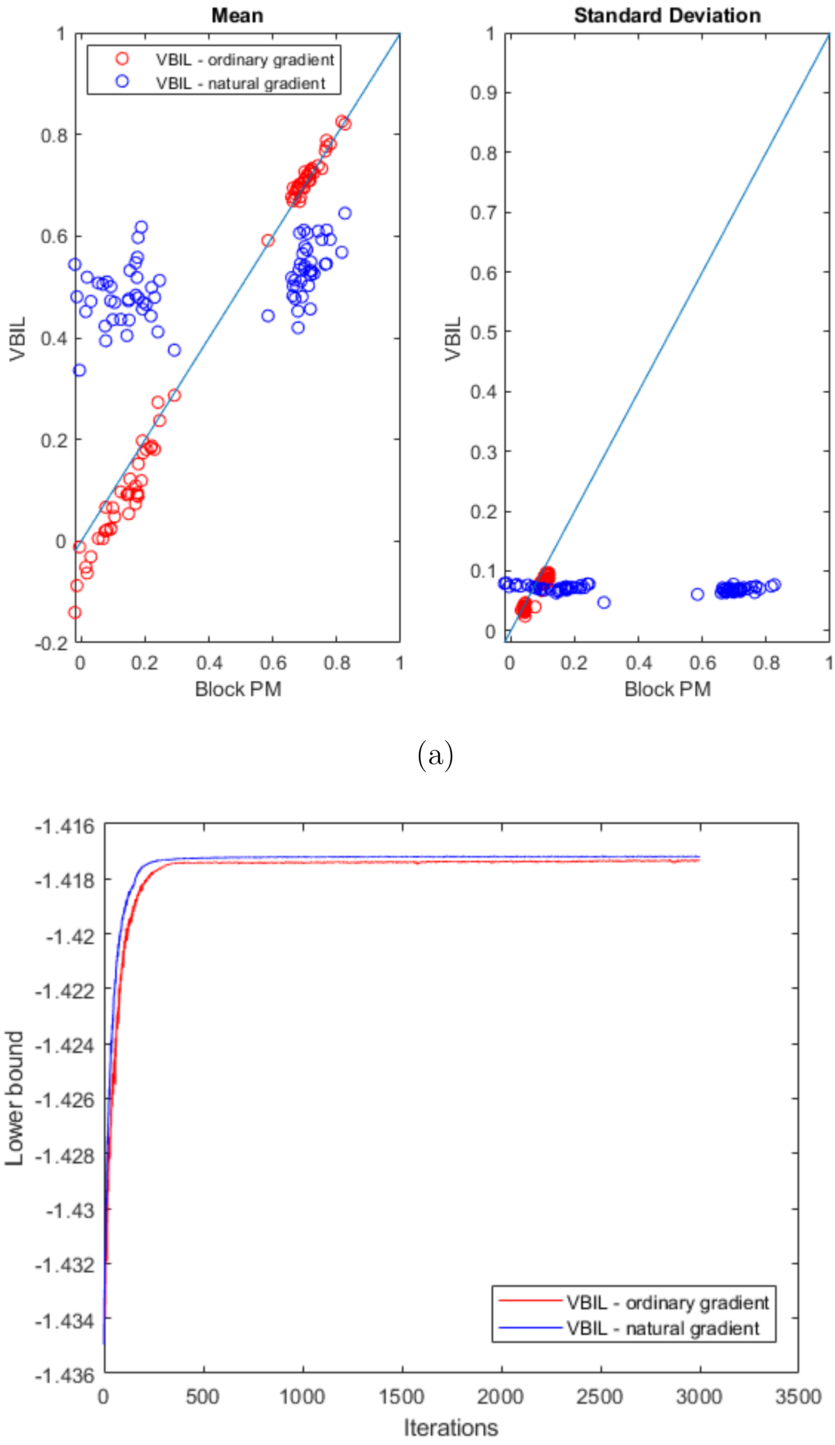
Figure 6.5 (b) shows for dimensions $J = 25$ and $N = 250$ that the means of the β estimates computed using the natural gradient are marginally better the ordinary case when compared to block PM. However, Figure 6.6 (b) shows for dimensions $J = 37$ and $N = 1200$ that the means of the natural gradient both under- and over-estimate the true posterior means of the copula parameters. In both cases, the standard deviations computed using the natural gradient underestimate the true values by more than the ordinary gradient. Therefore, despite faster convergence and higher lower bounds, ADADELTA with the ordinary gradient performs better than Ranganath's adaptive learning rate with the natural gradient when compared to block PM.



(a)

(b)

Figure 6.5: Plots of the Gaussian copula parameters based on ordinal data simulations with dimensions $J = 25$ and $N = 250$: (a) Mean and standard deviations of VBIL with ordinary and natural gradient against block PM; (b) Lower bounds of VBIL with ordinary and natural gradient against the number of iterations.



(a)

(b)

Figure 6.6: Plots of the Gaussian copula parameters based on ordinal data simulations with dimensions $J = 37$ and $N = 1200$: (a) Mean and standard deviations of VBIL with ordinary and natural gradient against block PM; (b) Lower bounds of VBIL with ordinary and natural gradient against the number of iterations.

Chapter 7
An Application to HILDA

This chapter models 36 discrete health questions from HILDA using a Gaussian copula. I use VBIL to approximate the posterior of the copula parameters. From the posterior distribution, information such as credibility intervals, analogous to confidence intervals, can be extracted to capture the uncertainty of the parameter estimates. In addition, from the posterior distribution of the copula parameters, the posterior distribution of any function of the parameters can be obtained. To demonstrate this and to show that VBIL can be used for inference and prediction, I construct a Bayesian multidimensional health index (MHI) to track health over time. To the best of my knowledge, there are no previous papers that construct a Bayesian application of the multidimensional index of Alkire and Foster (2011) solely to measure health. Cross-validation results and the prediction of missing responses can also be easily obtained, although I do not include these as they are applications of the prediction method. Secondly, I use the Gaussian copula to provide insight on the joint and conditional dependence across the dimensions of health. For example, the probability that a person is severely unhealthy, and one's future expectations of health given severe physical health.

7.1 Background Information on Australia's Health System

In Australia, healthcare is one of the largest Government and household spending items; \$185 billion in real terms was spent on healthcare in 2017-18, which is equivalent to \$7,485 per capita (Australian Institute of Health and Welfare, 2019). Of this, over 68 per cent was funded by the Government and the remainder by individuals, private health insurance providers and other non-government sources.

Starting in the 1990s, women's health has become a much discussed topic. In 1996, the government funded the Australian Longitudinal Study on Women's Health. Findings include adult women reporting higher levels of psychological distress than men, which is more common among women with little social support (Holden et al., 2013). More recently, building on the National Women's Health Policy 2010, the National women's health strategy 2020-2030 was established to improve the well-being and health of Australian women by addressing gaps in the provision and awareness of health services (Department of Health, 2018). Key areas of the strategy include improving awareness of mental health, recognising loneliness in older women, and increasing early intervention of chronic health conditions. In 2013, New South Wales Health also established a framework to provide services and an environment to assist in improving their health (NSW Ministry of Health, 2013). Given the recent developments in policies related to women's health, this chapter focuses on women over the age of 40 who live in New South Wales (NSW).

7.2 Existing Tools to Understand Health

Measuring public health is critical in resource allocation to improve health, manage health care costs, and address quality of life (Thacker et al., 2006). Since the health level of an individual or more generally, a population is unobservable, inference can only be made through observable variables. Statistics are popular tools used to track health over time because they summarise the overall health status of a population at a high-level. Example statistics include, numbers and rates of deaths, estimates of life expectancy at birth, disability adjusted life years, health returns to healthcare investment, availability and quality of healthcare, and expenditure on healthcare. Since there are many factors contributing to health, one of these statistics alone cannot provide a complete summary of health.

To gain a better understanding of public health, two or more dimensions can be combined. Due to synergies between the dimensions, the outcome may be different compared to interpreting the two dimensions separately. It is often of interest to determine the contribution of each dimension to a health outcome. Outcome variables related to health are usually discrete, for example the number of visits to a doctor, emergency room, or the number of hospital stays. Practitioners often rely on using regression to understand the relationships between the dimensions of health and health outcomes. However, numerous papers, including Manning et al. (1982), Windmenjer and Santos-Silva (1997), Lourenco (2007) and Van Ourti (2004) claim that regression without additional techniques such as instrumental variables leads to biased results, as there is often an endogenous relationship. For example, a regression with the number of doctor visits as the regressand and self-assessed health as the regressor is biased because both variables are driven by past healthcare receipts. Additionally, an increase in the number of doctor visits may lead to a negative self-assessment of health, while a negative self-assessment of health may also lead to an increase in the number of doctor visits. This example highlights endogeneity issues, which is confounding variables and simultaneity, that often exist in the health literature.

To avoid the endogeneity problem inherent in regression, a number of papers use copulas to obtain probabilistic statements on health. Murteira and Lourenço (2011) use a bivariate copula to model the relationship between the number of doctor visits and self-assessed health. Conlon et al. (2017) use a Gaussian copula to evaluate whether the effect of fluorouracil infusion on cancer progression can be used to indicate the effect on survival time. Since cancer progression is observed before survival time, inference on cancer progression leads to reduced time and cost in obtaining results from clinical trials. García-Gómeza et al. (2019) use an empirical copula to construct a multidimensional view of poverty for all European countries using health as a dimension. The authors use a copula to analyse the dependence between dimensions rather than within a dimension, as a high degree of dependence between dimensions increases poverty. They are then able to determine changing patterns of dependence over time.

Conversely, when only the dimensions are known, they can be combined to form a multidimensional index. While a large number of methods exist to construct a multidimensional index, the methodology of Alkire and Foster (2011) is one of the most used and well-studied in the literature. This methodology is used by large international agencies, like the United Nations and the World Bank, to identify poverty on an international scale. For example, the United Nations constructed the Global Multidimensional Poverty Index to build a comprehensive picture of poverty, based on health, education and standard of living dimensions, and to identify patterns in poverty within a country over time (Oxford Poverty Human Development Initiative). Over 18 countries have adopted this methodology, for example, Chile, Mexico and Vietnam use it to monitor poverty and design more effective government policy (Alkire et al., 2018). The popular use of this methodology lies in the identification of multiple deprivations at the individual level to construct an aggregate measure of poverty.

The World Health Organisation defines health as a ‘state of complete, physical, mental and social well-being and not merely the absence of disease or infirmity’. Based on this definition, health can also be thought of as a multidimensional issue because there are many factors that contribute to health to identify whether a person is healthy or not. To understand the health of Australian women, I model 36 health questions from HILDA using a Gaussian copula and then use the results to construct a Bayesian MHI to carry out inference.

7.3 Data: HILDA Overview

HILDA is a nationally representative survey of more than 17,000 people. It is a panel dataset with 17 waves, with each wave representing the year the data was collected. The first wave was collected in 2001, continuing from thereon on an annual basis. The primary objective of HILDA is to inform policy makers on the economic and personal well-being, labour dynamics and the family life of Australians. HILDA is conducted by the Melbourne Institute of Applied Economic and Social Research on behalf of the Department of Social Services.

As part of understanding the health and well-being of the Australian population, survey respondents are given a self-completion questionnaire and a face-to-face interview. The questionnaire is voluntary, meaning that not all respondents who participate in an interview complete the questionnaire. The general health section of the questionnaire has 36 questions aimed at understanding the physical and mental health of the respondents. All of the questions in this survey have remained the same since the first HILDA survey in 2001. Most of these questions relate to physical and mental health within the last four weeks of the respondent filling out the survey. Responses in the health questionnaire range from binary yes or no, to 6-option ordinal responses.

To track changes in health over time, I consider these three waves of HILDA: 2004, 2010, and 2016. These years are chosen because they are equally spaced and span the available years of HILDA. I expect changes in these years will be larger than any three consecutive years as a smaller proportion of participants remain in the sample after 6 years than after 1 year. Additionally, health problems that are faced by an individual in one year are more likely to be different in 6 years time than in the next year. The three waves are treated as a repeated cross-section to allow all the available information to be included. For example, new participants that were added into the 2011 survey as top-ups and those who have left the sample due to death can also be included. I use all 36 questions from the three waves in my analysis. These questions are:

1. In general, would you say your health is excellent, very good, good, fair, or poor?

2. Compared to one year ago, how would you rate your health in general now?

3. Does your health now limit you in these activities? If so, how much? Vigorous activities, such as running, lifting heavy objects, participating in strenuous sports; moderate activities, such as moving a table, pushing a vacuum cleaner,

bowling or playing golf; lifting or carrying groceries; climbing several flights of stairs; climbing one flight of stairs; bending, kneeling, or stooping; walking more than one kilometre; walking 100 metres; bathing or dressing yourself (10 questions)

4. During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of your physical health? Cut down the amount of time you spent on work or other activities; accomplished less than you would like; were limited in the kind of work or other activities; had difficulty performing the work or other activities; had difficulty performing the work or other activities (for example, it took extra effort) (4 questions)

5. During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)?; Cut down the amount of time you spent on work or other activities; accomplished less than you would like; didn't do work or other activities as carefully as usual (3 questions)

6. During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbours, or groups?

7. How much bodily pain have you had during the past 4 weeks?

8. During the past four weeks, how much did pain interfere with your normal work (including both work outside the home and housework)?

9. These questions are about how you feel and how things have been with you during the past 4 weeks. For each question, please give the one answer that comes closest to the way you have been feeling. How much of the time during the past 4 weeks: Did you feel full of life?; Have you been a nervous person?; Have you felt so down in the dumps that nothing could cheer you up?; Have you felt calm and peaceful?; Did you have a lot of energy?; Have you felt down? Did you feel worn out?; Have you been a happy person?; Did you feel tired? (9 questions)

10. During the past four weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting friends, relatives, etc.)?

11. How TRUE or FALSE is each of the following statements for you? I seem to get sick a little easier than other people; I am as healthy as anybody I know; I expect my health to get worse; my health is excellent. (4 questions)

To make the survey responses consistent in meaning and easy to interpret, responses to all questions are ordered from those implying a deterioration of health to excellent health. For a given question, the most negative response indicating a deterioration of health is assigned a value of 1, with the next in order assigned a value of 2, etc. Table 7.1 gives an example of responses to the question 'In general, would you say your health is?' before and after reordering.

Table 7.1: Reordering of HILDA responses: In General, would you say your health is

	Excellent	Very Good	Good	Fair	Poor
Unordered responses	1	2	3	4	5
Reordered responses	5	4	3	2	1

See the Appendix L for the histograms of the reordered responses. The histograms show little change across the three survey years. Responses are mainly negatively skewed with most people reporting positive health conditions.

Questions related to health were also asked in 2001, 2009, 2013, and 2017 face-to-face interviews. Although, Wilkins et al. (2019) compare the interview responses to questionnaire responses for these years and find that there is a greater tendency to report excellent health in the face-to-face interview. The differences in responses may be a result of acquiescence bias, where respondents are more likely to say positive results in front of an interviewer (Bowling, 2005). Since there is a risk that respondents will overstate their health in the face-to-face interview responses, analysis in this thesis is based on the questionnaire.

The analysis is based on individuals who have completed the questionnaire. Since each question should only have one response, participants who leave one of these questions unanswered or answer with more than one response are omitted from the sample. Of the remaining, only NSW women that are at least 40 are included in the sample.

7.3.1 Sets of Health Questionnaire Questions

Broadly, the 36 questions can be classified into 5 groups: questions that relate to physical health, mental health, impact of physical health on work, impact of mental health on work, and general reflections on health. In all questions, the respondents are only asked to select one response in each question. The responses vary in length from 2 to 6.

I can form 2 sets of questions from the 5 groups. The first set consists of 14 questions and captures information on physical and mental health faced by the individuals, such as whether the respondent is able to walk 100 metres and whether they have been happy in the last four weeks. Not all questions are included as some are similar in meaning to others, or imply responses to other questions. See Appendix L.2 for a complete discussion on choosing the physical and mental health questions for the first set and table L.3 for the questions. The second set of questions includes all 36 questions from the health questionnaire. Although this set includes the questions above that have similar meanings, it is chosen to illustrate the performance of VBIL for high-dimensional data.

7.4 A Bayesian Multidimensional Health Index

This section shows how to construct the MHI of Alkire and Foster (2011) in a Bayesian context for the three waves of HILDA.

7.4.1 MHI construction

Following Alkire and Foster (2011), let N be the number of persons and $J \geq 2$ be the number of dimensions. Let $y = [y]_{nj}$ be a $N \times J$ matrix of survey responses for the set of individuals in a wave of HILDA, with row n denoting person n and column j denoting dimension j.

Each dimension is associated with a cut-off. A person is deprived in a dimension if their selected response is below a predefined dimension cut-off. Let the dimension cut-off for jth dimension be Z_j and the vector of all the dimension cut-offs be Z. The nth person is deprived in the jth dimension if $y_{nj} \leq Z_j$, and is not otherwise. For the HILDA questions, the dimension cut-offs are chosen to divide the responses indicating poor health to those with fair and above health in a dimension. For example, the question ‘Compared to one year ago, how would you rate your health in general now?’ has responses (with the reordered responses in brackets): ‘much worse now than one year ago’ (1); ‘somewhat worse now than one year ago’ (2); ‘about the same as one year ago’ (3); ‘somewhat better than a year ago’ (4); and ‘much better now than a year ago’ (5). I choose the cut-off in this dimension to be the response: ‘somewhat worse now than one year ago’ (2). Therefore, respondents who answer either ‘much worse now than one year ago’ (1), or ‘somewhat worse now than one year ago’ (2) will be deprived in this dimension. See Appendix L.3 for the cut-offs in each dimension. In general, the choice of the dimension cut-offs is subjective (Alkire et al., 2018).

I construct a corresponding deprivation matrix $g \in G$ by assigning $g_{nj} = 1$ when person n is deprived in the jth dimension and $g_{nj} = 0$ otherwise. Let the deprivation vector g_n be the nth row of the deprivation matrix g. The sum of all the elements in g_n represents the number of dimensions the nth person is deprived in.

In addition to the dimension cut-offs, the aggregate cut-off determines whether a person is healthy. Let the aggregate cut-off be 'k'. A person is identified as unhealthy if they have more than k deprivations and is identified as healthy otherwise.

Using the deprivation matrix and the aggregate cut-off, the multidimensional measure is constructed using identification and aggregation steps:

1. The identification step determines whether a person is healthy. The nth person with deprivation vector g_n is identified as unhealthy if they have more than k deprivations. Let the deprivation function be $\phi : G \times \mathbb{R}^+ \to \{0, 1\}$. The nth person is classified as healthy if $\phi(g_n, k) = 0$ and unhealthy otherwise.

2. The aggregation step combines information from the identification step to construct a multidimensional health measure. From the identification step, the deprivation matrix is censored by setting the rows in the deprivation matrix corresponding to healthy persons to zero. Let the censored deprivation matrix with aggregate cut-off k be denoted as $g^{(0)}(k)$. The multidimensional measure M_0 is the mean of the elements of $g^{(0)}(k)$. Therefore, M_0 explains the share of deprivations faced by persons that are unhealthy out of the total possible deprivations that can be faced by the population.

The MHI can be constructed by computing the multidimensional health measure for each of the three HILDA waves.

7.4.2 Choosing the weights

The multidimensional index of Alkire and Foster (2011) implicitly uses equal weights for each dimension. While this is appropriate for dimensions that have equal importance, this is not the case when some dimensions are considered to be worse than others. For example, being limited a lot in bathing or dressing can be considered as more severe than not being able to participate in vigorous activities, such as running or strenuous sports. To reflect the severity of each dimension on health, the dimensions can be re-weighted. Following Alkire and Foster (2011) and Alkire et al. (2017), I choose the weights subjectively. Weights can also be chosen to reflect priorities and goals of government; see (Alkire et al., 2018) for more information.

