BAYESIAN ESTIMATION OF MIXED LOGIT MODELS: SELECTING AN APPROPRIATE PRIOR FOR THE COVARIANCE MATRIX

Deniz Akinc [a]

Email: deniz.akinc@kuleuven.be

Martina Vandebroek [a,b]

Email: martina.vandebroek@kuleuven.be

[a] Faculty of Business and Economics, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium

[b] Leuven Statistics Research Centre, KU Leuven, Celestijnenlaan 200B, 3001 Leuven, Belgium

BAYESIAN ESTIMATION OF MIXED LOGIT MODELS: SELECTING AN APPROPRIATE PRIOR FOR THE COVARIANCE MATRIX

Abstract

Maximum likelihood and Bayesian estimation are both frequently used to fit mixed logit models to choice data. The type and the number of quasi-random draws used for simulating the likelihood and the choice of the priors in Bayesian estimation have a big impact on the estimates. We compare the different approaches and compute the relative root mean square errors of the resulting estimates for the mean, covariance matrix and individual parameters in a large simulation study. We focus on the prior for the covariance matrix in Bayesian estimation and investigate the effect of Inverse Wishart priors, the Separation Strategy, Scaled Inverse Wishart and Huang Half-t priors. We show that the default settings in many software packages can lead to very unreliable results and that it is important to check the robustness of the results.

Keywords: Mixed Logit Model, Hierarchical Bayesian Estimation, Separation Strategy, Inverse Wishart Distribution, Scaled Inverse Wishart Distribution, Huang Half-t Distribution

1 Introduction

The mixed logit model (also called random parameter logit model) has rapidly become the standard model to analyze choice behavior within health economics, marketing and the trans-

portation research literature. The model extends and improves the standard multinomial logit model by focusing on the distribution of individual-level preferences rather than on average preferences (Revelt and Train, 1998). To fit the model, one often uses maximum (simulated) likelihood estimation or Bayesian estimation, where the latter is using Monte Carlo Markov Chain (MCMC) methods to compute the joint posterior distribution of the parameters.

In theory, both methods should converge to the true parameters if the sample size increases and the model is correctly specified. With smaller samples, one hopes for similar estimates from both approaches which was supported by the results in Huber and Train (2001). However, the evidence of this similarity is not very large as their findings are based on a single dataset. Besides Huber and Train (2001), there are other studies that have investigated the similarity based on only one dataset (Regier *et al.*, 2009; Haan et al., 2015). More recently, Elshiewy *et al.* (2017) reported on a somewhat larger study but their conclusion is based on a model with only a few random parameters and a dataset with a huge number of choice sets.

In this paper we investigate how close the estimates from both approaches are based on a large simulation study. We examine the effect of the priors in Bayesian estimation and of the number of Halton draws for maximum simulated likelihood. As it turns out that the prior on the covariance matrix has more impact than expected, we study the effect of some frequently used priors as well as some recently suggested priors.

The most commonly used prior for the covariance matrix in the Bayesian approach is the Inverse Wishart distribution. Though this prior is easy to implement because of it conjugacy property for the multivariate normal distribution, it has some objectionable issues which will be discussed later. Some alternative priors have been suggested: the Separation Strategy (Barnard,*et al.*, 2000), the Scaled Inverse Wishart distribution (O'Malley and Zaslavsky, 2008) and the Huang Half-t distribution (Huang, *et al.*, 2013). We will investigate how and to what extent the results depend on these priors.

This paper is organized as follows. In section 2, the mixed logit model and both the maximum simulated likelihood and the Bayesian approaches are described. We explain the undesirable issues related to the Inverse Wishart distribution as a prior for the covariance matrix and

describe some other prior distributions that have been proposed in the literature in section 3. Section 4 contains a simulation study with different scenarios and an overview of the relative root mean square errors (RRMSE) of the parameters. Finally, the discussion of the results and the main conclusions are given in section 5.

2 The mixed logit (MXL) model

As stated before, the mixed logit model describes the heterogeneity in the population by the distribution of the individual-level preferences rather than relying on average preferences. So the individual-level parameters, $\boldsymbol{\beta}_n$, associated with the attributes are assumed to vary according to a probability distribution $\boldsymbol{\beta}_n \sim f(\boldsymbol{\beta}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In the sequel, we will assume that the heterogeneity distribution $f(\boldsymbol{\beta}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a multivariate normal distribution.

Conditional on $\boldsymbol{\beta}_n$, the probability that person n chooses alternative k in choice set s is

$$p_{ksn}(\boldsymbol{\beta}_n) = \frac{\exp(\mathbf{x}'_{ksn}\boldsymbol{\beta}_n)}{\sum_{i=1}^{K} \exp(\mathbf{x}'_{isn}\boldsymbol{\beta}_n)}, \tag{1}$$

where $\mathbf{x}_{ksn}$ is a p-dimensional vector characterizing the attribute levels of alternative k in choice set s for respondent n with p number of coefficients in the model. The choice is stored in the variable, y_{ksn}, a binary variable that equals one if respondent n chooses alternative k in choice set s and zero otherwise. Let $\mathbf{y}_n$ contain all the choices from respondent n corresponding to all S choice sets. The probability, unconditional on $\boldsymbol{\beta}_n$, of a respondent n's choices $\mathbf{y}_n$ is

$$\pi_n(\mathbf{y}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int_{\boldsymbol{\beta}_n} \left(\prod_{s=1}^{S} \prod_{k=1}^{K} \left(p_{ksn}(\boldsymbol{\beta}_n)\right)^{y_{ksn}} \right) f(\boldsymbol{\beta}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{\beta}_n. \tag{2}$$

The mixed logit (MXL) model takes into account the correlation of the probabilities for a single respondent in multiple choices and is therefore also called the panel mixed logit model.

The log-likelihood of the MXL model is

$$\begin{aligned} LL(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{y}) &= \sum_{n=1}^{N} \ln\left(\pi_n(\mathbf{y}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})\right), \\ &= \sum_{n=1}^{N} \ln\left[\int_{\boldsymbol{\beta}_n} \left(\prod_{s=1}^{S}\prod_{k=1}^{K} [p_{ksn}(\boldsymbol{\beta}_n)]^{y_{ksn}}\right) f(\boldsymbol{\beta}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{\beta}_n\right], \end{aligned} \tag{3}$$

where $\mathbf{y} = [\mathbf{y}_1, ..., \mathbf{y}_N]$ denotes the matrix of choices from all N respondents.

The following sections review the two main procedures for estimating the hyperparameters, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, of the heterogeneity distribution and the related individual-level parameters $\boldsymbol{\beta}_n$.

2.1 Maximum simulated likelihood estimation

In this approach, the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ are considered fixed but unknown. The estimators of these parameters are obtained by maximizing the (log-) likelihood. As this maximum depends on the sample at hand the estimators are stochastic. As the log-likelihood denoted in equation (3) contains a multivariate integral over the individual-level parameters which cannot be computed analytically, random draws of individual-level parameters are required to approximate this multivariate integral. The corresponding simulated probability is defined as $\hat{\pi}_n(\mathbf{y}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\hat{\pi}_n(\mathbf{y}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{R}\sum_{r=1}^{R}\left(\prod_{s=1}^{S}\prod_{k=1}^{K} (p_{ksn}(\boldsymbol{\beta}_n^r))^{y_{ksn}}\right), \tag{4}$$

with $\boldsymbol{\beta}_n^r$ random draws from $f(\boldsymbol{\beta}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The simulated log-likelihood is then defined as

$$SLL(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{y}) = \sum_{n=1}^{N} \ln\left(\hat{\pi}_n(\mathbf{y}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})\right). \tag{5}$$

Either Pseudo-Monte Carlo (PMC) or Quasi-Monte Carlo (QMC) random draws are used to evaluate these integrals. The quasi-random draws decrease the computation time noticeably by providing better estimates with a smaller number of draws and there exists an abundant literature on the performance of different types of QMC draws (for a comparison in a choice

modeling context, see for instance Bhat, 2001; Sándor & Train, 2004; Bliemer *et al.*, 2008; Yu *et al.*, 2010).

Table 1: Default settings for some software packages.