Instead of assuming equal weights for each dimension, the deprivation matrix g of zeros and ones can be replaced by a weighted deprivation matrix. By definition, the sum of the weights across all the dimensions must be equal to the number of dimensions considered. For example, if there are 14 questions, then the sum of the weights must be 14. For the questionnaire, I put more weight on the physical health dimensions than on the mental health dimensions because physical health conditions are unlikely to be reversible, while mental health conditions usually are. Additionally, a respondent who feels worn out, or is nervous a good bit of the time, may not be too unhealthy because they could still be happy all of the time. Table L.3 in the appendix lists the dimensions and their corresponding weights I have chosen for the analysis.

One should interpret the following results, including the conditional probabilities, probabilistically rather than as causal. These results aim to illustrate that VBIL for a Gaussian copula with discrete margins can be used to understand high-dimensional survey datasets. The dimension cut-offs are shown in Appendix L.3 and are the same for each survey year.

7.4.3 Bayesian version of the MHI construction

A Bayesian version of the MHI gives policy-makers more information on health than directly applying the Alkire and Foster (2011) method outlined in Section 7.4.1. Directly applying the Alkire and Foster (2011) method gives a point estimate, whereas under a Bayesian framework, the posterior distribution of M_0 is obtained. Furthermore, properties of the posterior distribution provide information on the uncertainty of the M_0 estimate. For example, the 95% credibility intervals, which is analogous to confidence intervals can be used to quantify uncertainty.

Using VBIL, I model the 36 health questions in HILDA and obtain the log transformed copula parameters θ. Given the posterior of θ, the posterior of any function of θ, including the multidimensional index of Alkire and Foster (2010) can be obtained. To estimate the posterior of M_0:

1. Generate $\theta^{(i)} \sim q_\lambda(\theta)$ and obtain $\beta^{(i)}$ by taking the exponential of the elements corresponding to the diagonal of $\text{vech}(\beta)$ and converting it into a lower triangular matrix.

2. Predict $\hat{y}^{(i)}_{N\times J}$ from the Gaussian copula with associated covariance matrix $\Sigma^{(i)} = \beta^{(i)}\beta^{(i)\intercal} + I$.

3. Calculate $M_0^{(i)}$. From the data matrix, determine the matrix of deprivations,

then set the rows corresponding to those who are healthy to zero (unweighted deprivation matrix). For example when $k = 1$, a person is identified as unhealthy if their total deprivations is greater than 1:

$$\text{persons}\left\{\begin{pmatrix} \hat{y}_{11} & \hat{y}_{12} & \hat{y}_{13} \\ \hat{y}_{21} & \hat{y}_{22} & \hat{y}_{23} \\ \hat{y}_{31} & \hat{y}_{32} & \hat{y}_{33} \end{pmatrix}\right. \Rightarrow g(1) = \overbrace{\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}}^{\text{dimensions}} \Rightarrow M_0^{(i)} = \mu\left(\begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}\right).$$

Replace g with a weighted deprivation matrix when the dimensions are of different severity.

4. Repeat until $i = B$ (where I take $B = 1000$).

As a result, $M_0^{(1)}, M_0^{(2)}, \dots, M_0^{(B)}$ are obtained. The posterior distribution of M_0 can be constructed by fitting a kernel density estimate to the M_0 sample. The posterior mean can be obtained by calculating the mean of the sample, the posterior median by calculating the median of the sample, and the 95% credibility interval by extracting the 25th and 975th values of the sample. Despite the posterior of θ following a normal distribution, the posterior of M_0 is not necessarily normally distributed.

In estimating the posterior distribution of M_0, the posterior distribution of the copula parameters is obtained in step 1. Similarly to M_0, the properties of the posterior distribution of β can be inferred from the β samples. For example, the 95% credibility interval can be used to capture the uncertainty of the β estimates. Compared to classical bootstrap approaches for application to high-dimensional datasets, it is easier to obtain more accurate estimates of uncertainty under a Bayesian framework (Pham, 2006).[10]

7.4.4 Results

I construct a Bayesian MHI with a weighted deprivation matrix for questions in the first set (14 dimensions) for the survey years 2004, 2010, and 2016. Consider first the posterior of M_0 when the aggregate cut-off is 3 for year 2016 (Figure 7.1). To gauge the accuracy of VBIL in estimating the posterior of M_0, I also compute the parameters using block PM. Block PM did not converge within 48 hours and

[10] Here, the classical bootstrap refers to using the parametric bootstrap to obtain confidence intervals of the β estimates. Each bootstrap sample requires computation of the likelihood, which is computationally expensive, as shown in Section 3.2. Although the non-parametric bootstrap is an alternative method which can be used to quantify the uncertainty of M_0 when directly applying the method of Alkire and Foster (2011), this example demonstrates how inference on M_0 is carried out using a Bayesian approach.

required using initial parameter values that were closer to the true posterior than VBIL. Despite differences in initial parameter values, Figure 7.1 shows that the kernel density of M_0 estimated using VBIL largely overlaps with the kernel density estimated using block PM, indicating that VBIL provides a good approximation to the true posterior distribution of M_0. From the posterior distribution of M_0, I can obtain any property of the posterior distribution. For example, the posterior mean is 0.044, the posterior median is 0.044, the 50 per cent credibility interval is equal to [0.041, 0.046], and the 95 per cent credibility interval is equal to [0.036, 0.052]. The posterior mean implies that 4.4 per cent of deprivations are faced by the unhealthy out of all the possible deprivations that can be faced by Australian women over the age of 40 who live in NSW. The small value suggests that these women are generally healthy. The 95% credibility interval corresponds to the 2.5th and 97.5th quantiles of the posterior estimates and therefore, the credibility interval captures 95% of the M_0 estimates. The point estimate 0.045, estimated by directly applying Alkire and Foster (2011) method, also lies within the 95% credibility interval.[11]

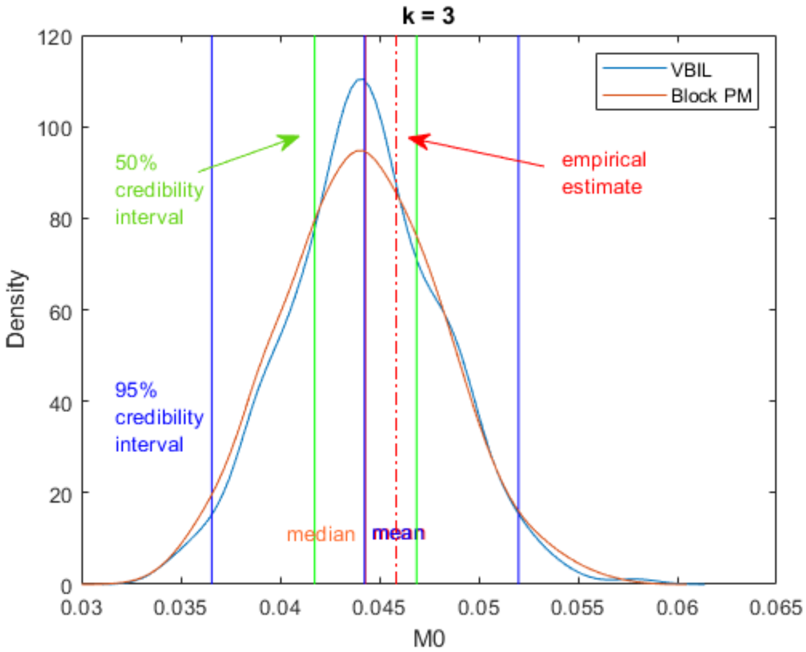


Figure 7.1: Estimates of the posterior distribution, posterior mean, posterior median, and 50 and 95 percent credibility intervals of the multidimensional measure M_0 for 2016 using an aggregate cut-off of 3. The vertical line in red denotes the point estimate.

Setting the aggregate cut-off equal to 3 means that a person is identified as unhealthy when they have 3 or more deprivations. Here, a respondent who answers they are

[11] The point estimate and the posterior mean are not the same because of differences in the estimation method and the choice of prior distribution.

‘limited a lot’ in bathing or dressing, walking 100 metres, climbing one flight of stairs, and in doing vigorous activities, will be deprived in four dimensions, if all of the dimension cut-offs are ‘limited a lot’. Additionally, since the aggregate cut-off is 3, the respondent will be identified as unhealthy.

Since the aggregate cut-off is chosen somewhat arbitrarily, I consider the cut-offs 3, 5 and 7 in Figure 7.2. Here, all the posterior distributions are constructed using a weighted deprivation matrix. As the aggregate cut-off increases, the corresponding posterior distribution shifts to the left and the posterior mean of M_0 decreases. This can be explained by the censoring of data. When the aggregate cut-off increases, the number of people who are identified as unhealthy falls and more rows in the deprivation matrix are set to zero. The mean of the matrix in this case is much smaller than the mean of a matrix corresponding to a lower aggregate cut-off. Intuitively, this occurs because increasing the threshold for a person to be determined unhealthy results in less people being classified as unhealthy.

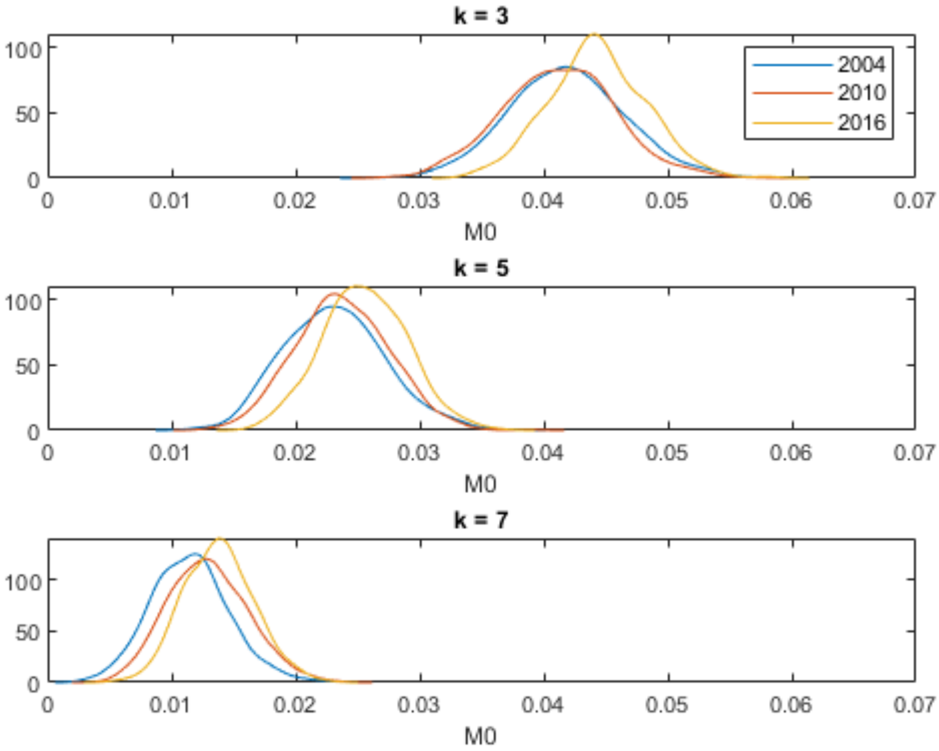


Figure 7.2: Posterior distributions of the weighted M_0 estimates for year 2004, 2010, 2016 with an aggregate cut-off of 3, 5, 7

Table L.2 in the appendix shows shifts in the posterior mean for different aggregate cut-offs to illustrate the effect of aggregate cut-offs on the identification of the unhealthy. For 2004, an increase in the aggregate cut-off from $k = 3$ to $k = 5$ is associated with a fall in M_0 from 0.041 to 0.023. Similarly, an increase from $k = 5$ to $k = 7$ is associated with a fall in M_0 from 0.023 to 0.011. Since the difference between $k = 5$ and $k = 3$ is around 0.02 and between $k = 7$ and $k = 5$ is around 0.01, the asymmetric change in the posterior means implies that individuals are more likely to have deprivations between 3 to 5 deprivations than 5 to 7 deprivations. Asymmetric change in the posterior means between the three years are also observed for 2010 and 2016.

A 1% increase in the posterior mean implies that the number of deprivations faced by those that are unhealthy as a share of the total possible deprivations that can be faced by the population increases by a multiple of 1.01. To determine the impact of an increase in M_0, consider the case when M_0 is calculated using an unweighted deprivation matrix. If the number of deprivations faced by the unhealthy is 1,000 then, for the same population, a one per cent increase in M_0 is associated with an increase of $0.01 \times 1,000 = 10$ deprivations faced by the unhealthy. An increase in M_0 may be a reflection of (1) more respondents being classified as unhealthy; (2) increased deprivations for respondents who are already classified as unhealthy; and for weighted M_0, (3) an increase in the severity of the deprivations faced by respondents who are already classified as unhealthy. The interpretation of an M_0 based on a weighted deprivation matrix is harder because changes reflect both the number and severity of the deprivations. In Figure 7.2, which shows the posterior distributions estimated using a weighted deprivation matrix, the posterior mean of M_0 for $k = 7$ in 2004 is 0.011 and is around 0.002 larger in 2016. Since this difference is almost zero, this suggests that the health of women changed very little over the years.

Additionally, the posterior distributions for the 2004, 2010 and 2016 surveys all overlap across the three cut-offs, suggesting that the health of women has not changed significantly within the twelve years. The point estimates of M_0 obtained from implementing Alkire and Foster (2011) also fall within the credibility intervals of the posteriors (Appendix table L.2).

To illustrate that this method applies to even higher data dimensions, Figure 7.3 constructs the Bayesian MHI for all 36 equally weighted dimensions of the questionnaire with aggregate dimensions 5, 10 and 15. Similarly to the weighted deprivation matrix constructed for questions in the first set. Figure 7.2 shows that the kernel densities for all years overlap, implying that health responses have not changed significantly within the twelve years.

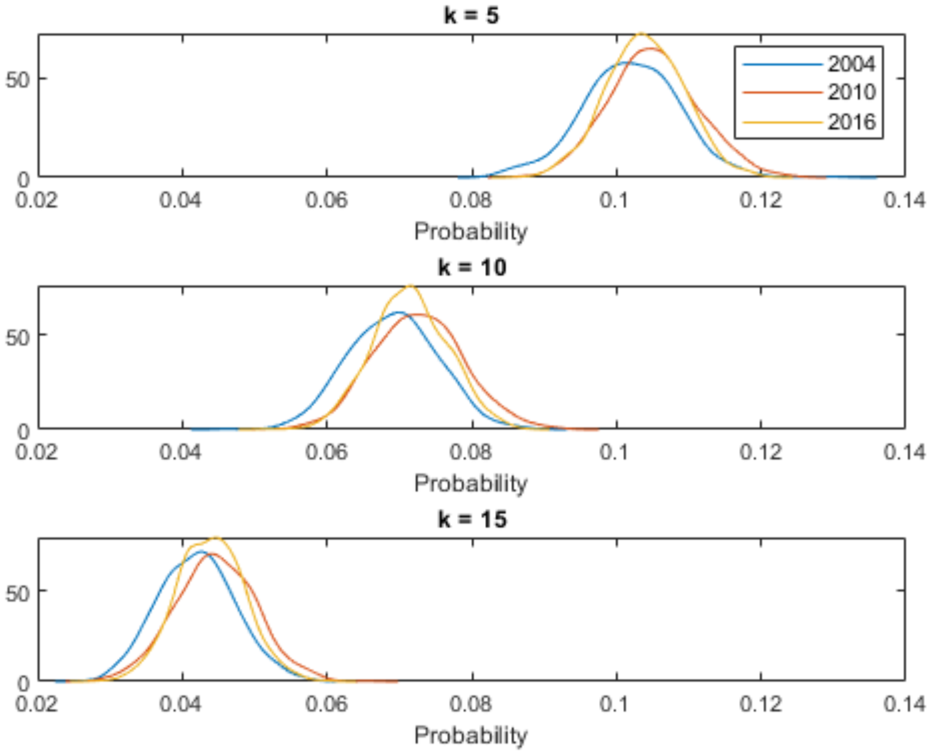


Figure 7.3: Posterior distributions of unweighted M_0 estimates for year 2004, 2010 and 2016 with aggregate cut-offs 5, 10 and 15.

To evaluate the efficiency and accuracy of VBIL, I constructed a Bayesian MHI and benchmarked the results to block PM, which involved modelling the 36 health questions from HILDA, estimating the posterior of the Gaussian copula parameters using VBIL, and using prediction to estimate the posterior of M_0. Similarly to the simulation studies in Chapter 6, I found that VBIL produced similar results in terms of accuracy but required significantly less computation time than block PM. The overlapping posterior distributions for the three survey years indicate that health has not changed significantly over time. Compared to directly applying the Alkire and Foster (2011) method, the Bayesian approach is more informative because inference is from a posterior distribution rather than from a point estimate. I also obtained, from the posterior distributions of M_0, the posterior mean, posterior median, 50 and 95% credibility intervals. Policy makers may benefit from the additional information provided in the Bayesian approach, such as the 95% credibility intervals to measure uncertainty.

7.5 Direct Inference from the Copula

7.5.1 Joint and conditional distributions

Aside from obtaining a Bayesian MHI, results from the Gaussian copula can also be used to determine the probability that a respondent selects responses equal to or below the dimension cut-offs, defined in the previous section. The joint distribution of Y is

$$P(Y_1 \leq c_1, Y_2 \leq c_2, \ldots, Y_d \leq c_d) = C_\Lambda(F_1(c_1), ..., F_d(c_d)), \tag{7.1}$$

where c_d is the cut-off of the random variable Y_d corresponding to the dth question, and C_Λ is a copula with correlation matrix Λ. Since the number of respondents with responses below cut-offs is small in each survey year, a Bayesian framework is more useful than an empirical estimate because it is able to capture uncertainty.

The posterior distribution of $P(Y_1 \leq c_1, Y_2 \leq c_2, \ldots, Y_d \leq c_d)$ is obtained by first evaluating Equation 7.1 for $\Lambda^{(i)}$, where $i = 1, \ldots, 1000$ and then fitting a kernel density estimation over the sample points. The correlation matrix $\Lambda^{(i)} = D^{(i)}\Sigma^{(i)}D^{(i)}$, where $\Sigma^{(i)}$ is defined in Section 7.4.3 and $D^{(i)}$ is a diagonal matrix with entries $\frac{1}{\sqrt{\Sigma^{(i)}_{jj}}}, j = 1, \ldots, d$.

Using the posterior distribution of $P(Y_1 \leq c_1, Y_2 \leq c_2, \ldots, Y_d \leq c_d)$, I examine the probability that a respondent selects responses below the cut-offs for all mental health questions in the first set (such as feeling full of life, being a nervous person). I also examine the probability that a respondent selects responses below the cut-offs for all physical health questions in the first set (such as walking 100 meters, climbing several flight of stairs). Figure 7.4 shows plots of these posterior distributions for the three survey years.

The posterior mean of the probability of selecting responses below the cut-offs for all mental health questions is 0.0016, 0.0022 and 0.0022, while the posterior mean associated with physical health questions is 0.0064, 0.0075 and 0.0087 for 2004, 2010 and 2016. According to the posterior mean estimates, on average both the probability of selecting responses below the cut-offs for mental health questions and physical health questions increased from 2004 to 2010. From 2010 to 2016, the posterior mean associated with mental health questions remained the same, while for the posterior mean associated with physical health increased. The larger posterior mean for 2016 compared to 2004 and 2010 suggests that physical health is a driver of the increased number and severity of deprivations faced by the unhealthy in Figure 7.2.

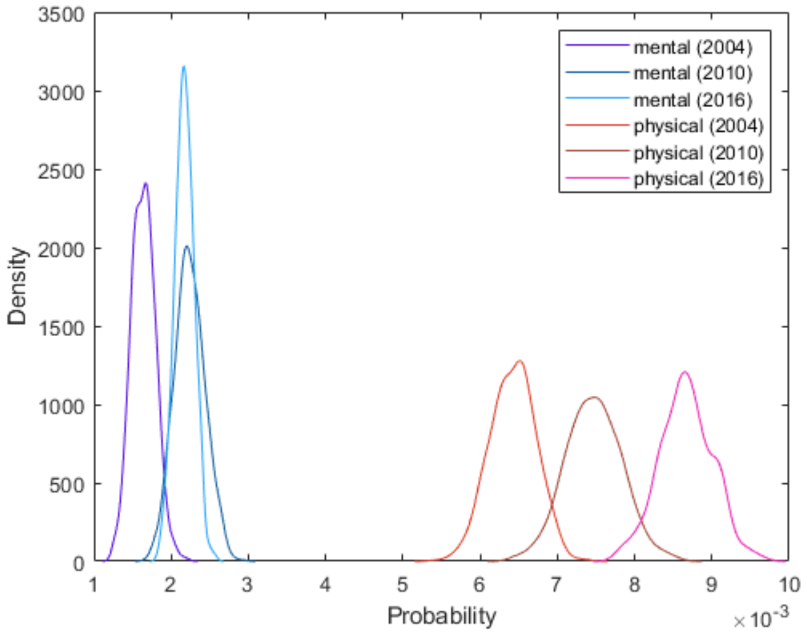


Figure 7.4: Kernel density plots of the probability that a respondent selects responses below the cut-offs for mental health questions (blue) and physical health questions (red).