Software Package	Version	Type of Random Draws	Number of draws
NLOGIT	NLOGIT 6	Pseudo MC	100
R-**mlogit**	0.2-4	Pseudo MC	40
R-**gmnl**	1.1-3	Halton	40
STATA-**mixlogit**	STATA 14.2	Halton	50

Table 1 has the default settings for some frequently used software packages. In this paper, we will use the R-package **gmnl** and only use Halton draws with varying number of draws. We will only use the results obtained by the MSL estimation as a benchmark for the estimates obtained by Bayesian estimation in Section 4.

Maximizing the likelihood function only yields estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Estimates for the individual parameters $\boldsymbol{\beta}_n$ can then be obtained using Bayes' theorem (Train, 2003):

$$g(\boldsymbol{\beta}_n|\mathbf{y}_n, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\pi(\mathbf{y}_n|\boldsymbol{\beta}_n) f(\boldsymbol{\beta}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\pi(\mathbf{y}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})}, \tag{6}$$

where $g(\boldsymbol{\beta}_n|\mathbf{y}_n, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the posterior distribution of the individual parameters $\boldsymbol{\beta}_n$ conditional on the observed sequence of choices and on the unconditional distribution $f(\boldsymbol{\beta}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ which is approximated by using the estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ from maximizing the likelihood function.

2.2 Bayesian estimation

Unlike in the previous approach, the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are now considered stochastic. This Bayesian technique was developed by Allenby (1997) and generalized by Train (2003). To obtain the posterior distribution of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, one needs to specify prior distributions $m_0(\boldsymbol{\mu})$ and $q_0(\boldsymbol{\Sigma})$ and combine these with the likelihood function of the data. Gibbs sampling, in combination with the Metropolis-Hasting algorithm, is then used to obtain draws from the joint posterior distribution. More precisely, the Gibbs sampling steps are used to update the parameters when their conditional posterior distributions are available and Metropolis-Hastings

steps are considered otherwise.

In most applications, a non-informative prior distribution is used for the mean $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. The joint posterior distribution of $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta}_n$, for all N respondents can be written as

$$f(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_N | \mathbf{y}) \propto \prod_{n=1}^{N} L(\mathbf{y}_n | \mathbf{X}_n, \boldsymbol{\beta}_n)\ f(\boldsymbol{\beta}_n | \boldsymbol{\mu}, \boldsymbol{\Sigma})\ m_0(\boldsymbol{\mu})\ q_0(\boldsymbol{\Sigma}). \tag{7}$$

In commonly used software packages as the R-package *bayesm*, SAWTOOTH and STATA, $m_0(\boldsymbol{\mu})$ is a multivariate normal distribution with zero mean and a diagonal covariance matrix with large variances and $q_0(\boldsymbol{\Sigma})$ is an Inverse Wishart distribution with ν degrees of freedom and a p-dimensional scale matrix $\mathbf{T}$, IW(ν,$\mathbf{T}$).

Draws from this posterior can be obtained by iteratively taking draws from the conditional posterior distributions of some of the parameters, given all other parameters:

- The conditional posterior distribution of $\boldsymbol{\mu}$, given $\boldsymbol{\beta}_n$ and $\boldsymbol{\Sigma}$ is $\mathcal{N}(\bar{\boldsymbol{\beta}}, \boldsymbol{\Sigma}/N)$, with $\bar{\boldsymbol{\beta}}$ representing the sample mean of the current $\boldsymbol{\beta}_n$ values, and $\mathcal{N}(.)$ denoting the multivariate normal distribution.

- The conditional posterior distribution of $\boldsymbol{\Sigma}$, given $\boldsymbol{\beta}_n$ and $\boldsymbol{\mu}$ is $IW(\nu+N,\ \mathbf{T}+\sum_{n=1}^{N}(\boldsymbol{\beta}_n - \boldsymbol{\mu})(\boldsymbol{\beta}_n - \boldsymbol{\mu})')$.

- The posterior distribution of $\boldsymbol{\beta}_n$ conditional on $\mathbf{y}_n$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ has no closed form, and we use a Metropolis-Hasting random walk procedure to take draws from a normal proposed distribution. By means of the Metropolis-Hastings algorithm, a draw $\boldsymbol{\beta}_n^t$ is obtained for each individual n separately in the t^{th} iteration:

  - Stack p independent standard normal values in a vector η and compute a trial draw $\tilde{\boldsymbol{\beta}}_n^t = \boldsymbol{\beta}_n^{t-1} + \rho L \eta$, with L the Choleski factor of $\boldsymbol{\Sigma}^t$ and ρ a scalar fixed by the researcher.

  - Compute the ratio

$$F = \frac{L(\tilde{\boldsymbol{\beta}}_n^t)\ \phi(\tilde{\boldsymbol{\beta}}_n^t | \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)}{L(\boldsymbol{\beta}_n^{t-1})\ \phi(\boldsymbol{\beta}_n^{t-1} | \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)},$$

with $L(\boldsymbol{\beta}_n)$ the likelihood of the conditional logit model and ϕ the standard normal density.

– Draw a value r from the standard uniform distribution. If $r \leq F$ then accept the trial $\boldsymbol{\beta}_n^t = \tilde{\boldsymbol{\beta}}_n^t$. In case $r > F$, $\boldsymbol{\beta}_n^t = \boldsymbol{\beta}_n^{t-1}$.

Starting from initial values $\boldsymbol{\mu}^0$, $\boldsymbol{\Sigma}^0$ and $\boldsymbol{\beta}_n^0$ $\forall n$, one iteratively obtains the draws $\boldsymbol{\mu}^t$, $\boldsymbol{\Sigma}^t$ and $\boldsymbol{\beta}_n^t$ $(\forall n)$ in the t^{th} iteration according to the given Gibbs and Metropolis-Hastings steps. Executing these steps many times, the draws converge to draws from the joint posterior distribution of $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta}_n$ $(\forall n)$. Means of the draws after the burn-in period act as estimates for the model parameters.

As the simulation study to be described later revealed a large impact of the prior $q_0(\boldsymbol{\Sigma})$ on the results, we will have a closer look at the Inverse Wishart and some other priors for $\boldsymbol{\Sigma}$.

3 Prior for covariance matrices

3.1 Inverse Wishart distribution

Although the Inverse Wishart distribution is often used in Bayesian estimation because of its natural conjugacy property with the multivariate normal distribution, its properties are not commonly known. The density function of an IW(ν,$\mathbf{T}$) is defined as

$$q(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu+p+1)/2} \exp\left(-\frac{1}{2} tr(\mathbf{T}\boldsymbol{\Sigma}^{-1})\right). \tag{8}$$

This distribution is defined for ν larger than or equal to p but the mean $E(\boldsymbol{\Sigma}) = \mathbf{T}/(\nu-p-1)$ only exists if ν is larger than $p+1$. To get more insight, it is useful to know that (denote the variances by σ_i^2 and the correlations by ρ_{ij}):

- σ_i^2 has a scaled inverse $\chi^2(\nu - p + 1, \frac{T_{ii}}{\nu-p+1})$ distribution, with T_{ii} the i^{th} diagonal entry of $\mathbf{T}$, which has a very low density near zero.

- ρ_{ij} has a complicated density but if $\mathbf{T}$ is a diagonal matrix, the density is proportional

to $(1-\rho_{ij}^2)^{(\nu-p+1)/2}$. This implies that if $\mathbf{T}$ is a diagonal matrix and $\nu = p+1$, the correlations are uniformly distributed on $[-1,1]$ (Barthelmé, 2012) .

- When $\nu > p+1$ the marginal distribution of each correlation is unimodal around zero (O'Malley and Zaslavsky, 2008).
- As long as $\nu \in [p, p+1)$, the correlations have a bimodal distribution with high density on -1 and 1 (O'Malley and Zaslavsky, 2008).
- The correlations, ρ_{ij}, and the variances, σ_i^2, are correlated.

Remark that there are some issues with Inverse Wishart priors. A first issue is that a single hyperparameter ν controls several distributional properties of the variances as well as of the correlations. It influences for instance the location and the variance of the scaled inverse χ^2 distributions of all the variances simultaneously. It controls at the same time the prior densities of all the correlation coefficients. Furthermore, IW priors induce correlation between the variances and the correlations. For instance, if $\mathbf{T} = \mathbf{I}_p$, small variances are associated with correlations near zero and larger variances with correlations near -1 or 1. So, choosing the hyperparameter to control the prior density of the variances often has unexpected consequences on the prior for the correlations.

Because of the low density near zero of a scaled inverse χ^2 distribution if $\mathbf{T} = \mathbf{I}_p$, it can be expected that when true heterogeneity is small, an Inverse Wishart prior will cause bias toward larger variances.