In addition to the posterior means, which provide information on the average probability estimates of selecting responses below cut-offs, the probabilities need to be interpreted with their uncertainty estimates. Two overlapping posterior distributions with one posterior having a larger mean, does not imply that the posterior with the larger mean has a strictly larger probability of having responses below the cut-offs than the other. Instead, it implies that on average the probability of having responses below the cut-offs is larger than the other, but there is still a chance that they may be statistically the same. This highlights the importance of inference from the entire posterior distribution rather than inference from means or empirical probabilities alone. Here, the overlapping posterior distributions in Figure 7.2 suggest that differences in mental and physical health over the survey years may be statistically insignificant. Although the mode of the posteriors associated with physical health is larger than that of mental health, which implies that more respondents are likely to select all of their responses below the cut-offs for physical health questions than mental health questions.

The probability that a respondent selects a response to a question given that they selected responses equal to or below the cut-offs for a set of other questions is

$$P(Y_1 = y_1 | Y_2 \leq c_2, \ldots, Y_d \leq c_d) = \frac{C_{\Lambda_1}(F_1(y_1), ..., F_d(c_d)) - C_{\Lambda_1}(F_1(y_1^-), ..., F_d(c_d))}{C_{\Lambda_2}(F_2(c_2), ..., F_d(c_d))}, \tag{7.2}$$

where c_d is the cut-off of the random variable Y_d corresponding to the dth question, and Y_1 is the random variable corresponding to the variable of interest. The correlation matrix Λ_1 is for variables from Y_1 to Y_d, whereas Λ_2 is for variables from Y_2 to Y_d.

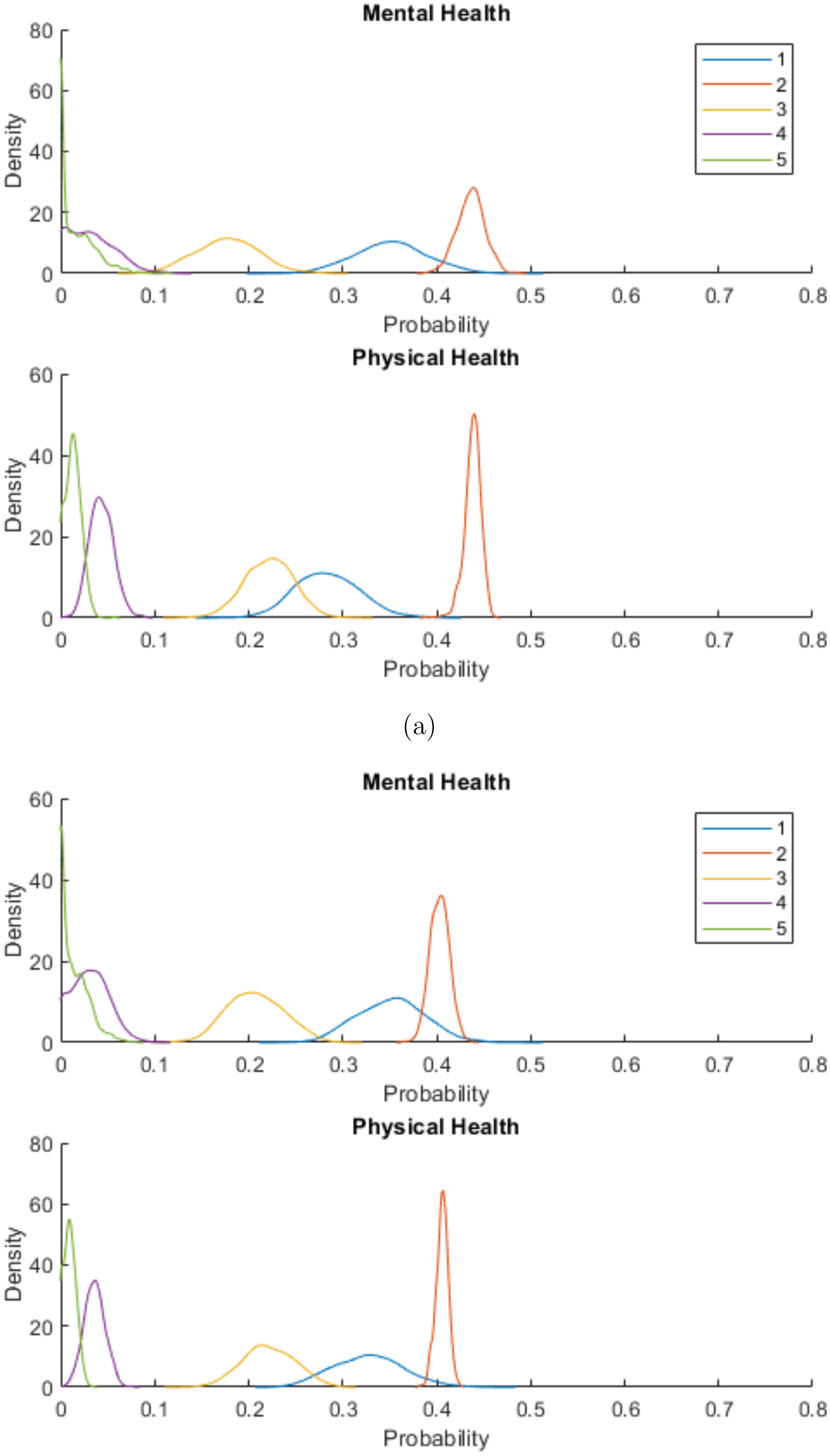
The posterior distribution of $P(Y_1 = y_1 | Y_2 \leq c_2, \ldots, Y_d \leq c_d)$ is obtained by first evaluating Equation 7.2 at $\Lambda_1^{(i)}$ and $\Lambda_2^{(i)}$, for $i = 1, \ldots, 1000$ and then fitting a kernel density over the sample points. $\Lambda_1^{(i)}$ is the matrix obtained by extracting the rows and columns corresponding to Y_1 to Y_d from $\Lambda^{(i)}$ given by $\Sigma^{(i)}$ as in from Section 7.4.3. Similarly, $\Lambda_2^{(i)}$ is obtained by extracting the rows and columns corresponding to Y_2 to Y_d from $\Lambda^{(i)}$.

Using the posterior distribution of $P(Y_1 = y_1 | Y_2 \leq c_2, \ldots, Y_d \leq c_d)$, I consider a person's future expectation on health (Y_1) given that the person is limited in being able to do physical activities (Y_2 to Y_d). I also examine future expectations given mental health responses below the cut-offs, defined in Table L.3. Panels (a)-(c) of Figure 7.5 plot these posterior probabilities for the years 2004, 2010 and 2016.

In all the three survey years, the ordering of responses according to their posterior probabilities is the same. For the mental health questions, in 2004, around 43 per cent believe it is 'mostly true' that their health will get worse, 35 per cent believe it is 'definitely true', 20 per cent respond they 'don't know', and a low proportion believe 'it is false'. The expectations on future health are similar for questions related to physical health except more people believe they 'don't know' whereas, less believe that it is 'definitely true' that their health will worsen. The similarity in the conditional posterior probabilities on future expectations in health across mental and physical health suggests that once a respondent obtains responses below the defined cut-offs, in either physical or mental health, they will have similar future expectations. For 2010, the posterior probability of a respondent responding 'mostly true' to a deterioration in health given they have selected responses below the cut-offs for mental health questions, decreased to around 40 per cent from levels observed in 2004. Similarly, for the same year, this is observed for physical health questions. In 2016, this reverted to previous levels seen in 2004 for the physical health questions.

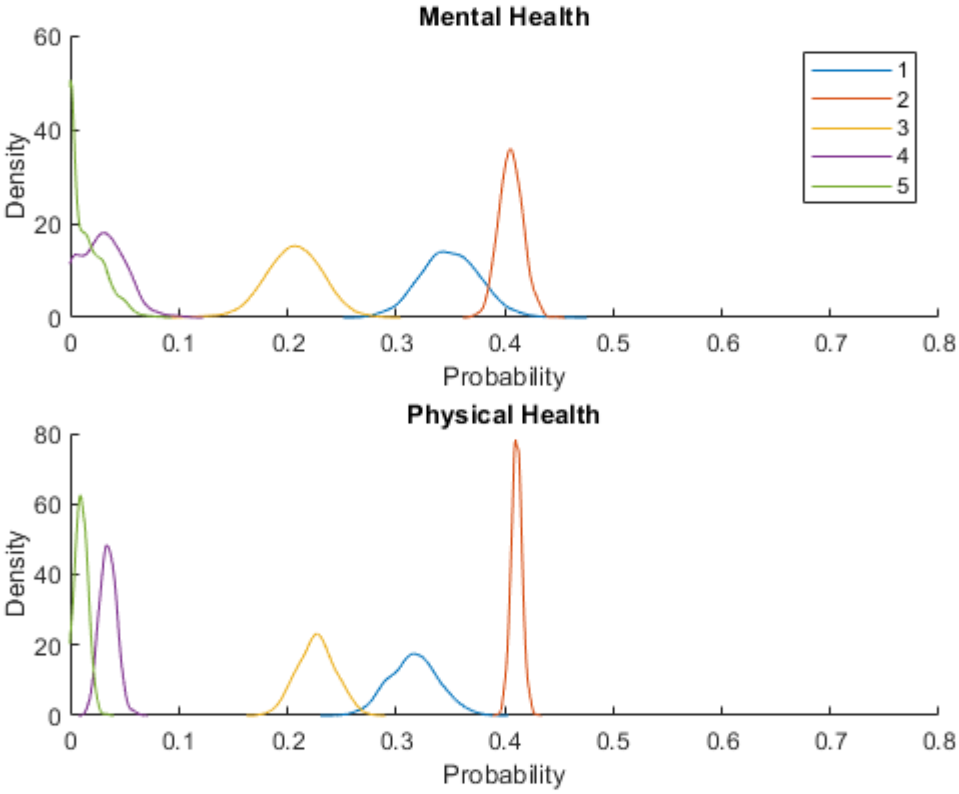
One explanation for the small change in health observed for NSW women over the age of 40 is the panel nature of the dataset. Persons in each of the survey years are predominantly the same. To reduce the panel effect, the amount of overlap between people in each of the three years can be reduced. For example, by including only those who are 50 to 55 years of age ensures that persons in the sample are different each year. Another alternative, is to randomly sample the population of women to increase variation in the physical and mental health responses. A second explanation for the small change in health is that health may be 'sticky'. To ensure that the results are robust, additional checks should be completed for these empirical examples, such as choosing different sets of weights and cut-offs, as well as different combinations of dimensions. These extensions are not explored in this thesis because this chapter is used to illustrate methodology.

This chapter provides further evidence to that in Chapter 6 that VBIL is an efficient estimation method. I have shown that VBIL can be used to conduct fast inference and prediction, such as to estimate a Bayesian MHI and to infer relationships between health dimensions using a copula. These empirical examples illustrate the possible types of analysis that can be conducted using VB methods for Gaussian copulas with discrete margins.


Mental Health
Physical Health
Density
Probability
0
20
40
60
80
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
1
2
3
4
5

(a)

(b)



(c)

Figure 7.5: Kernel density plots of 'I expect my health to get worse' given that responses to mental and physical health responses are below cut-offs for year (a) 2004; (b) 2010; (c)2016. Response (1) refers to 'definitely true', (2) refers to 'mostly true', (3) refers to 'don't know', (4) refers to 'mostly false', and (5) refers to 'definitely false'.

Chapter 8
Conclusion

This thesis extends the VB methodology proposed by Gunawan et al. (2019) and Loaiza-Maya and Smith (2019) to the case of the Gaussian copula with discrete margins. I overcame the computational challenges of estimating the Gaussian copula parameters by using an augmented posterior distribution and a factor structure covariance matrix. Similarly, I imposed a factor structure on the covariance matrix of the Gaussian variational distribution to reduce the number of parameters to be estimated. For larger data and parameter dimensions than those considered by Gunawan et al. (2019) and Loaiza-Maya and Smith (2019), this thesis shows that the proposed VB methods yield good approximations of the true posterior distribution of the copula parameters. Out of the two VB methods, VBIL was found to be the most computationally efficient, providing significant improvement in computation time with little trade off in accuracy. This thesis also extended VBIL to ordinal data applications. To illustrate this extension, a Bayesian MHI to track health over time for 36 health questions in HILDA was constructed. I showed that the Bayesian approach to construct a multidimensional health measure was more informative than directly applying Alkire and Foster's method (2011), as inference is from a posterior distribution rather than from a point estimate. I also showed how to extract properties of the posterior distribution, such as the 95% credibility interval to capture the uncertainty of the M_0 estimate. Finally, I demonstrated that a Gaussian copula can be used to estimate the probability that a person is severely deprived in physical and mental health, and their outlook on life given they are severely deprived.

8.1 Future Research

This thesis assumes for simplicity that the number of factors r for the covariance matrix associated with the Gaussian copula, and $\tilde{r}$ for the covariance matrix of the variational distribution are fixed. However, the number of factors should depend on the dataset, with more complex datasets requiring a larger number of factors to fully capture the dependence structure. One method to choose r is to use the log predictive Bayes factors (Kastner et al., 2017). However, I find that setting $\tilde{r} = 1$ is sufficient in all my applications because the VBIL results match those from block

PM. In future applications, where $\tilde{r}$ may not be 1, the value of $\tilde{r}$ can be chosen by running the VBIL algorithm on increasing values of $\tilde{r}$ and stopping when the results stabilise (Ong et al., 2018b). Allowing r and $\tilde{r}$ to vary leads to an increase in computation time, although the total time to run VB will still be much less than that required for MCMC.

Subsampling can be used to reduce the computation time of the MCMC and the VBIL algorithm for larger datasets than those considered here (Quiroz et al., 2019). Subsampling involves sampling at random $\tilde{n} \ll n$ observations, where n is the original number of observations used to compute the likelihood at each iteration. Sampling a smaller set of observations will reduce the computation cost of estimating the likelihood (Dang et al., 2019; Gunawan et al., 2018).

Appendix A
Deriving the Unbiased Likelihood Estimate

Given the form of the augmented posterior distribution in Section 3.2, I show how to derive an unbiased estimate of the likelihood, $\widehat{p}_M(y|\theta, u)$, for a Gaussian copula with two discrete margins. The number of discrete margins can easily be generalised to higher dimensions. To estimate the likelihood, I first show how to estimate the probability at a point (also referred to as the likelihood term). Suppose that $p(Y_1 = y_1, Y_2 = y_2)$ is the probability at a point given θ. Notation-wise, I omit θ for clarity.

Let $Y \in \mathbb{R}^{\mathrm{N \times J}}$ be the data matrix with $J = 2$ with Y^* a continuous latent vector such that when $y^- < Y^* \leq y$ then $Y = y$. Denote the joint density of Y^* as $f(\cdot, \cdot)$.[12] The probability at a point is given by,

$$\begin{aligned} p(Y_1 = y_1, Y_2 = y_2) &= \mathbb{P}(y_1^- < Y_1^* \leq y_1, y_2^- < Y_2^* \leq y_2) \\ &= \int_{y_1^-}^{y_1} \int_{y_2^-}^{y_2} f(y_1^*, y_2^*) \, dy_1^* \, dy_2^*. \end{aligned} \tag{A.1}$$

Rewriting $f(\cdot, \cdot)$ in terms of a copula with correlation matrix Λ,

$$\begin{aligned} f(y_1^*, y_2^*) &= \frac{\partial^2}{\partial y_1^* \partial y_2^*} F(y_1^*, y_2^*) \\ &= \frac{\partial^2}{\partial y_1^* \partial y_2^*} C(u_1, u_2 | \Lambda), \text{ where } u_j = F_j(y_j^*) \\ &= c(u_1, u_2|\Lambda) \prod_{j=1}^{2} f(y_j^*). \end{aligned}$$

Substitute into Equation A.1 and, for simplicity of notation, let $a_j = F_j(y_j^-)$ and

[12] The use of latent variables is similar to the latent-variable model in logistic regression.

$b_j = F_j(y_j)$, for $j = 1, 2$, I obtain that

$$\begin{aligned}
p(Y_1 = y_1, Y_2 = y_2) &= \int_{a_1}^{b_1} \int_{a_2}^{b_2} c(u_1, u_2|\Lambda) \frac{f_1(y_1^*) f_2(y_2^*)}{f_1(y_1^*) f_2(y_2^*)} \, du_1 \, du_2 \\
&= \int_{a_1}^{b_1} \int_{a_2}^{b_2} c(u_1, u_2|\Lambda) \, du_1 \, du_2 \\
&= \int_{a_1}^{b_1} \int_{a_2}^{b_2} \frac{\partial^2}{\partial u_1 \partial u_2} \Phi_\Lambda(\Phi^{-1}(u_1), \Phi^{-1}(u_2)) \, du_1 \, du_2 \\
&= \int_{a_1}^{b_1} \int_{a_2}^{b_2} \phi_\Lambda(\Phi^{-1}(u_1), \Phi^{-1}(u_2)) \frac{dz_1^*}{du_1} \frac{dz_2^*}{du_2} \, du_1 \, du_2 \\
&= \int_{a_1}^{b_1} \int_{a_2}^{b_2} \phi_\Lambda(\Phi^{-1}(u_1), \Phi^{-1}(u_2)) \prod_{j=1}^{2} \frac{1}{\phi(\Phi^{-1}(u_j))} \, du_1 \, du_2,
\end{aligned}$$

where $z^* = \Phi^{-1}(u)$, so that $\Phi(z^*) = u$. To rewrite the correlation matrix Λ in terms of the covariance matrix Σ, let D be a diagonal matrix such that $\Lambda = D\Sigma D$. The diagonal entries of D are equal to $D_{jj} = 1/\sigma_{jj}$, where σ_{jj} is the standard deviation of Σ. Now, let $z = D^{-1} z^*$, where $z_j^* \sigma_{jj} = x_j$. Thus,

$$\begin{aligned}
p(Y_1 = y_1, Y_2 = y_2) &= \int_{\Phi^{-1}(a_1)}^{\Phi^{-1}(b_1)} \int_{\Phi^{-1}(a_2)}^{\Phi^{-1}(b_2)} \phi_\Lambda(z_1^*, z_2^*) \, dz_1^* \, dz_2^* \\
&= \int_{\sigma_{11}\Phi^{-1}(a_1)}^{\sigma_{11}\Phi^{-1}(b_1)} \int_{\sigma_{22}\Phi^{-1}(a_2)}^{\sigma_{22}\Phi^{-1}(b_2)} (2\pi)^{-1} |D\Sigma D|^{-1/2} e^{-1/2 x^\intercal D^\intercal (D\Sigma D)^{-1} D x} \prod_{j=1}^{2} \frac{1}{\sigma_{jj}} \, dx_1 \, dx_2 \\
&= \int_{\sigma_{11}\Phi^{-1}(a_1)}^{\sigma_{11}\Phi^{-1}(b_1)} \int_{\sigma_{22}\Phi^{-1}(a_2)}^{\sigma_{22}\Phi^{-1}(b_2)} \phi(x|\Sigma) \, dx_1 \, dx_2.
\end{aligned} \tag{A.2}$$

More generally, the likelihood term for J-dimensions is

$$p(Y_{1,n} = y_1, \ldots, Y_{J,n} = y_J|\Sigma) = \int_{\sigma_{11}\Phi^{-1}(a_1)}^{\sigma_{11}\Phi^{-1}(b_1)} \ldots \int_{\sigma_{JJ}\Phi^{-1}(a_J)}^{\sigma_{JJ}\Phi^{-1}(b_J)} \phi_\Sigma(x) \, dx_1 \ldots \, dx_J.$$

Each likelihood term can be estimated unbiasedly using a modified version of the Genz's algorithm (1992) or Botev's algorithm (2017) (see Appendix D for information on a modified version of the Genz (1992) algorithm). Similarly to Gunawan et al. (2019), I can obtain an unbiased estimate of each likelihood term using importance sampling with a uniform density proposal.

Since each likelihood term can be estimated unbiasedly, the likelihood

$$\widehat{p}_M(Y|\theta,u) = \prod_{n=1}^{N} \widehat{p}_M(Y_{1,n} = y_1, Y_{2,n} = y_2, \ldots, Y_{J,n} = y_J)$$

is also an unbiased estimate because

$$\begin{aligned} \mathbb{E}[\widehat{p}_M(y|\theta,u)] &= \mathbb{E}[\prod_{n=1}^{N} \widehat{p}_M(Y_{1,n} = y_1, Y_{2,n} = y_2, \ldots, Y_{J,n} = y_J)] \\ &= \prod_{n=1}^{N} \mathbb{E}[\widehat{p}_M(Y_{1,n} = y_1, Y_{2,n} = y_2, \ldots, Y_{J,n} = y_J)] \\ &= \prod_{n=1}^{N} p(Y_{1,n} = y_1, Y_{2,n} = y_2, \ldots, Y_{J,n} = y_J) \\ &= p(Y_{1,n} = y_1, Y_{2,n} = y_2, \ldots, Y_{J,n} = y_J), \end{aligned} \tag{A.3}$$

where $Y_n, n = 1, .., , N$ are independent.

Appendix B
Gaussian Copula: Identification of a Covariance Matrix with a Factor Structure

A factor structure can be used to reduce the number of parameters, although it suffers from identification issues unless further restrictions are made. Let β be a $J \times r$ ($r \ll J$) matrix of factor loadings, $z \sim N(0, I)$ be a $r \times 1$ factor matrix (I is an identity matrix), and $\epsilon \sim N(0, I^2)$ be a $r \times 1$ matrix of idiosyncratic noise. A factor structure for the covariance of x assumes that x is linear, where $x = \beta z + \epsilon$ (Murray et al., 2013). Without restrictions on β,

$$\begin{aligned} x &= \beta V V^{\intercal} z + \epsilon \\ &= \beta^* z^* + \epsilon \end{aligned}$$

and the variance of x given β is

$$\begin{aligned} \operatorname{cov}(x|\beta) &= \beta\beta^{\intercal} + I^2 \\ &= \beta V V^{\intercal} \beta^{\intercal} + I^2 \\ &= (\beta V)(\beta V)^{\intercal} + I^2 \\ &= \beta^* \beta^{*\intercal} + I^2, \end{aligned}$$

where V is an orthogonal matrix satisfying $VV^{\intercal} = I$, $\beta^* = \beta V$ is an orthogonal transformation of β, and $z^* = V^{\intercal} z$ is an orthogonal transformation of z. Without restrictions, the factor loading matrix and the factor matrix are only identified up to orthogonal transformations, including rotations, reflections, label switching, and permutations i.e. β^* is obtained instead of β.

Several solutions have been proposed to identify the factor loading and factor matrix (1) restrict the matrix to be lower triangular and place parameter restrictions on the factor loading matrix; (2) restrict the matrix to be lower triangular and correct the posterior iterates aposteriori. Geweke and Zhou (1996) were one of the first to propose using a positive diagonal and lower triangular restriction on the factor loading matrix. They show that these restrictions lead to identification of the factor loading matrix and the factor matrix, as the identity matrix becomes the only possible orthogonal transformation (this is equivalent to $V = I$) (refer to Appendix

B.1 to understand the intuition). This solution is the most widely used in the literature (Aguilar and West, 2000; Chib et al., 2006; Murray et al., 2013). On the other hand, Kastner et al. (2017) and Aßmann et al. (2012) show that one can restrict the factor loading matrix to be lower triangular and deal with the sign restrictions aposteriori.

B.1 Intuition for Positive Lower Triangular Condition

The positive and lower triangular restriction (PLT) is important for identification of factor loadings and factors (Geweke and Zhou, 1996; Murray et al., 2013; Aguilar and West, 2000; Chib et al., 2006). To build intuition for why the PLT condition is used for identification, consider a two-dimensional example. Suppose that the factor loading matrix has PLT restriction given by $A = \begin{pmatrix} a & 0 \\ b & c \end{pmatrix}$, where $a, c > 0$. In general, the orthogonal matrix is of the form $V = \begin{pmatrix} e & f \\ g & h \end{pmatrix}$, where $e, f, g, h \in \mathbb{R}$. By definition, the orthogonal transformation needs to satisfy condition B.1:

$$\begin{pmatrix} e & f \\ g & h \end{pmatrix} \begin{pmatrix} e & g \\ f & h \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{B.1}$$

An orthogonal transformation of A with the resulting matrix also having the PLT restriction satisfies condition B.2:

$$\begin{pmatrix} a & 0 \\ b & c \end{pmatrix} \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} a' & 0 \\ b' & c' \end{pmatrix}, \tag{B.2}$$

where $a', c' > 0$.

Condition B.2 implies that, $ae = a'$, $af = 0$, $be+cg = b'$ and $bf+ch = c'$. Now $f = 0$ as $a > 0$. When $f = 0$, condition B.1 implies that $e^2 = 1$, $eg = 0$ and $g^2 + h^2 = 1$. Since $e = \pm 1$, it must be that $g = 0$ and therefore $h^2 = 1$. The orthogonal matrix is therefore, a diagonal matrix of $+1$ or -1. Since a and $ae = a'$ are positive then $e = 1$. Similarly, $h = 1$. Thus, when the PLT restriction holds, the identity matrix is the only possible orthogonal transformation (i.e. $V = I$).

As an extension, I assume that only the lower triangular restriction is imposed and derive the form of the orthogonal transformation to convert A into a matrix with the PLT restriction aposteriori. Suppose that A has no sign restriction prior to running the VBIL algorithm. Since the matrix which is obtained aposteriori should satisfy the PLT restriction, let $a', c' > 0$. From condition B.2, given that $ae = a'$, and the sign of a' is positive, then a and e must have the same sign. Similarly, c and h must have the same sign. This implies that an orthogonal transformation of the form

$$\begin{pmatrix} \mathrm{sign}(a)1 & 0 \\ 0 & \mathrm{sign}(c)1 \end{pmatrix}$$

can be used to convert A into a matrix with PLT restriction aposteriori. I refer to this method as 'sign-transformation' and provide an illustration in Appendix I.

Appendix C
Derivatives

The derivatives in steps 1(b) and 2(b) of the VBIL algorithm are given below. In the thesis, Γ is a vector. More generally, Γ can be a lower triangular matrix of size $R \times \tilde{r}$, for $\tilde{r} < R$. In general, let $\xi = \text{vech}(\Gamma)$ and following previous notation, let d be the diagonal of D. The variational parameter vector is equal to $\lambda = \{\eta^\top, \xi^\top, d^\top\}^\top$. The gradient of $\log(q_\lambda(\theta))$ is $\nabla_\lambda \log(q_\lambda(\theta)) = \{\nabla_\eta \log(q_\lambda(\theta))^\top, \nabla_\xi \log(q_\lambda(\theta))^\top, \nabla_d \log(q_\lambda(\theta))^\top\}^\top$, where

$$\begin{aligned}
&\nabla_\eta \log(q_\lambda(\theta)) = (\Gamma\Gamma^\top + D)^{-1}(\theta - \eta),\\
&\nabla_\xi \log(q_\lambda(\theta)) = \text{vech}\big(-(\Gamma\Gamma^\top + D)^{-1}\Gamma + (\Gamma\Gamma^\top + D)^{-1}(\theta-\eta)(\theta-\eta)^\top(\Gamma\Gamma^\top + D)^{-1}\Gamma)\big),\\
&\nabla_d \log(q_\lambda(\theta)) = \text{diag}\big(-(\Gamma\Gamma^\top + D)^{-1}D + (\Gamma\Gamma^\top + D)^{-1}(\theta-\eta)(\theta-\eta)^\top(\Gamma\Gamma^\top + D)^{-1}D\big).
\end{aligned}$$

Following Ong et al. (2018b), the Woodbury formula can be used to compute the inverse of the covariance matrix faster

$$(\Gamma\Gamma^\top + D)^{-1} = D^{-2} - D^{-2}\Gamma(I + \Gamma D^{-2}\Gamma)^{-1}\Gamma^\top D^{-2}.$$

An additional derivative is required to implement the YJ-transformation for VBIL. The variational parameter vector is equal to $\lambda = \{\eta^\top, \xi^\top, d^\top, \gamma^\top\}^\top$, where $0 < \gamma < 2$. The gradient of $\log(q_\lambda(\theta))$ is

$$\nabla_\lambda \log(q_\lambda(\theta)) = \{\nabla_\eta \log(q_\lambda(\theta))^\top, \nabla_\xi \log(q_\lambda(\theta))^\top, \nabla_d \log(q_\lambda(\theta))^\top, \nabla_\gamma \log(q_\lambda(\theta))^\top\}^\top.$$

The first three derivatives are obtained by replacing θ from above with ψ. The derivative with respect to γ is given by

$$\nabla_\gamma \log(q_\lambda(\theta)) = \Big(-\frac{\partial t_\gamma(\theta)}{\partial \gamma}(\Gamma\Gamma^\top + D)^{-1}(\psi - \eta_\psi) + \big(\frac{\partial t'_{\gamma_1}(\theta_1)}{\partial \gamma_1}\frac{1}{t'_{\gamma_1}(\theta_1)}, \dots, \frac{\partial t'_{\gamma_R}(\theta_R)}{\partial \gamma_R}\frac{1}{t'_{\gamma_R}(\theta_R)}\big)^\top\Big)$$

where

$$\frac{\partial t_\gamma(\theta)}{\partial \gamma} = \text{Diag}\big(\frac{\partial t_{\gamma_1}(\theta_1)}{\partial \gamma_1}, \dots, \frac{\partial t_{\gamma_R}(\theta_R)}{\partial \gamma_R}\big).$$

See Smith et al. (2019) for the derivatives of t_γ corresponding to the YJ-transformation.

Appendix D
Modification of the Genz Algorithm

The likelihood of the Gaussian copula with discrete margins involves computing J integrals for each likelihood term (Equation A). The Genz's algorithm (1992) provides an accurate and efficient solution to numerically compute multivariate normal probabilities of the form

$$\frac{1}{\sqrt{(}|\Sigma|(2\pi)^J}\int_{a_1}^{b_1}\int_{a_2}^{b_2}\dots\int_{a_J}^{b_J} e^{-1/2x^\intercal\Sigma^{-1}x}dx_1,\dots,dx_J, \tag{D.1}$$

where Σ is a symmetric positive definite covariance matrix. To simply the computation, several transformations are used to make the integral simpler to calculate. In the last transformation, the integral is defined on a unit hypercube, which allows the integral to be computed efficiently using standard numerical integration algorithms.

The algorithm specifies the use of a randomised lattice rule, which is a quasi-random Monte Carlo technique, to generate the random numbers used to compute the integral in Equation D.1. Instead of generating random numbers using a randomised lattice rule, I simulate uniform random variables using the *rand* function in MATLAB, which is known as importance sampling with a uniform proposal. The result produces an unbiased estimate of Equation D.1.

In block PM, importance sampling with a uniform proposal allows for easier manipulation of u, a set of random numbers used to compute the likelihood. A block of u can be replaced by simulating independent $[0, 1]$ uniform random numbers from *rand*. See Genz (1992) for more information and the MATLAB code.

Appendix E
Kullback-Leibler Divergence Proof

The variational parameter, λ, can be found by minimising the KL divergence and using stochastic gradient methods. Let $\tilde{p}(\theta, u|y) \propto h(\theta, u)g_M(z|\theta)$, where $h(\theta, u) = \hat{p}_M(y|\theta, u)p(\theta)$ and note that $h(\theta, u)$ is independent of λ.

The KL divergence for VBIL is given by (Tran et al., 2017),

$$\begin{aligned} \mathrm{KL}(q_\lambda(\theta, z)||\tilde{p}(\theta, u|y)) &= \iint \log\big(\frac{q_\lambda(\theta, z)}{\tilde{p}(\theta, u|y)}\big)q_\lambda(\theta, z)\, dz\, d\theta \\ &= \log p(y) - \iint \log\big(\frac{h(\theta, u)g_M(z|\theta)}{q_\lambda(\theta, z)}\big)q_\lambda(\theta, z)\, dz\, d\theta. \end{aligned} \tag{E.1}$$

Minimising the KL divergence is equivalent to maximising the lower bound,

$$\begin{aligned} LB &= \iint \log\big(\frac{h(\theta, u)g_M(z|\theta)}{q_\lambda(\theta, z)}\big)q_\lambda(\theta, z)\, dz\, d\theta \\ \nabla_\lambda LB &= \iint \big(- q_\lambda(\theta)^{-1}\nabla_\lambda q_\lambda(\theta)\big)q_\lambda(\theta)g(z|\theta) + \\ &\qquad \big(\log(h(\theta, u)g_M(z|\theta)) - \log q_\lambda(\theta, z)\big)\nabla_\lambda q_\lambda(\theta)g(z|\theta)\, dz\, d\theta \\ &= \iint \big(\log(h(\theta, u)g_M(z|\theta)) - \log(q_\lambda(\theta)g(z|\theta))\big)\nabla_\lambda q_\lambda(\theta)g(z|\theta)\, dz\, d\theta \\ &= \iint \big(\log(h(\theta, u)g_M(z|\theta)) - \log(q_\lambda(\theta)g(z|\theta))\big)\nabla_\lambda \log(q_\lambda(\theta))q_\lambda(\theta)g(z|\theta)\, dz\, d\theta \\ &= \mathbb{E}_{\theta, z|\theta}\big[\big(\log(\hat{p}_M(y|\theta, u)p(\theta)) - \log q_\lambda(\theta)\big)\nabla_\lambda \log(q_\lambda(\theta))\big]. \end{aligned} \tag{E.2}$$

The proof involves using $\iint \nabla_\lambda q_\lambda(\theta)g_M(z|\theta)\, dz\, d\theta = 0$ and $\nabla_\lambda \log q_\lambda(\theta) = q_\lambda(\theta)^{-1}\nabla q_\lambda(\theta)$.

Appendix F
Comparison of VBIL and VBDA

Figure F.1 (Blei et al., 2017) shows the amount of posterior information captured in VBDA when independence is assumed between $q_{\lambda_\alpha}(\theta)$ and $q_{\lambda_\beta}(u)$. In the figure, the information captured is in green and is labelled as 'mean-field approximation'.

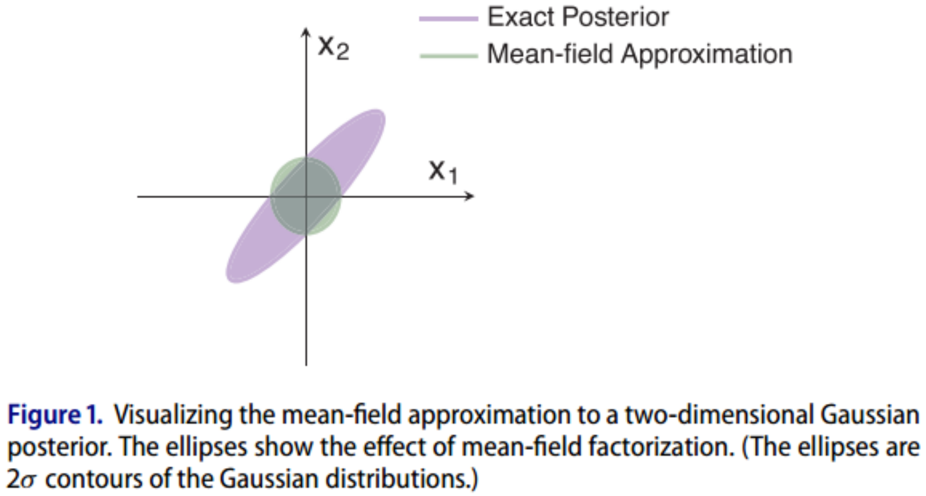
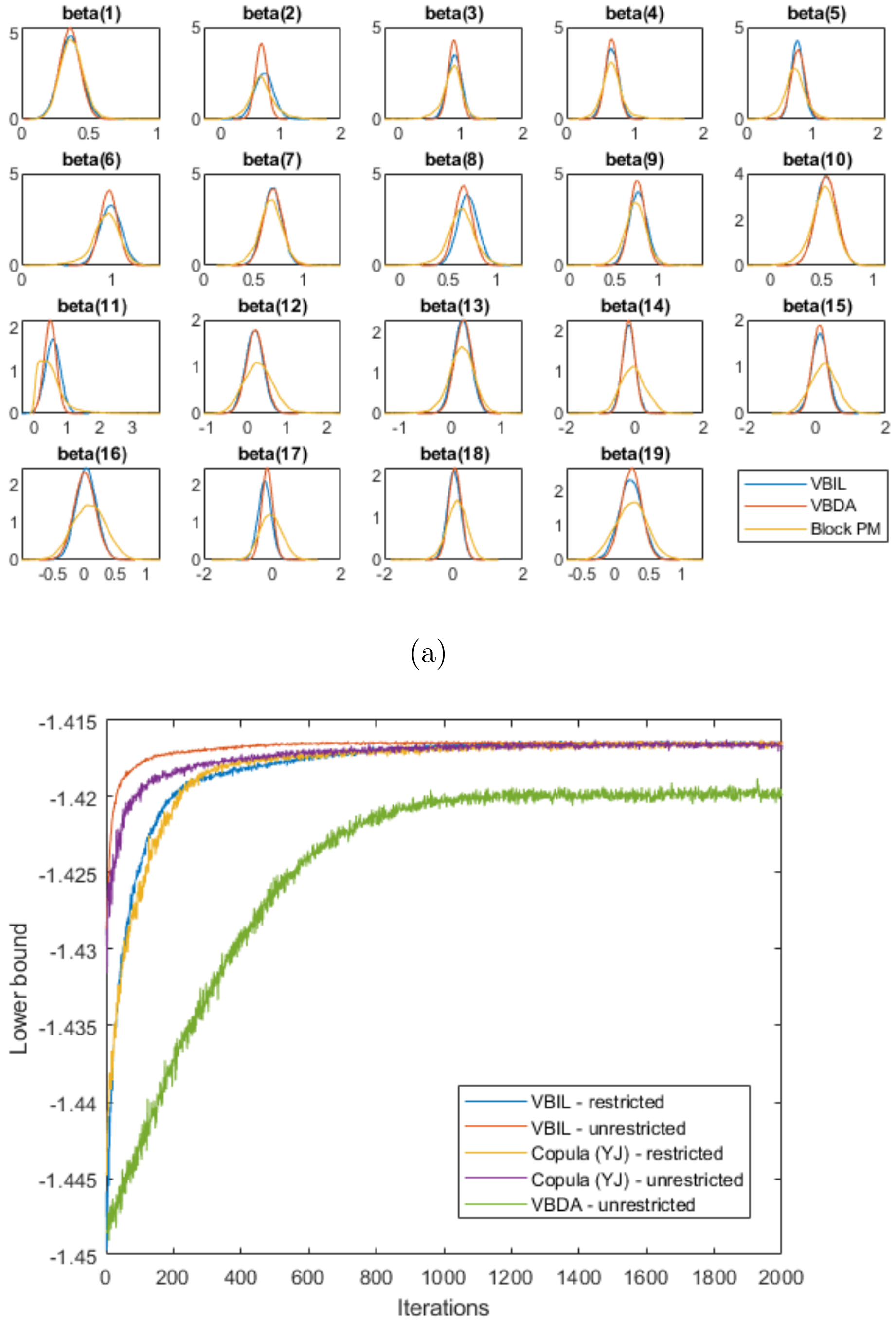


Figure 1. Visualizing the mean-field approximation to a two-dimensional Gaussian posterior. The ellipses show the effect of mean-field factorization. (The ellipses are 2σ contours of the Gaussian distributions.)