Although, there is some discussion in the literature about the best specification of an uninformative prior for correlations, a uniform or unimodal distribution is clearly to be preferred over the bimodal which means that ν should at least be equal to $p+1$.

We will illustrate some often used Inverse Wishart priors by showing the implied priors on the variances and on the correlations as well as the correlation between these parameters. Train (2003), for instance, used $\nu = p$, $\mathbf{T} = \nu\mathbf{I}_p$ and $\mathbf{T} = \mathbf{I}_p$. As these are improper priors, other researchers, such as Rossi, et al. (1996 & 2005), have used $\nu = p+3$ or $\nu = p+4$. Balcombe, et al. (2009) tried to avoid the resulting inflated estimates by using $\nu = p(p+1)/2$

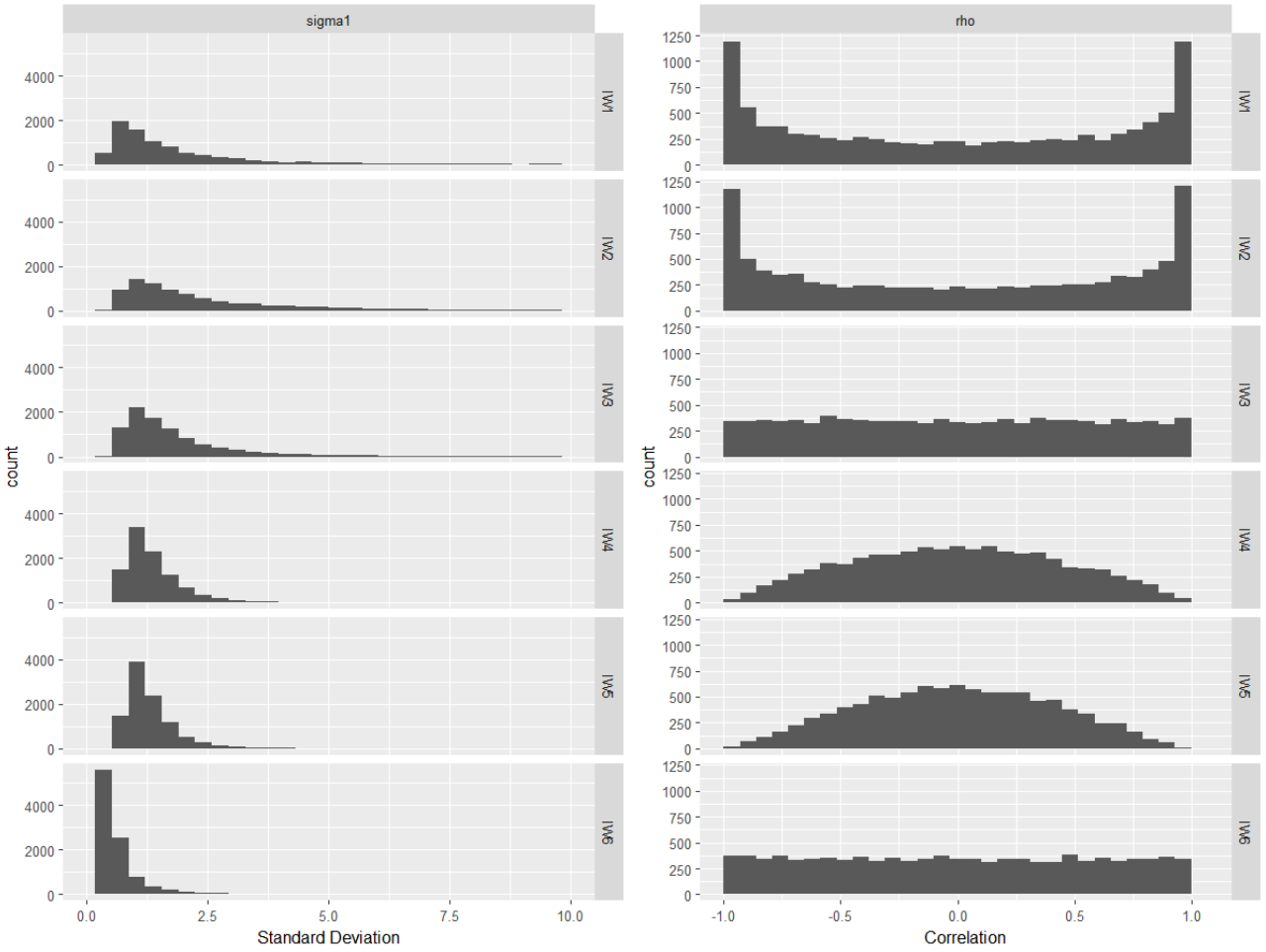

Figure 1: Prior distributions of the standard deviation (σ_1) and the correlation (ρ_{12}) implied by different Inverse Wishart distributions on the covariance matrix in case $p = 2$ (based on 10000 draws); (i) $IW1 = IW(\nu = p, \mathbf{I}_p)$; (ii) $IW2 = IW(\nu = p, \nu\mathbf{I}_p)$; (iii) $IW3 = IW(\nu = p + 1, \nu\mathbf{I}_p)$; (iv) $IW4 = IW(\nu = p + 3, \nu\mathbf{I}_p)$; (v) $IW5 = IW(\nu = p + 4, \nu\mathbf{I}_p)$; (vi) $IW6 = IW(\nu = 0.5p(p + 1), 0.1\nu\mathbf{I}_p)$.

and $\mathbf{T} = 0.1\nu\mathbf{I}_p$. The plots in the left panel of Figure 1 illustrate the corresponding prior densities for the standard deviations in case $p = 2$ and the plot on the right panel shows that the related priors on the correlation can have rather unexpected patterns. It is clear that IW1 and IW2 have long tailed distributions for the standard deviations combined with bimodal distributions for the correlation coefficients. IW3 provides a flat distribution for the correlations since $\nu = p + 1$ with a similar prior for the standard deviations. IW4 and IW5 combine relatively small values for the standard deviations with unimodal distribution for the correlations. Finally, IW6 represents a belief in small values for the standard deviations with a uniform prior for the correlations.

Figure 2 shows the correlation between the standard deviations and the correlations sho-