Figure F.1: Comparison of mean-field approximation and the exact posterior.

F.1 Ordinal Results



(a)

(b)

Figure F.2: Plots of the Gaussian copula parameters based on ordinal data simulations with dimensions $J = 10$ and $N = 250$: (a) Kernel densities of VBIL and VBDA against block PM; (b) Lower bounds of VBIL and VBDA against the number of iterations.

Appendix G
Adaptive Learning Rates

The learning rate can be set adaptively using different methods, such as ADADELTA (Zeiler, 2012) and Ranganath's adaptive learning rate (Ranganath et al., 2013).

G.0.1 ADADELTA

ADADELTA involves updating the step size $\Delta\lambda_i^{(t)}$ for each element in the vector of variational parameters λ through the equation

$$\lambda_i^{(t+1)} = \lambda_i^{(t)} + \Delta\lambda_i^{(t)},$$

with $\Delta\lambda_i^{(t)} = \rho_i^{(t)} g_{\lambda_i}^{(t)}$ and $g_{\lambda_i}^{(t)}$ denotes the ith element of $\nabla_\lambda \widehat{\text{LB}}(\lambda_t)$. The learning rate is given by

$$\rho_i^{(t)} = \frac{\sqrt{\mathbb{E}(\Delta_{\lambda_i}^2)^{t-1} + \epsilon}}{\sqrt{\mathbb{E}(g_{\lambda_i}^2)^t + \epsilon}}.$$

The terms $\mathbb{E}(\Delta_{\lambda_i}^2)^t$ and $\mathbb{E}(g_{\lambda_i}^2)^t$ are updated recursively as

$$\mathbb{E}[\Delta_{\lambda_i}^2]_t = \zeta \mathbb{E}[\Delta_{\lambda_i}^2]_{t-1} + (1-\zeta)\Delta_{\lambda}{}_i^{(t)2}$$

$$\mathbb{E}[g_{\lambda_i}^2]_t = \zeta \mathbb{E}[g_{\lambda_i}^2]_{t-1} + (1-\zeta) g_{\lambda_i}^{(t)2}.$$

Following Ong et al. (2017), I set $\epsilon = 10^{-6}$, $\zeta = 0.95$, $\mathbb{E}[\Delta_{\lambda_i}^2]^{(0)} = 0$ and $\mathbb{E}[g_{\lambda_i}^2]^{(0)} = 0$.

G.0.2 Ranganath

Ranganath's adaptive learning rate (Ranganath et al., 2013) involves estimating the natural gradient of the lower bound, $\hat{n}_t = I_F(\lambda^{(t)})^{-1}\nabla_\lambda \widehat{\text{LB}}(\lambda^{(t)})$. A running average of the values of $\hat{n_t}$ and $\hat{n_t}^\intercal \hat{n_t}$ are equal to

$$\bar{n}_t = (1-\alpha_t)\bar{n}_{t-1} + \alpha_t \hat{n}_t$$

$$\bar{c}_t = (1-\alpha_t)\bar{c}_{t-1} + \alpha_t \hat{n}_t^\intercal \hat{n}_t,$$

where α_t is a discounting factor. The learning rate $\rho_t = \frac{\bar{n}_t^\intercal \bar{n}_t}{\bar{c}_t}$ and the discounting factor are updated adaptively as $\alpha_{t+1}^{-1} = \alpha_t^{-1}(1-\rho_t) + 1$. The initial values $\hat{n}_0$ and

$\hat{c}_0$ are set equal to K independent computations of the gradient estimates at the initial starting values of the variational parameters. The parameter α_0 is initialised as $1/K$. For the intuition of how this method works, see Ong et al. (2018a).

Appendix H
Trace Plots Examples

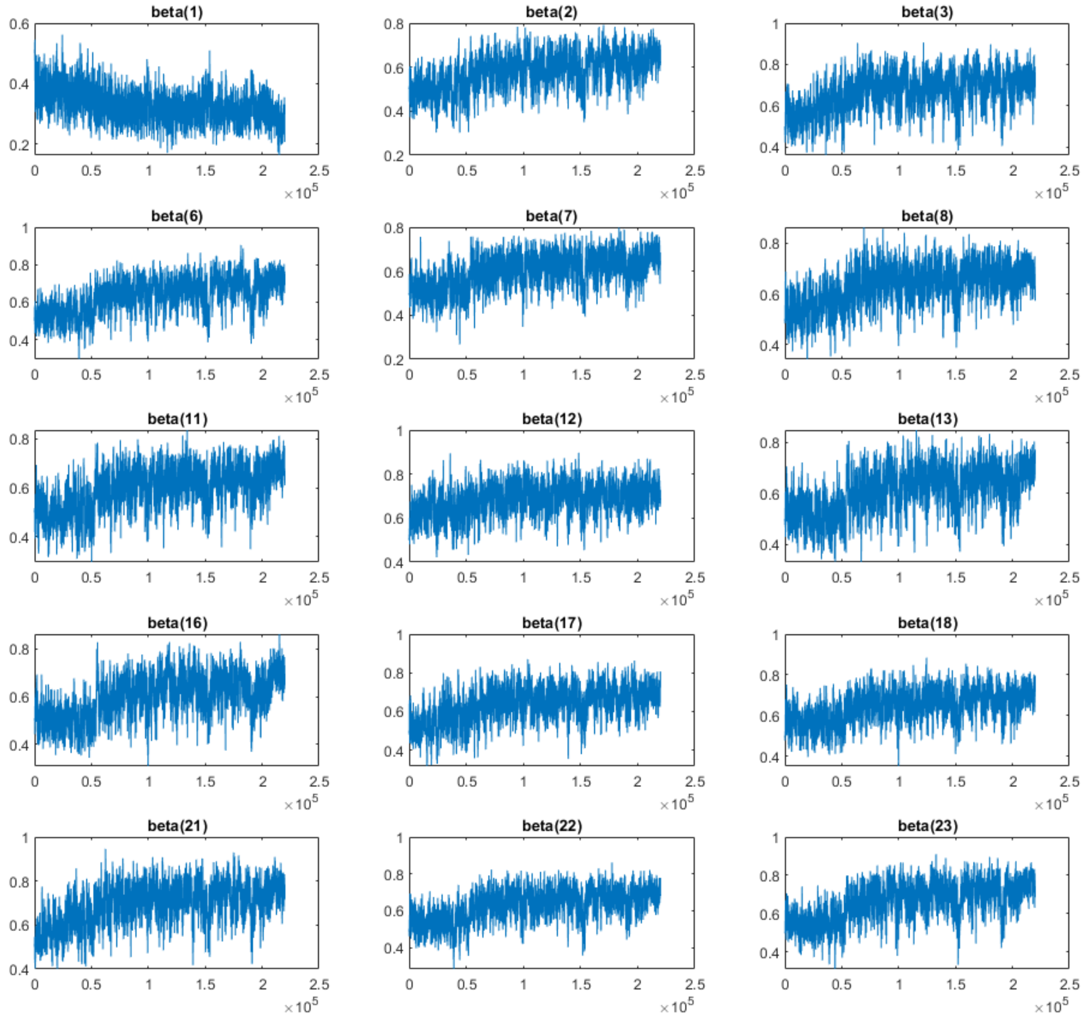


Figure H.1: Trace plots of select βs of the Gaussian copula parameters for ordinal data simulations with dimensions $J = 37$ and $N = 1200$. The initial values of the chain are the same as the initial values of VB algorithms. The trace plots do not look stationary.

Appendix I
YJ and Gaussian Copula Parameter Transformation Examples

This section presents two extensions. The first provides results of ordinal and Bernoulli simulations modelled using a Gaussian copula with YJ-transformations, hereafter referred to as Copula (YJ) (Section 5.4). The second extension provides evidence of whether the choice of (sign) transformation for β has an effect on VBIL and Copula (YJ) results. This addresses a remark made by Tan and Nott (2018) who found, but did not provide evidence, that different transformations of the variational distribution parameters yield different results. For this extension, two types of transformations are considered: the log transformation from Section 2.2.3 and the sign-transformation from Section B.1. In addition, the case where β is unrestricted (no sign transformation) is also considered. For block PM, the sign restriction of β is imposed through log transformations for all simulations.

Figures I.1 and I.2 (a) and (b) plot the ordinal simulation results for $J = 10$ and $N = 250$. For these figures, the sign restriction for VBIL and Copula (YJ) is imposed through log transformations and is referred to as ‘restricted’ in Figure I.2 (b). Both the VBIL and Copula (YJ) approximations provide a good fit to the true posterior. Figure I.2 (a) shows that Copula (YJ) provides a similar fit to VBIL in terms of mean but underestimates the variance more. For Copula (YJ), a small improvement in skewness is observed. The lower bounds of VBIL and Copula (YJ) for both restricted (log transformed) and unrestricted β converge to the same value. In this example, the sign-transformation results are similar to the log transformation results and are not included in the thesis.

Plots of the Bernoulli simulations for dimensions $J = 10$ and $N = 250$ are included below. Compared to the ordinal simulations, imposing sign restrictions through log transformations for the Bernoulli simulations lead to a worse fit for VBIL and Copula (YJ) (Figure I.3). Improvement in the posterior fit is observed for VBIL when β is unrestricted (Figure I.4). Similarly to the ordinal simulations, Copula (YJ) underestimates the variance more than VBIL and has a small improvement in skewness. In addition, the lower bounds of VBIL and Copula (YJ) for both the log transformation (restricted) and unrestricted case converge to the same value (Figure I.5). A further improvement in the posterior fit of the mean is observed for both VBIL and Copula (YJ) when a sign-transformation is used (Figure I.6). However, larger overestimation and underestimation of the posterior variance and skewness for Copula (YJ) are also observed.

Overall, I find that Copula (YJ) provides similar results to VBIL with small improvement in skewness. Therefore, there is little added benefit in using the YJ-transformation for the examples considered. For both VBIL and Copula (YJ), the log transformation, sign-transformation and the case of unrestricted β provide similar results in the ordinal simulation. In contrast, despite the lower bound of VBIL and Copula (YJ) converging to the same value, different transformations yield different posterior results in the Bernoulli simulations. Therefore, these simulations provide supporting evidence to the remark made by Tan and Nott (2018) that different transformations of the variational distribution parameters yield different results.

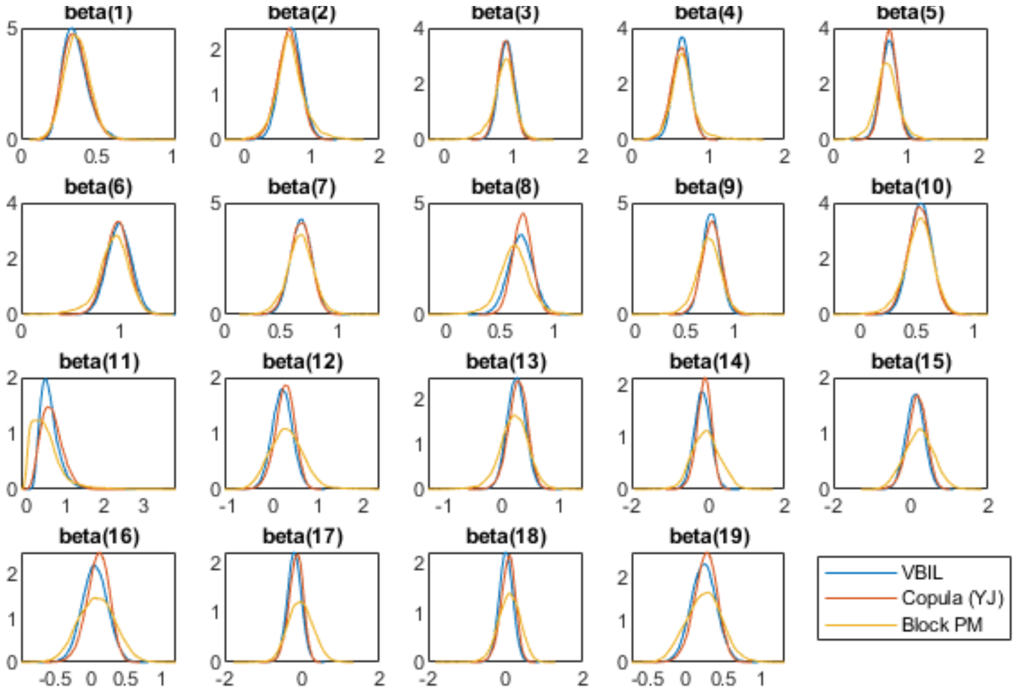
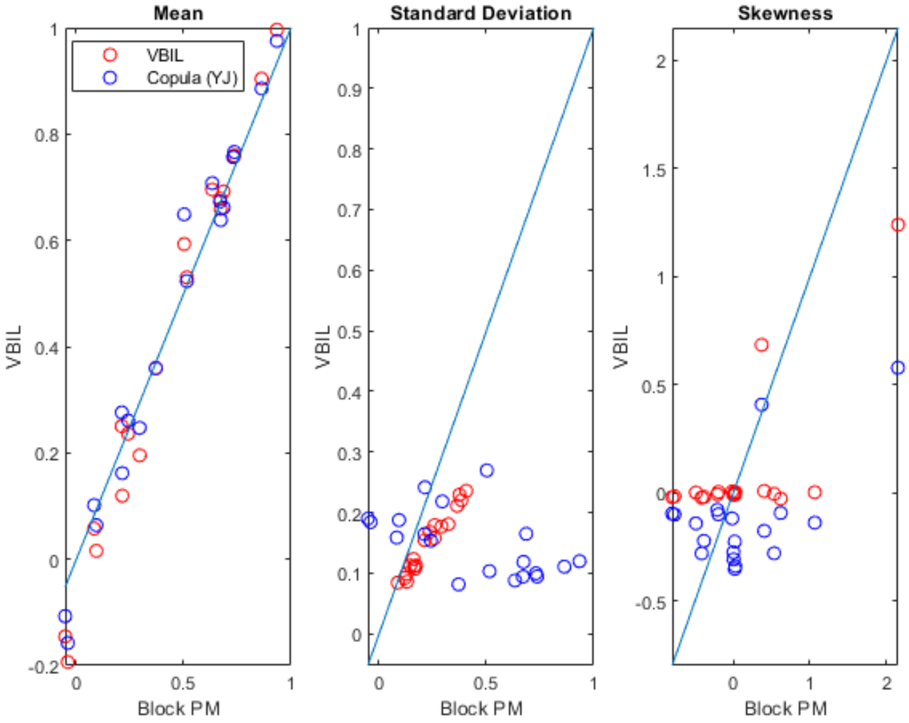
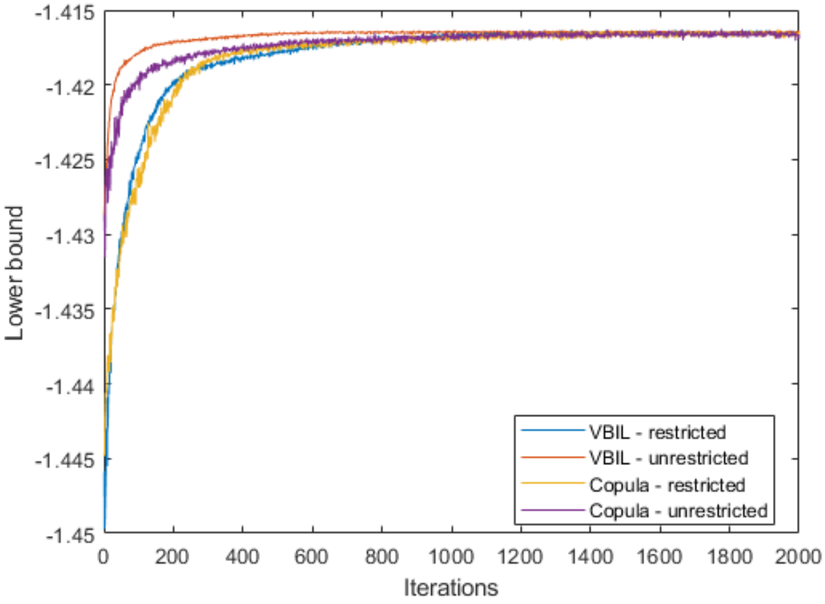


Figure I.1: Kernel density plots based on ordinal data simulations with dimensions $J = 10$ and $N = 250$ of VBIL, Copula (YJ) and block PM for log transformed β.

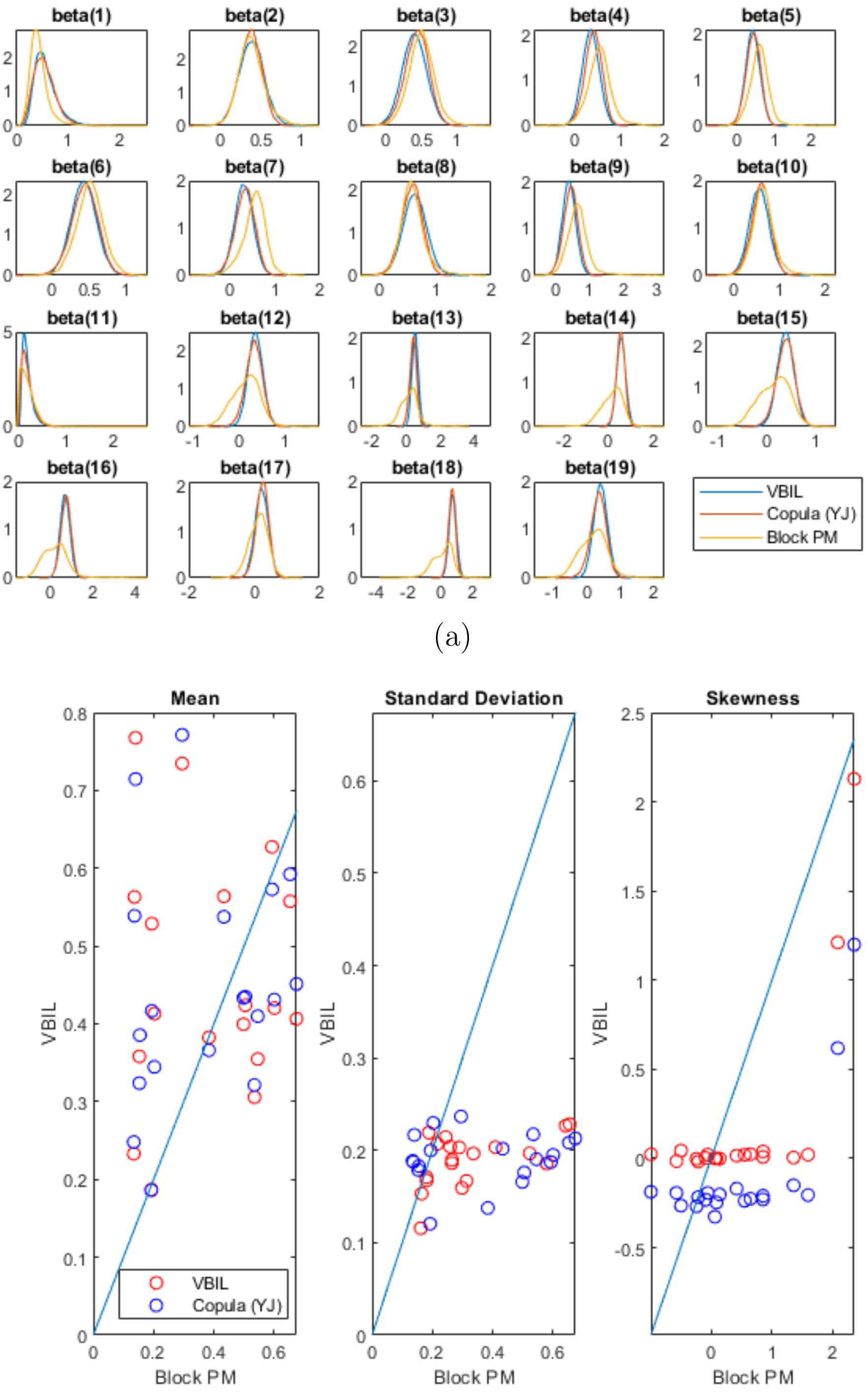


(a)