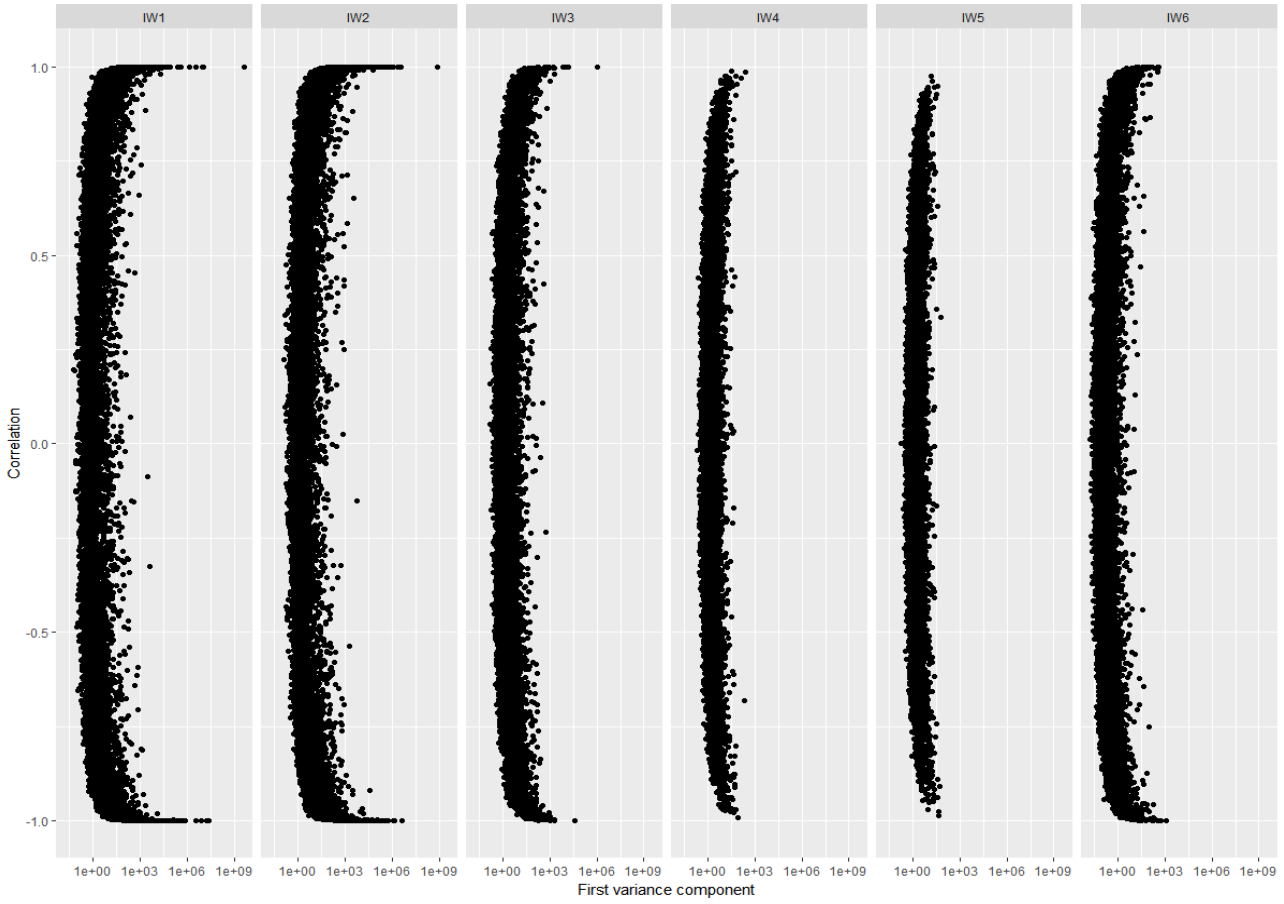

Figure 2: Scatterplot of the correlation coefficient vs. the first variance component implied by different Inverse Wishart distributions on the covariance matrix in case $p = 2$ (based on 10000 draws); Prior Distributions: (i) $IW1 = IW(\nu = p, \mathbf{I}_p)$; (ii) $IW2 = IW(\nu = p, \nu\mathbf{I}_p)$; (iii) $IW3 = IW(\nu = p + 1, \nu\mathbf{I}_p)$; (iv) $IW4 = IW(\nu = p + 3, \nu\mathbf{I}_p)$; (v) $IW5 = IW(\nu = p + 4, \nu\mathbf{I}_p)$; (vi) $IW6 = IW(\nu = 0.5p(p + 1), 0.1\nu\mathbf{I}_p)$.

wing that large standard deviations are assumed to go together with large absolute values for the correlation as was pointed out by Barthelmé (2012). This association seems to be less strong for the IW3 and IW4. Remark that the plots in Figure 1 and 2 can easily be obtained by the R-code included on the website (https://github.com/dakinc/estimation-comparison/) which relies heavily on the R-code given by Matt Simpson (2012) on his website.

In order to solve some undesirable issues related to IW priors, some alternative prior distributions have recently been proposed in the literature.

3.2 Separation Strategy

To get rid off the correlation between the standard deviations and the correlations, Barnard *et al.* (2000) developed a Separation Strategy (SS) which models these components of the

covariance matrix independently. The covariance matrix is decomposed as $\mathbf{\Sigma} = \mathbf{\Delta R \Delta}$ where $\mathbf{\Delta}$ is a diagonal matrix with elements σ_i and $\mathbf{R}$ is a correlation matrix. Separate priors are used for both components as follows:

(i) $log(\sigma_i) \overset{iid}{\sim} \mathcal{N}(b_i, \xi_i^2)$

(ii) $\mathbf{R}$ starts from an $IW(\nu, \mathbf{T})$ distribution which is transformed into a correlation matrix.

As such, the priors on the standard deviations can be chosen without influencing the prior on the correlations. We denote this prior as $SS(\nu, \mathbf{T}, \boldsymbol{b}, \boldsymbol{\xi})$ where $\boldsymbol{\xi}$ is a vector of ξ_i and $\boldsymbol{b}$ is a vector of b_i. It can be shown that $\nu = p + 1$ still leads to a uniform prior on the correlations.

This prior seems to solve the main issues described earlier. The main disadvantage however is that taking draws from the posterior distribution of the mixed logit model parameters, which is easy in case of Inverse Wishart priors, becomes much more involved because of the transformation in step (ii). Therefore, this prior cannot be used in standard estimation algorithms.