(b)

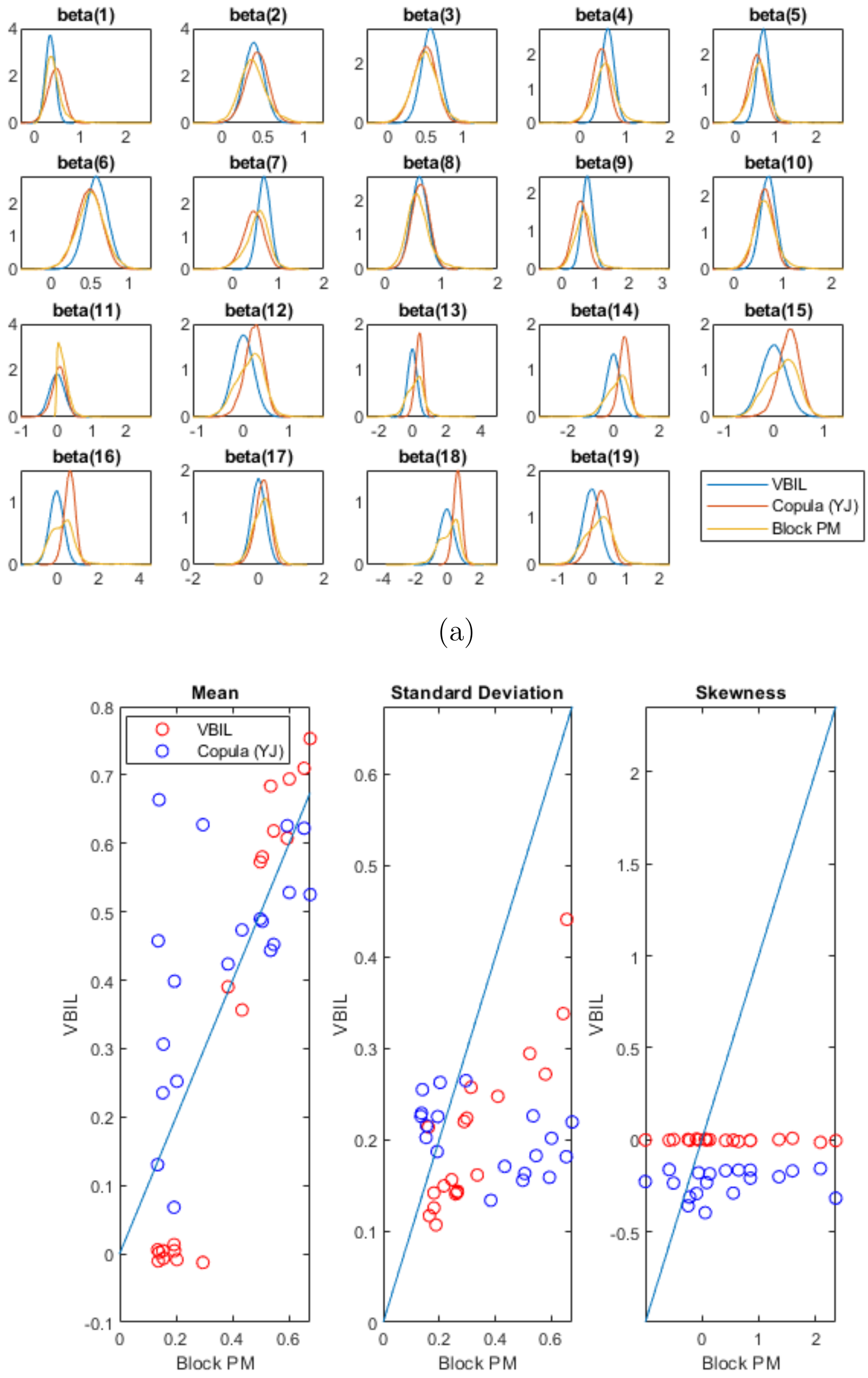
Figure I.2: Plots based on ordinal data simulations with dimensions $J = 10$ and $N = 250$: (a) Mean, standard deviation and skewness of VBIL and Copula (YJ) against block PM for log transformed β; (b) Lower bounds of VBIL and Copula (YJ) for unrestricted and restricted (log transformed) β against the number of iterations. The differences at the start reflect different choices of initial starting values. All four lower bounds converge to the same value.



(a)

(b)

Figure I.3: Plots based on Bernoulli data simulations with dimensions $J = 10$ and $N = 250$: (a) Kernel density plots of VBIL, Copula (YJ) and block PM for log transformed β; (b) Mean, standard deviation and skewness of VBIL and Copula (YJ) against block PM for log transformed β.



(a)

(b)

Figure I.4: Plots based on Bernoulli data simulations with dimensions $J = 10$ and $N = 250$: (a) Kernel density plots of VBIL, Copula (YJ) and block PM for unrestricted β; (b) Mean, standard deviation and skewness of VBIL and Copula (YJ) against block PM for unrestricted β.

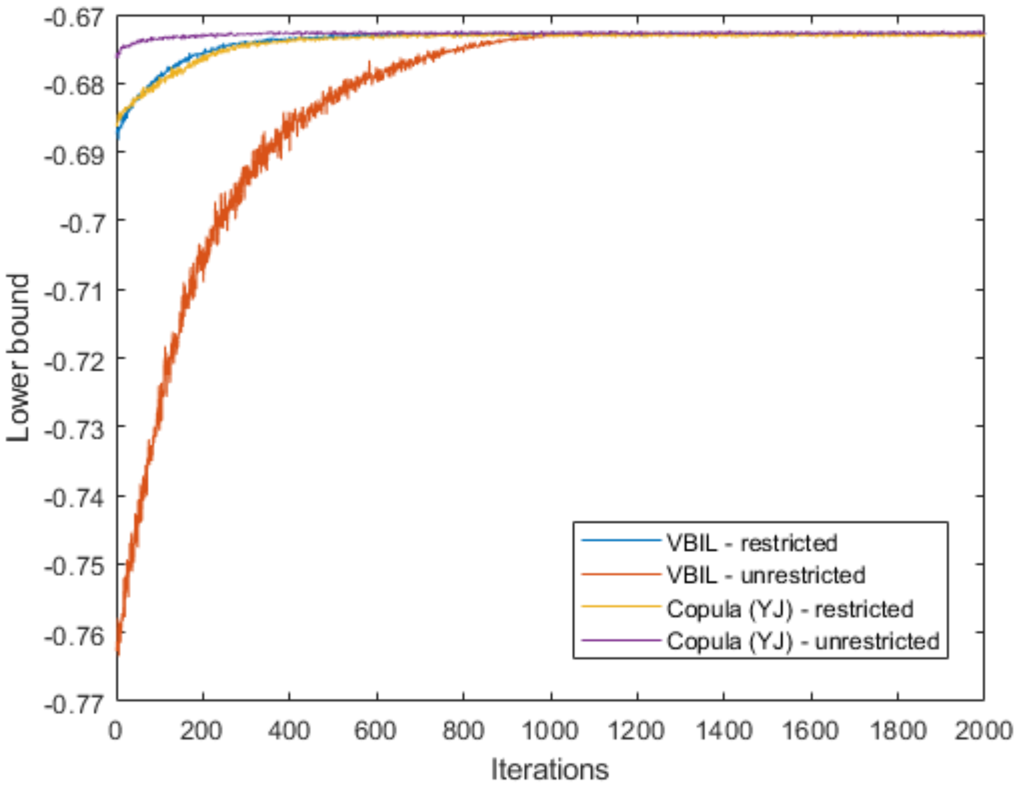
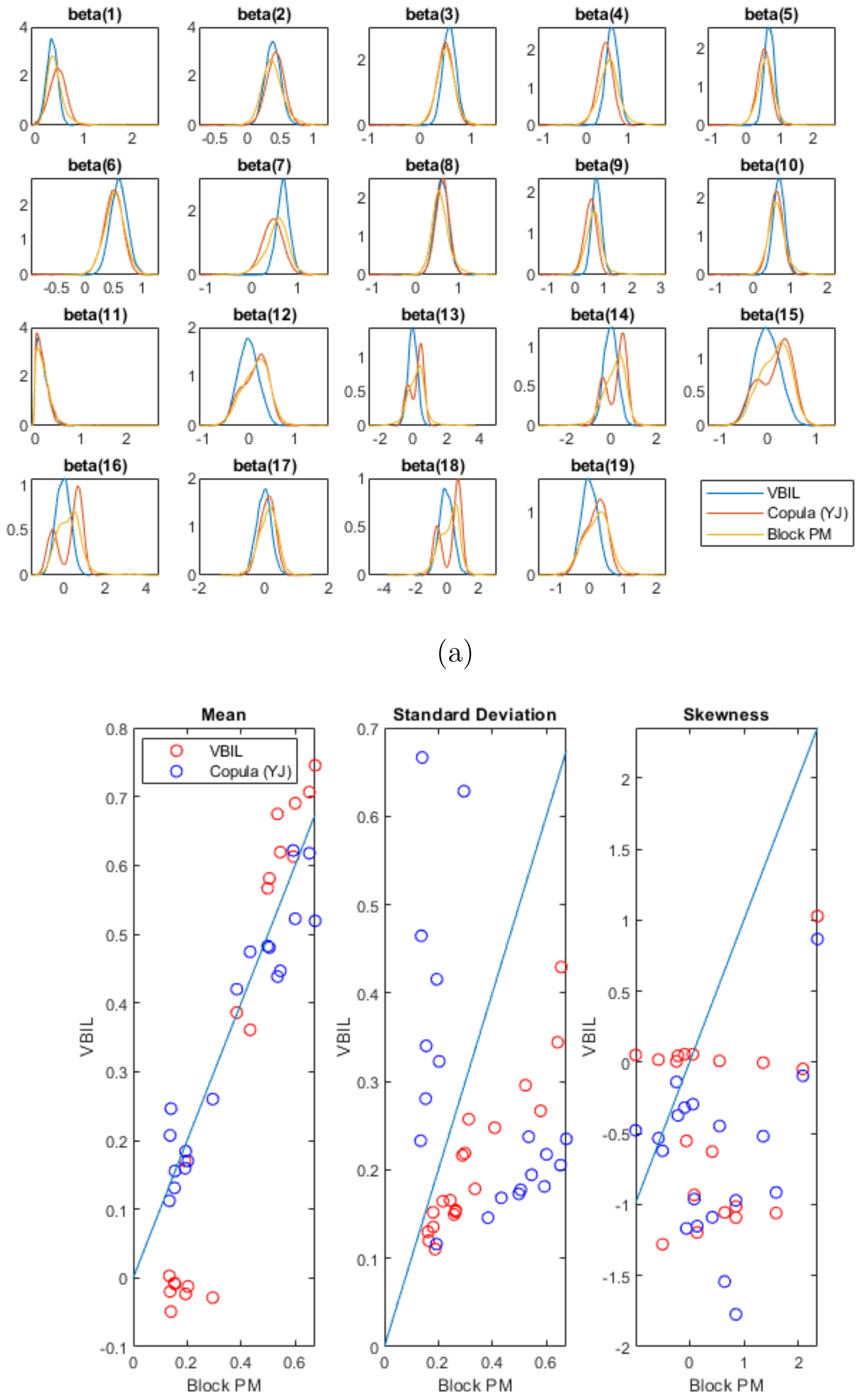


Figure I.5: Lower bound based on Bernoulli data simulations with dimensions $J = 10$ and $N = 250$ of VBIL and Copula (YJ) for unrestricted and restricted (log transformed) β against the number of iterations. The differences at the start reflect different choices of initial starting values. All four lower bounds converge to the same value.



(a)

(b)

Figure I.6: Plots based on Bernoulli data simulations with dimensions $J = 10$ and $N = 250$: (a) Kernel density plots of VBIL, Copula (YJ) and block PM for sign-transformed β; (b) Mean, standard deviation and skewness of VBIL and Copula (YJ) against block PM for sign-transformed β.

Appendix J
Variational Bayes Data Augmentation (VBDA)

Following Smith and Khaled (2012), Loaiza-Maya and Smith (2019) introduce an extra J-dimensional vector, v, into the posterior to make the likelihood tractable. Let $v_j = F_j(y_j)$. Using Bayes rule, the joint posterior is

$$\begin{aligned} p(\theta, v|y) &= p(y|v,\theta)c(v|\theta)p(\theta)/p(y) \\ &= \prod_j \mathbb{I}(F_j(y_j^-) \leq v_j < F_j(y_j))c(v|\theta)p(\theta)/p(y), \end{aligned} \tag{J.1}$$

where $c(v|\theta)$ is a copula density, $F_j(.)$ is the CDF of the marginal distribution of $Y_{\cdot j}$ and $\mathbb{I}$ is an indicator function equal to one if v_j is between $F_j(y_j^-)$ and $F_j(y_j)$ and zero otherwise. The notation $Y_{\cdot j}$ refers to the jth column of the data matrix. Since v_j is a function of y_j, given v_j, the chance that v_j falls inside the interval $[F_j(y_j^-), F_j(y_j)]$ is known. The indicator function captures this relationship.

The Gaussian copula density is

$$\begin{aligned} c(v|\theta) &= \frac{\partial^J}{\partial v_1 ... \partial v_J} \Phi_\Lambda(z^*) \\ &= \phi_\Lambda(z^*) \prod_{j=1}^{J} \frac{1}{\phi(z_j^*)} \\ &= \phi_\Lambda(z^*)/\phi(z^*) \\ &= \det(\Lambda)^{-1/2} e^{-1/2 z^{*\intercal}(\Lambda^{-1}-I)z^*}, \end{aligned} \tag{J.2}$$

where $z_j^* = \Phi^{-1}(v_j), j = 1, ..., J$ and Λ is the correlation matrix of the copula.

To approximate the posterior, Loaiza-Maya and Smith (2019) choose the corresponding variational distribution to be $q_\lambda(v,\theta) = q_{\lambda_\alpha}(\theta) q_{\lambda_\beta}(v)$, where $\lambda = \{\lambda_\alpha, \lambda_\beta\}$. Here, for ease of computation, v and θ are assumed to be independent.

The variational distribution for θ, $q_{\lambda_\alpha}(\theta)$, follows a multivariate Gaussian distribution given by $N(\mu_\alpha, B_\alpha B_\alpha^\intercal + D_\alpha^2)$ with a factor representation for the covariance matrix (see Section K for why factor structures are used). The authors show that a small number of factors in B_α is adequate and that there is little improvement in accuracy when the number of factors is increased.

Loaiza-Maya and Smith (2019) propose 3 variational distributions VA1, VA2 and VA3 to model $q_{\lambda_\beta}(v)$. VA1 assumes that $q_{\lambda_\beta}(v)$ is a uniform distribution, while VA2 assumes that v follows a transformed Gaussian distribution with a diagonal covariance matrix. As a generalisation of VA2, VA3 also assumes that v follows a transformed Gaussian distribution, but with a band one lower triangular Cholesky factor

$$q_{\lambda_\beta}(v) = \frac{\phi_J(z; \eta_\beta, (L_\beta L_\beta^\intercal)^{-1})}{\prod_{j=1}^J (b_j - a_j)\phi(z_j)} \tag{J.3}$$

where η_β is a mean vector, L_β is a band one lower triangular Cholesky factor of the inverse covariance matrix,

$$Z_j = \Phi^{-1}(\frac{V_j - a_j}{b_j - a_j}),$$

and

$$Z_j \sim N(\eta_\beta, (L_\beta L_\beta^\intercal)^{-1}).$$

VA2 is a special case of VA3 when the Cholesky factor is diagonal. The authors remark that when $(b_j - a_j)$ goes to zero for all j, the VA1, VA2 and VA3 approximations become exact.

In the simulation studies, I only consider VA3 for modelling $q_{\lambda_\beta}(v)$ because VA2 is a subset of VA3. I do not consider VA1 because it assumes the data dimensions are independent, which is unsuitable for modelling cross-sectional data. As an example, consider a cross-sectional dataset with N individuals and their responses to J survey questions. It is unlikely, that an individual's responses to the J survey questions are uncorrelated.

See Loaiza-Maya and Smith (2019) for the derivatives of $q_{\lambda_\alpha}(\theta)$.

Appendix K
Covariance Factorisations of the Variational distribution

Papers in the literature propose to reduce the number of parameters in the covariance matrix of a Gaussian variational distribution by assuming it has a particular structure. For example, Titsias and Lázaro-Gredilla (2014) reduce the number of parameters from J^2 to J by assuming the covariance matrix has a diagonal structure. Challis and Barber (2013) assume a banded covariance matrix, where the main diagonal and a fixed number of diagonals above and below the main diagonal are non-zero. While these covariance structures may be satisfactory for certain applications, zero covariance under a Gaussian distribution implies that two variables are independent. For the variational distribution of θ specified in Section 5.3.1, assuming a diagonal or banded covariance matrix is a strong assumption because it implies that some of the transformed factor loadings are independent. Since factor loadings play a similar role as linear regression coefficients, factor loadings and their transformations (θ) are likely to be correlated. Not accounting for the correlation between the factor loadings leads to underestimation of the posterior variance (Blei et al., 2017).

To minimise the loss of information, an alternative is to impose a factor structure on the covariance matrix. Variational distributions with factor structure covariance matrices have been used by Seeger (2000), Loaiza-Maya and Smith (2019) and Ong et al. (2018b). A factor structure reduces the number of parameters and does not assume independence between variables. In addition, when relationships between the variables are known, such as in random effects or state space models, it is possible to impose a structure on a transformed covariance matrix to reflect the relationship without losing information. For example, Tan and Nott (2018) and Tan et al. (2019) impose sparsity on the inverse covariance matrix to reflect conditional independence when modelling random effects using Generalised Linear Mixed Models and state space models. Another structure which does not lead to a loss of information is the full cholesky factor of the inverse covariance matrix. Although compared to factor structures, the number of parameters to be estimated is generally much larger, requiring more computational effort. I consider both the factor structure of the covariance matrix and the full Choleksy factor of the inverse covariance matrix in Section 5.3.1.

Appendix L
Appendix for HILDA

L.1 Data

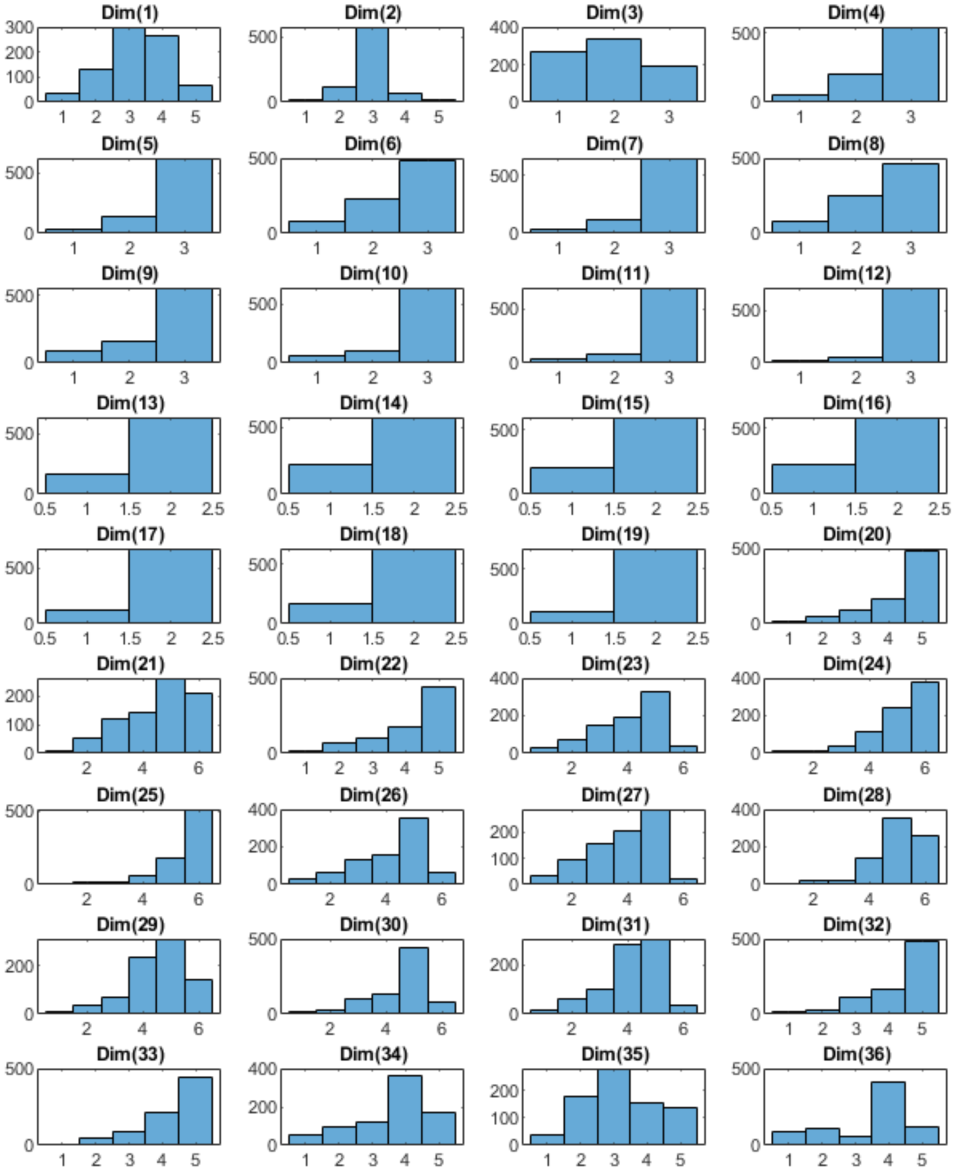


Figure L.1: 2004 survey histograms of the 36 dimensions of the scq.

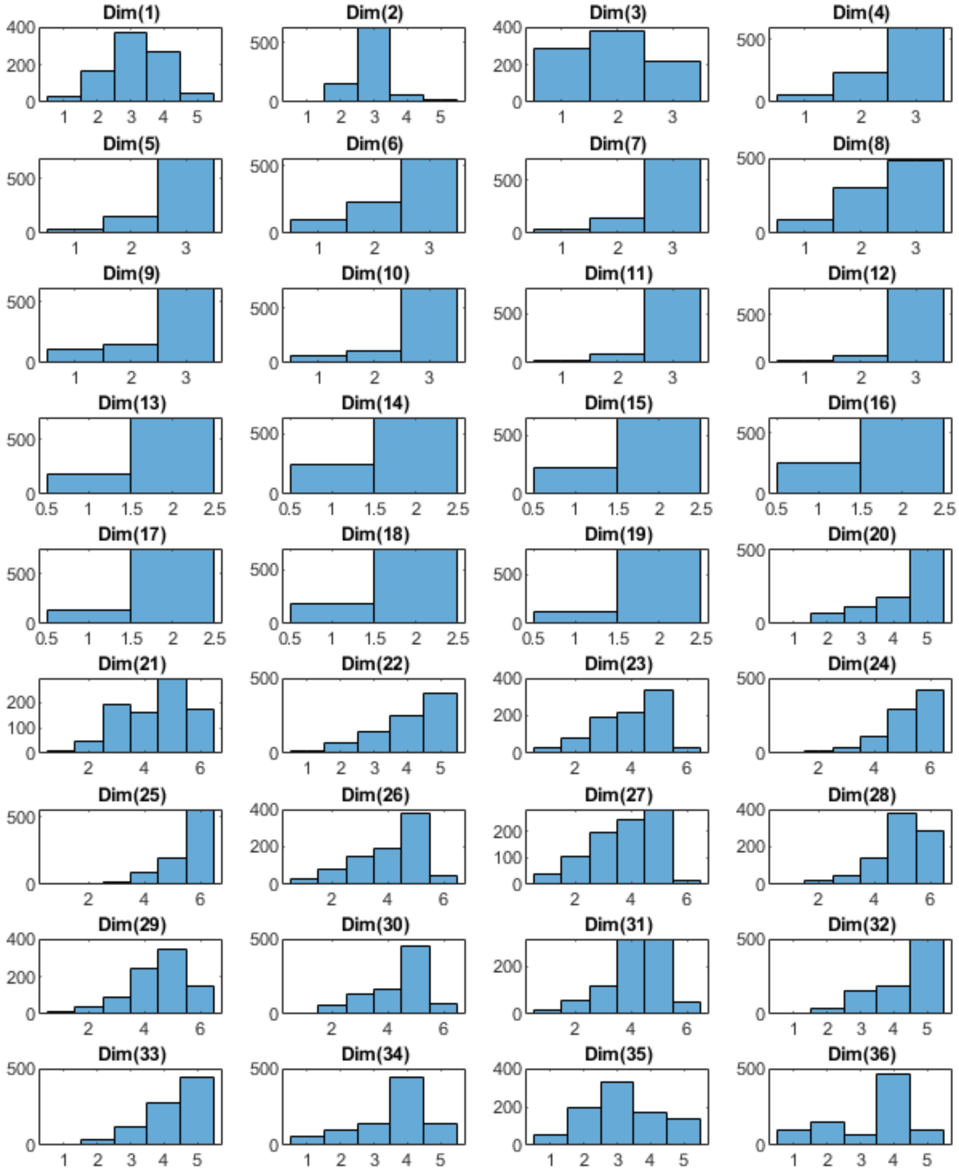


Figure L.2: 2010 survey: histograms of the 36 dimensions of the scq.

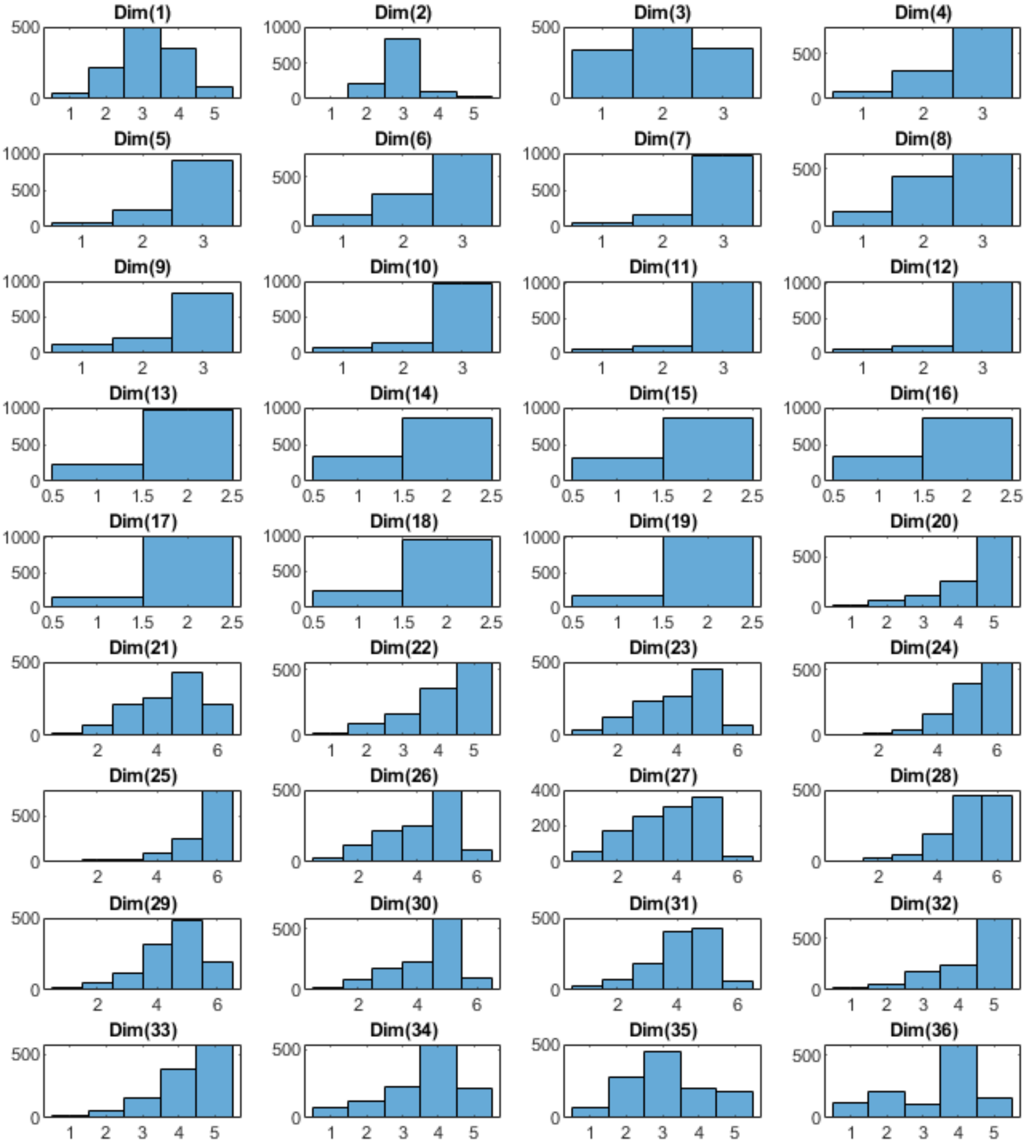


Figure L.3: 2016 survey: histograms of the 36 dimensions of the scq.

L.2 Choosing the questions for the first set

I exclude questions that are similar and questions with a response that implies the response of another question with almost a probability of one. Questions that are too similar provide little additional information about the respondents health. Like regression, selecting variables which are almost orthogonal is optimal, although the variables can still be correlated. As an example of questions that are too similar, consider the two questions: during the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbours, or groups?; and during the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting friends, relatives, etc.)? Both questions are similar in that they ask about the impact of physical health and emotional problems, albeit slightly different. The first question is aimed at the impact on their ability to do normal activities, while the second question is related to time with no specification of whether it is normal or general activities. Despite the slight differences, the differences are not large enough to be of interest as an additional dimension.

The second type of questions that are excluded outlines a subset problem. I adopt a general rule where the most severe condition given a subset of questions is chosen. For example, three questions relate to walking, including being able to walk more than one kilometre, walk for half a kilometre, and walk 100 metres. Out of the three, I retain the last question because not being able to walk 100 metres is more severe than the other two, and if a respondent is limited a lot in their ability to walk 100 metres, they are also unlikely able to walk one or more kilometres.

L.3 Weights and Cut-off

Category	Dimension	Weight (Set 1)	Weight (Set 2)	Cut-off
Physical health	Vigorous activities (such as running, lifting heavy objects, participating in strenuous sports)	0.25	1	1
	Moderate activities (such as moving a table, pushing a vacuum cleaner, bowling or playing golf)	1	1	1
	Lifting or carrying groceries	1.5	1	1
	Climbing one flight of stairs	0.5	1	1
	Bending, kneeling, or stooping	1	1	1
	Walking 100 metres	0.75	1	1
	Bathing or dressing yourself	2	1	1
	Severe bodily pain	2	1	1
Mental health	Felt full of life	0.5	1	2
	Been a nervous person	0.5	1	3
	Felt so down in the dumps that nothing could cheer you up	2	1	3
	Did not feel calm and peaceful	0.5	1	2
	Felt worn out	0.5	1	3
	Not been a happy person	1	1	2
Physical health on work	Cut down the amount of time you spent on work or other activities		1	1
	Accomplished less than you would like		1	1
	Limited in the kind of work or other activities		1	1
	Had difficulty performing the work or other activities (for example, it took extra effort)		1	1
	Pain interfering with your normal work (including both work outside the home and housework)		1	2

Category	Dimension	Weight (Set 1)	Weight (Set 2)	Cut-off
Mental health on work	Cut down the amount of time you spent on work or other activities		1	1
	Accomplished less than you would like		1	1
	Didn't do work or other activities as carefully as usual		1	1
Other	In general, would you say your health is		1	1
	Compared to one year ago, how would you rate your health in general now?		1	2
	Climbing several flights of stairs		1	1
	Walking more than one kilometre		1	1
	Walking half a kilometre		1	1
	Physical health and emotional problems interfered with normal social activities		1	2
	Did not have a lot of energy		1	2
	Felt down		1	3
	Felt tired		1	3
	How much of the time have physical health and emotional problems interfered with social activities		1	3
	Gets sick a little easier than other people		1	2
	Healthy as anybody I know		1	2
	Expect health to get worse		1	3
	Health is excellent		1	3

L.4 Comparison of MCMC and VB Kernel Densities

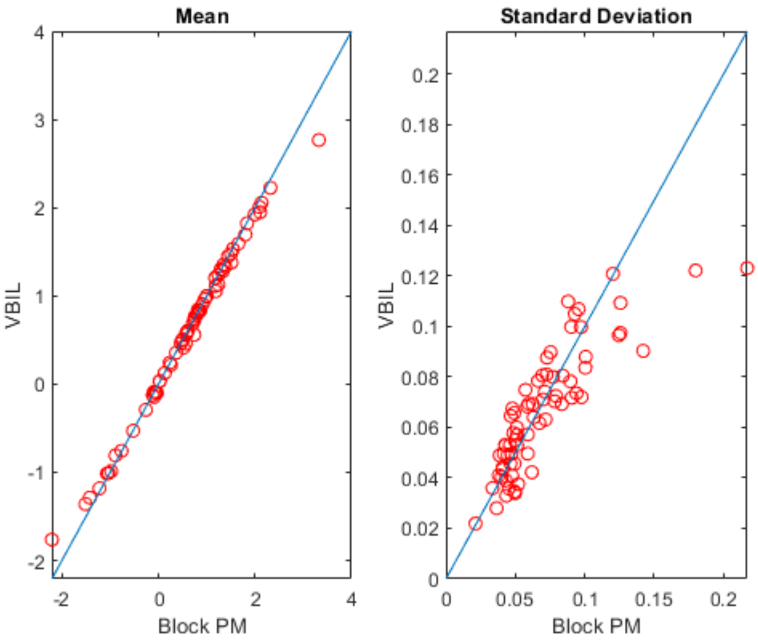


Figure L.4: Mean and standard deviation plot of Gaussian copula parameters based on HILDA 2016 survey. Results show VBIL estimates against block PM.

L.5 MHI results for the Weighted Deprivation Matrix

Table L.2: M_0 posterior means and credibility estimates

k	Year	M_0	Posterior mean (M_0)	Credibility interval (M_0)
3	2004	0.046	0.041	[0.032, 0.051]
	2010	0.045	0.041	[0.032, 0.050]
	2016	0.045	0.044	[0.036, 0.052]
5	2004	0.023	0.023	[0.016, 0.031]
	2010	0.028	0.023	[0.016, 0.031]
	2016	0.029	0.025	[0.019, 0.032]
7	2004	0.009	0.011	[0.005, 0.018]
	2010	0.011	0.012	[0.007, 0.019]
	2016	0.017	0.013	[0.008, 0.019]

Bibliography

Aguilar, O. and West, M. (2000). Bayesian dynamic factor models and portfolio allocation. *Journal of Business & Economic Statistics*, 18(3):338–357.

Alkire, S., Conconi, A., Pinilla-Roncancio, M., and Vaz, A. (2018). How to build a national multidimensional poverty index (MPI): using the MPI to inform the SDGs.

Alkire, S. and Foster, J. (2011). Counting and multidimensional poverty measurement. *Journal of public economics*, 95(7-8):476–487.

Alkire, S., Roche, J. M., and Vaz, A. (2017). Changes over time in multidimensional poverty: methodology and results for 34 countries. *World Development*, 94:232–249.

Amari, S.-I. (1998). Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276.

Andrieu, C., Roberts, G. O., et al. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725.

Aßmann, C., Boysen-Hogrefe, J., and Pape, M. (2012). The directional identification problem in Bayesian factor analysis: an ex-post approach. Technical report, Economics Working Paper.

Australian Institute of Health and Welfare (2019). Health expenditure australia 2017–18.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. *Journal of public health*, 27(3):281–291.

Challis, E. and Barber, D. (2013). Gaussian Kullback-Leibler approximate inference. *The Journal of Machine Learning Research*, 14(1):2239–2286.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335.

Chib, S., Nardari, F., and Shephard, N. (2006). Analysis of high dimensional multivariate stochastic volatility models. *Journal of Econometrics*, 134(2):341–371.

Choroś, B., Ibragimov, R., and Permiakova, E. (2010). Copula estimation. In *Copula theory and its applications*, pages 77–91. Springer.

Conlon, A., Taylor, J., and Elliott, M. (2017). Surrogacy assessment using principal stratification and a Gaussian copula model. *Statistical methods in medical research*, 26(1):88–107.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904.

Dang, K.-D., Quiroz, M., Kohn, R., Tran, M.-N., and Villani, M. (2019). Hamiltonian Monte Carlo with energy conserving subsampling. *Journal of machine learning research*, 20(100):1–31.

Deligiannidis, G., Doucet, A., and Pitt, M. K. (2018). The correlated pseudo-marginal method. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):839–870.

Department of Health (2018). National women's health strategy 2020-2030.

Dunson, D. B. (2010). Flexible Bayes regression of epidemiologic data. *The Oxford Handbook of Applied Bayesian Analysis*, pages 3–24.

García-Gómeza, C., Pérez, A., and Prieto-Alaiz, M. (2019). Copula-based analysis of multivariate dependence patterns between dimensions of poverty in Europe. Technical report, Economics Working Paper.

Garthwaite, P. H., Fan, Y., and Sisson, S. A. (2016). Adaptive optimal scaling of Metropolis-Hastings algorithms using the Robbins-Monro process. *Communications in Statistics-Theory and Methods*, 45(17):5098–5111.

Genest, C. and Nešlehová, J. (2007). A primer on copulas for count data. *ASTIN Bulletin: The Journal of the IAA*, 37(2):475–515.

Genz, A. (1992). Numerical computation of multivariate Normal probabilities. *Journal of computational and graphical statistics*, 1(2):141–149.

Geweke, J. and Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *The review of financial studies*, 9(2):557–587.

Gill, J. and Casella, G. (2004). Dynamic tempered transitions for exploring multimodal posterior distributions. *Political Analysis*, 12(4):425–443.

Gunawan, D., Dang, K.-D., Quiroz, M., Kohn, R., and Tran, M.-N. (2018). Subsampling sequential Monte Carlo for static Bayesian models. *arXiv preprint arXiv:1805.03317*.

Gunawan, D., Tran, M.-N., Suzuki, K., Dick, J., and Kohn, R. (2019). Computationally efficient Bayesian estimation of high-dimensional Archimedean copulas with discrete and mixed margins. *Statistics and Computing*, 29(5):933–946.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications.

Holden, L. E., Dobson, A. J., Byles, J., Loxton, D., Dolja-Gore, X., Hockey, R., Lee, C. E., Chojenta, C., Reilly, N., Mishra, G. D., et al. (2013). Mental health: findings from the Australian longitudinal study on women's health.

Honkela, A., Raiko, T., Kuusela, M., Tornio, M., and Karhunen, J. (2010). Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes. *Journal of Machine Learning Research*, 11(Nov):3235–3268.

Kastner, G., Frühwirth-Schnatter, S., and Lopes, H. F. (2017). Efficient Bayesian inference for multivariate factor stochastic volatility models. *Journal of Computational and Graphical Statistics*, 26(4):905–917.

Loaiza-Maya, R. and Smith, M. S. (2019). Variational Bayes estimation of discrete-margined copula models with application to time series. *Journal of Computational and Graphical Statistics*, pages 1–17.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Murray, J. S., Dunson, D. B., Carin, L., and Lucas, J. E. (2013). Bayesian Gaussian copula factor models for mixed data. *Journal of the American Statistical Association*, 108(502):656–665.

Murteira, J. M. and Lourenço, Ó. D. (2011). Health care utilization and self-assessed health: specification of bivariate models using copulas. *Empirical Economics*, 41(2):447–472.

Nelsen, R. B. (2007). *An introduction to copulas.* Springer Science & Business Media.

NSW Ministry of Health (2013). NSW health framework for women's health 2013.

Ong, V. M., Nott, D. J., Tran, M.-N., Sisson, S. A., and Drovandi, C. C. (2018a). Variational Bayes with synthetic likelihood. *Statistics and Computing*, 28(4):971–988.

Ong, V. M.-H., Nott, D. J., and Smith, M. S. (2018b). Gaussian variational approximation with a factor covariance structure. *Journal of Computational and Graphical Statistics*, 27(3):465–478.

Panagiotou, D. and Stavrakoudis, A. (2015). Price asymmetry between different pork cuts in the usa: a copula approach. *Agricultural and Food Economics*, 3(1):6.

Patton, A. J. (2006). Modelling asymmetric exchange rate dependence. *International economic review*, 47(2):527–556.

Pham, H. (2006). *Springer handbook of engineering statistics.* Springer Science & Business Media.

Pitt, M., Chan, D., and Kohn, R. (2006). Efficient Bayesian inference for Gaussian copula regression models. *Biometrika*, 93(3):537–554.

Pitt, M. K., dos Santos Silva, R., Giordani, P., and Kohn, R. (2012). On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171(2):134–151.

Quiroz, M., Kohn, R., Villani, M., and Tran, M.-N. (2019). Speeding up MCMC by efficient data subsampling. *Journal of the American Statistical Association*, 114(526):831–843.

Ranganath, R., Wang, C., David, B., and Xing, E. (2013). An adaptive learning rate for stochastic variational inference. In *International Conference on Machine Learning*, pages 298–306.

Salimans, T., Knowles, D. A., et al. (2013). Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis*, 8(4):837–882.

Seeger, M. (2000). Bayesian model selection for support vector machines, Gaussian processes and other kernel classifiers. In *Advances in neural information processing systems*, pages 603–609.

Smith, M. S. (2011). Bayesian approaches to copula modelling. *arXiv preprint arXiv:1112.4204*.

Smith, M. S. and Khaled, M. A. (2012). Estimation of copula models with discrete margins via Bayesian data augmentation. *Journal of the American Statistical Association*, 107(497):290–303.

Smith, M. S., Loaiza-Maya, R., and Nott, D. J. (2019). High-dimensional copula variational approximation through transformation. *arXiv preprint arXiv:1904.07495*.

Stander, J., Dalla Valle, L., Taglioni, C., Liseo, B., Wade, A., and Cortina-Borja, M. (2019). Analysis of paediatric visual acuity using Bayesian copula models with sinh-arcsinh marginal densities. *Statistics in medicine*.

Tan, L. S., Bhaskaran, A., and Nott, D. J. (2019). Conditionally structured variational Gaussian approximation with importance weights. *arXiv preprint arXiv:1904.09591*.

Tan, L. S. and Nott, D. J. (2018). Gaussian variational approximation with sparse precision matrices. *Statistics and Computing*, 28(2):259–275.

Thacker, S. B., Stroup, D. F., Carande-Kulis, V., Marks, J. S., Roy, K., and Gerberding, J. L. (2006). Measuring the public's health. *Public health reports*, 121(1):14–22.

Titsias, M. and Lázaro-Gredilla, M. (2014). Doubly stochastic variational Bayes for non-conjugate inference. In *International conference on machine learning*, pages 1971–1979.

Tran, M.-N., Kohn, R., Quiroz, M., and Villani, M. (2016). The block pseudo-marginal sampler. *arXiv preprint arXiv:1603.02485*.

Tran, M.-N., Nguyen, N., Nott, D., and Kohn, R. (2019). Bayesian deep net GLM and GLMM. *Journal of Computational and Graphical Statistics*, pages 1–17.

Tran, M.-N., Nott, D. J., and Kohn, R. (2017). Variational Bayes with intractable likelihood. *Journal of Computational and Graphical Statistics*, 26(4):873–882.

Tukey, J. W. (1977). Modern techniques in data analysis. In *Proceedings of the NSF-Sponsored Regional Research Conference*, volume 7. Southern Massachusetts University.

Wilkins, R., Lab, I., Butterworth, P., and Vera-Toscano, E. (2019). *The household, income and labour dynamics in Australia survey: selected findings from waves 1 to 17*. Melbourne Institute of Applied Economic and Social Research, The University of Melbourne.

Yeo, I.-K. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959.

Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

```
%% Standard Psuedo Marginal method - Gaussian Copula with Factor↙
Structure -- proposal has an adaptive sigma & Alan Genz

clear; clc;
parpool(2)

%   Data
load('GaussianCopula_10dim_250m_2factors_ordinal.mat'); %columns↙
are dependent
data_col = size(y,2);
data_row = size(y,1);

%   Define my variables
iter = 220000;
burn = 20000;
accept = 1;
lratio_save = zeros(iter, 1);

%   parameters of prior distribution -- beta
Bv = 2;  % variance of beta prior -- normal dist
G = 25;  %number of groups -- if 1 then this is the standard PM↙
method

%   parameters for copula parameter -- beta (has factor structure)
num_factors = 2; %number of factors to estimate the covariance↙
matrix of the Gaussian copula
beta = 0*vech(ones(data_col, num_factors));  %inital parameters --↙
vech converts into lower triangular
beta_store = zeros(length(beta), iter);

%   Variational parameters -- nita & gamma (has factor structure)
num_free_beta_parm = length(beta);
omega = eye(num_free_beta_parm); %inital parameters

%   parameters for unbiased likelihood & factor model
num_samples = 20; %can be super small because of low variance in↙
llh given by Alan Genz
[lower_limit, upper_limit] = VBIL_own_limit_ordinal(y, data_row,↙
data_col);
ns = 1;
u = rand(data_col, num_samples, data_row); %inital value --↙
changed data structure!!
```

```
%Adaptive update of omega
t_beta = 1;
m = num_free_beta_parm;
p = 0.234;
alpha = - norminv(p/2);
cov = 1;
c = cov*( (1-1/m)*(2*pi)^0.5*exp(alpha^2/2)/(2*alpha) + 1/(m*p*(1-↙
p))) ;

%Defining this simplies the MH -- inital values
rest = [1,data_col + 1 ];
beta_exp = beta ;
beta_exp(rest) = exp(beta(rest)) ;
llh = Block_PM_GaussianCopula_llh(beta_exp, data_row, data_col,↙
num_factors, u, lower_limit, upper_limit, ns, num_samples) ;
lprior = -0.5*num_free_beta_parm*(log(2*pi*Bv)) - 0.5*(1/Bv*↙
(beta_exp')*beta_exp) + log( prod( beta_exp(rest) ) ) ;  %assumes↙
each element in beta is indep and distribution norm(0,Bv)
lposterior_est = llh + lprior;  %the p(u) & beta proposal cancels↙
with the denominator


%   Independent Metropolis Hastings
tic
for i=1:iter

    i
%Current Values
%current beta & u
c_beta = beta;
c_omega = omega;
c_u = u;
c_llh = llh;
c_lposterior_est = lposterior_est ;


%Proposed Values
%proposed u
beta = c_beta + chol(omega, 'lower')*randn(num_free_beta_parm, 1)↙
;
beta_exp = beta ;
```

```
beta_exp(rest) = exp(beta(rest)) ;

num_elts_in_group = floor(data_row/G);
group_to_update = randperm(G,1);  %selects a number from 1:G
first_person_to_update = (group_to_update - 1)*num_elts_in_group +↙
1;
last_person_to_update = (group_to_update*num_elts_in_group);
persons_to_update = first_person_to_update:last_person_to_update;↙
%selects the persons to update
u(:, :, persons_to_update) = rand(data_col, num_samples,↙
num_elts_in_group);

llh = Block_PM_GaussianCopula_llh(beta_exp, data_row, data_col,↙
num_factors, u, lower_limit, upper_limit, ns, num_samples) ;
lprior = -0.5*num_free_beta_parm*(log(2*pi*Bv)) - 0.5*(1/Bv*↙
(beta_exp')*beta_exp) + log( prod( beta_exp(rest) ) ) ;
lposterior_est = llh + lprior;


% Metropolis-Hasting acceptance
lratio = lposterior_est - c_lposterior_est  ;

laccept = min(0, lratio);
lratio_save(i) = lratio;

if (laccept < log(rand(1)) )
   beta = c_beta ;
   u = c_u;
   lposterior_est = c_lposterior_est;
   llh = c_llh;
   accept = accept + 1

   if(i <= 200)
      cov = cov - c*p/i ;    %calc sigma -- tuning parameter of↙
the random walk
   else
      cov = cov - c*p/max(200,i/m) ;
   end

else
    if(i <= 200)
        cov = cov + c*(1-p)/i;  %robbin's Monro Process
    else
        cov = cov + c*(1-p)/max(200,i/m);  %robbin's Monro Process
    end

end

beta_store(:, i) = beta;


%Update omega
c = cov*( (1-1/m)*(2*pi)^0.5*exp(alpha^2/2)/(2*alpha) + 1/(m*p*(1-↙
p))) ;

 %calc recursion beta
      t_beta_m1 = t_beta; %previous value of t_beta
      t_beta = 1/i*((i-1)*t_beta_m1 + beta);

  %calc cov recursion
  if (i <= 100)
      est_omega = eye(m);
  else
      est_omega = (i-2)/(i-1)*est_omega + t_beta_m1*t_beta_m1' -↙
i/(i-1)*(t_beta*t_beta') + 1/(i-1)*(beta*beta') ;
  end

    omega = cov^2.*(est_omega + cov^2*eye(m)./i) ;


end

CPU_time = toc
beta_store_final = beta_store(:, burn+1:end); %without thinning
beta_store1 = beta_store_final ;
beta_store1(rest,:) = exp(beta_store_final(rest,:)) ;


num_beta_elts = num_free_beta_parm;
IACT_store = zeros(num_beta_elts, 1);
for i=1:num_beta_elts

        IACT_store(i) = IACT(beta_store1(i,:));
```

```
end
IACT_store
```

```
function [lower_limit, upper_limit] = VBIL_own_limit_ordinal(y,↙
data_row, data_col)

lower_limit = zeros(data_row, data_col);
upper_limit = zeros(data_row, data_col);

%empirical prob
for i=1:data_col

    [C,ia,ic] = unique(y(:,i)); % C is the unique values
    a_counts = accumarray(ic,1); %counts of each integer -- this↙
is ordered
    a_cumsum = cumsum(a_counts); %cumulative sum
    prob_int = a_cumsum./data_row;

    %construction of upper and lower limits for the data
    num_unique_values = length(C);
    lower_limit(:,i) = [y(:,i) == C(1)].* 0.0001;
    upper_limit(:,i) = [y(:,i) == C(end)].* 0.9999;

    for unique_value = 1:num_unique_values-1
        lower_limit(:, i) = lower_limit(:, i) + [y(:,i) == C↙
(unique_value + 1)].* prob_int(unique_value);
        upper_limit(:, i) = upper_limit(:, i) + [y(:,i) == C↙
(unique_value)].* prob_int(unique_value);
    end


end

end
```

```
%% Gaussian Copula with factor structure on the Covariance matrix

clear; clc;
parpool(20)

%   Data
load('GaussianCopula_10dim_250m_2factors_ordinal'); %dep in the↙
cols

data_col = size(y,2);
data_row = size(y,1);

%   Define my variables
t = 1;
iter = 4000;
S = 200;
stop = false;
lower_bound = zeros(iter, 1);
lower_bound_scale = data_row*data_col;

%parameters of prior distribution -- beta
Bv = 2;  % variance of beta prior -- indep. normal dist
lprior = zeros(S,1);

%parameters for copula parameter -- beta (has factor structure)
num_factors = 2; %number of factors to estimate the covariance↙
matrix of the Gaussian copula
beta = vech(ones(data_col, num_factors));  %inital parameters --↙
vech converts into lower triangular!
beta_store = zeros(iter, length(beta));
tuneparm = 1;

%Variational parameters -- nita & gamma (has factor structure)
num_free_beta_parm = length(beta);
q_nita = 0.5*ones(num_free_beta_parm, 1);  %inital parameters --↙
mean of beta
qnum_factors = 1;
q_gamma = 0.1*vech(ones(num_free_beta_parm, qnum_factors)); %↙
inital parameters
q_D = 0.2*ones(num_free_beta_parm, 1); %diag matrix with squared↙
terms
lambda = [q_nita', q_gamma', q_D'];  %--row vector
```

```
dim_lambda = length(lambda);
lambda_store = zeros(iter, dim_lambda);


%parameters for unbiased likelihood
grad_lq = zeros(S, dim_lambda);
scalar = zeros(S,1);
g_lambda = zeros(S, dim_lambda);
[lower_limit, upper_limit] = VBIL_own_limit_ordinal(y, data_row,↙
data_col);
num_samples = 30; %parameters for numerically calc the gaussian↙
likelihood
ns = 1; %parameters for numerically calc the gaussian likelihood

%parameters for control variate
c = zeros(iter+1, dim_lambda);

% %ADADELTA parameters
E_g2_param=0;
E_delta2_param=0;
adapt_rho=0.95;
adapt_eps=10^-6;

%%
%Step 1: Control variate
tic

        q_D = sparse(1:num_free_beta_parm, 1:num_free_beta_parm,↙
q_D);
        q_omega = q_gamma*q_gamma' + q_D.^2;


parfor s = 1:S

        chol_beta = chol(tuneparm*q_omega, 'lower');
        beta = q_nita + chol_beta*randn(num_free_beta_parm, 1) ; %↙
generate inital sample of size s -

        lprior = -0.5*num_free_beta_parm*(log(2*pi*Bv)) - 0.5*↙
(1/Bv*(beta')*beta);  %assumes each element in beta is indep and↙
distribution norm(0,Bv)
        llh = VBIL_GaussianCopula_llh(beta, data_row, data_col,↙
num_factors, lower_limit, upper_limit, ns, num_samples); %just↙
using Alan Genz

        invomegagamma = (q_omega)\q_gamma;
        invomegaD = (q_omega)\q_D;
        grad_qD2 = diag( -invomegaD + q_omega\(beta - q_nita)*↙
(beta - q_nita)'*invomegaD) ;
        grad_lq(s,:)= [ (q_omega\(beta - q_nita))', ...  %grad of↙
q_nita
                        ( -invomegagamma + (q_omega)\(beta -↙
q_nita)*(beta - q_nita)'*invomegagamma )', ... %grad of q_gamma
                        grad_qD2' ]; %grad of q_D2


        scalar(s) =  lprior + llh + 0.5*( num_free_beta_parm*log↙
(2*pi) + log(det(q_omega)) + ...
                                            (beta - q_nita)'/↙
(q_omega)*(beta - q_nita) ) ;
end


for k=1:dim_lambda
        temp = cov(scalar.*grad_lq(:,k), grad_lq(:,k));
        c(1,k) = temp(1,2)/temp(2,2);
end


%%
%Step 2: KL

while ~stop

           q_omega = q_gamma*q_gamma' + q_D.^2;

    parfor s=1:S

        chol_beta = chol(tuneparm*q_omega, 'lower');
        beta = q_nita + chol_beta*randn(num_free_beta_parm, 1) ; %↙
generate inital sample of size s -

        lprior = -0.5*num_free_beta_parm*(log(2*pi*Bv)) - 0.5*↙
(1/Bv*(beta')*beta);   %assumes each element in beta is indep and↙
```

```
distribution norm(0,Bv)
        llh = VBIL_GaussianCopula_llh(beta, data_row, data_col,↙
num_factors, lower_limit, upper_limit, ns, num_samples); %using↙
Alan Genz

        invomegagamma = (q_omega)\q_gamma;
        invomegaD = (q_omega)\q_D;
        grad_qD2 = diag( -invomegaD + q_omega\(beta - q_nita)*↙
(beta - q_nita)'*invomegaD) ;
        grad_lq(s,:)= [ (q_omega\(beta - q_nita))', ...  %grad of↙
q_nita
                        ( -invomegagamma + (q_omega)\(beta -↙
q_nita)*(beta - q_nita)'*invomegagamma )', ... %grad of q_gamma
                        grad_qD2' ]; %grad of q_D2


        scalar(s) =  lprior + llh + 0.5*( num_free_beta_parm*log↙
(2*pi) + log(det(q_omega)) + ...
                                            (beta - q_nita)'/↙
(q_omega)*(beta - q_nita) ) ;

        g_lambda(s,:) = grad_lq(s,:) .* ( scalar(s) - c(t,:) );

    end

    %compute the grad_KL
        grad_KL = mean(g_lambda) %--row vector with lambda↙
dimension columns

    %compute control variate t continued to be used in next↙
iter...
    for k=1:dim_lambda
        temp = cov(scalar.*grad_lq(:,k), grad_lq(:,k));
        c(t+1,k) = temp(1,2)/temp(2,2);
    end

    lower_bound(t) = mean(scalar)/lower_bound_scale;

    %update lambda using ADADELTA
    E_g2_param = adapt_rho*E_g2_param + (1-adapt_rho).*(grad_KL.↙
^2);
    delta = sqrt( E_delta2_param + adapt_eps) ./ sqrt( E_g2_param↙
+ adapt_eps) .* grad_KL;
    E_delta2_param = adapt_rho*E_delta2_param + (1-adapt_rho)*↙
(delta.^2);

    lambda = lambda + delta;
    lambda_store(t+1,:) = lambda;
    q_nita = lambda(1:num_free_beta_parm)' ;
    q_gamma = lambda(num_free_beta_parm + 1 :(num_free_beta_parm +↙
num_free_beta_parm*qnum_factors))';
    q_D = lambda((num_free_beta_parm +↙
num_free_beta_parm*qnum_factors + 1) : end)';
    q_D = sparse(1:num_free_beta_parm, 1:num_free_beta_parm, q_D);


    if t > iter
        stop = true;
    end

    t = t + 1

end

CPU_time = toc
```

```
function [lower_limit, upper_limit] = VBIL_own_limit_ordinal(y,↙
data_row, data_col)

lower_limit = zeros(data_row, data_col);
upper_limit = zeros(data_row, data_col);

%empirical prob
for i=1:data_col

    [C,ia,ic] = unique(y(:,i)); % C is the unique values
    a_counts = accumarray(ic,1); %counts of each integer -- this↙
is ordered
    a_cumsum = cumsum(a_counts); %cumulative sum
    prob_int = a_cumsum./data_row;

    %construction of upper and lower limits for the data
    num_unique_values = length(C);
    lower_limit(:,i) = [y(:,i) == C(1)].* 0.0001;
    upper_limit(:,i) = [y(:,i) == C(end)].* 0.9999;

    for unique_value = 1:num_unique_values-1
        lower_limit(:, i) = lower_limit(:, i) + [y(:,i) == C↙
(unique_value + 1)].* prob_int(unique_value);
        upper_limit(:, i) = upper_limit(:, i) + [y(:,i) == C↙
(unique_value)].* prob_int(unique_value);
    end


end

end
```

```
function llh = VBIL_GaussianCopula_llh(beta, data_row, data_col,↙
num_factors, lower_limit, upper_limit, ns, num_samples)


% Calculate log likelihood - proposed beta
beta = vec2mat(beta, data_col, num_factors);    %convert beta back↙
into a matrix
qsilatmvnv_cov_mat = beta*beta' + eye(data_col) ;  %unrestricted↙
covariance matrix
qsilatmvnv_diag = diag( qsilatmvnv_cov_mat ).^(0.5);
qsilatmvnv_lower_limit = qsilatmvnv_diag' .* norminv(lower_limit);↙
%dim= num_row*num_col
qsilatmvnv_upper_limit = qsilatmvnv_diag' .* norminv(upper_limit);

lprob_person = zeros(data_row,1);

u = rand(data_col, num_samples, data_row);

parfor person = 1:data_row     %first_person_to_update:↙
last_person_to_update %number of independent people in group to↙
update
    u_person = u(:, :, person);

    lprob_person(person) = log(qsilatmvnv_Anny_Block_PM(↙
qsilatmvnv_cov_mat,  qsilatmvnv_lower_limit(person, :)', ...
                                    qsilatmvnv_upper_limit(person,↙
:)', u_person, ns, num_samples));  %Alan Genz
%
%
%     prob_person  = mvncdf(qsilatmvnv_lower_limit(person, :)',↙
...   %Botev
%                                     qsilatmvnv_upper_limit↙
(person, :)',  qsilatmvnv_cov_mat,  1000);
%     lprob_person(person) = log( [prob_person.prob] );


end
llh = sum(lprob_person);

end
```