The recently developed software, STAN, which uses Hamiltonian Monte Carlo (HMC) (a Metropolis strategy that manages all parameters simultaneously to take draws more efficiently from the joint posterior distribution), can deal with this prior. In the STAN manual (Stan Development Team, 2016), the Separation Strategy is recommended with the following prior specifications:

(i*) $\sigma_i \overset{iid}{\sim} Cauchy^{+}(b_i, \xi_i)$

(ii*) $\mathbf{R} \propto \det |R|^{\eta - 1}$, the $LKJ(\eta)$ distribution with shape parameter η where LKJ stands for Lewandowski, Kurowicka and Joe (2009)

We will denote this distribution as $SS_{STAN}(\eta, \boldsymbol{b}, \boldsymbol{\xi})$. As Gelman (2006) suggested, the half-Cauchy distribution with mode b_i, and a large scale value ξ_i is used here as prior for standard deviations in order to have a long tailed positive distribution. It is possible to use a simple distribution as prior for the correlation, as the HMC approach does not require the conjugacy property.

The shape parameter η controls the prior distribution of the correlations expected among

the taste parameters $\boldsymbol{\beta}_n$. With $\eta = 1$ one gets uniformly distributed correlation coefficients. With $\eta > 1$, the correlations have a unimodal distribution around zero and as η increases, the correlations are more likely to have values close to zero. So, the LKJ prior is easier to tune than choosing the hyperparameters of an Inverse Wishart prior and then transforming it to a correlation matrix.

3.3 Scaled Inverse Wishart distribution

To avoid the computational problems that are related to the Separation Strategy of Barnard *et al.* (2000), O'Malley and Zaslavsky (2008; published as a working paper in 2005) proposed a Scaled Inverse Wishart (SIW) prior, which also uses a decomposition approach as in the Separation Strategy but avoids the problematic transformation step. Let the covariance matrix be $\boldsymbol{\Sigma} = \boldsymbol{\Delta}\boldsymbol{\Phi}\boldsymbol{\Delta}$ and assume for

(i) $\boldsymbol{\Delta}$ a diagonal matrix with elements δ_i and $log(\delta_i) \overset{iid}{\sim} \mathcal{N}(b_i, \xi_i^2)$

(ii) $\boldsymbol{\Phi}$ a classical $IW(\nu, \mathbf{T})$ distribution

We denote this prior as $SIW(\nu, \mathbf{T}, \boldsymbol{b}, \boldsymbol{\xi})$. By definition, $\boldsymbol{\Delta}$ and $\boldsymbol{\Phi}$ now jointly determine the standard deviations as $\sigma_i = \delta_i\sqrt{\Phi_{ii}}$, but only $\boldsymbol{\Phi}$ determines the correlations as $\rho_{ij} = \Phi_{ij}/\sqrt{\Phi_{ii}\Phi_{jj}}$. Gelman and Hill (2007) suggested to use for $\boldsymbol{\Phi}$ an $IW(\nu = p + 1, \mathbf{I}_p)$ yielding uniform priors on $[-1, 1]$ for the correlations.

To avoid the tuning of this many parameters, it is often assumed that $log(\delta_i) \overset{iid}{\sim} \mathcal{N}(0, 1)$. This provides for rather large values for the standard deviations which may cause convergence problems in MCMC if the true heterogeneity is small.

The main advantage of SIW is that one can set priors on standard deviations and on correlation coefficients semi-separately which gives much more flexibility than the default IW prior. A disadvantage of SIW is that the dependency between correlations and standard deviations has been diminished but is not completely eliminated.

3.4 Huang Half-t distribution

Huang and Wand (2013) proposed a hierarchical approach for the prior of a covariance matrix, instead of decomposing it. They start from an Inverse Wishart distribution and therefore retain the tractable properties of this prior in MCMC. They use the prior $IW(\nu + p - 1, 2\nu\boldsymbol{\Delta})$ with $\boldsymbol{\Delta}$ a diagonal matrix with elements λ_i which are assumed to be independently distributed as $Gamma\left(\frac{1}{2}, \frac{1}{\xi_i^2}\right)$. We use the notation $Hht(\nu, \boldsymbol{\xi})$ to refer to this prior.

It was shown in Huang and Wand (2013) that this prior $\boldsymbol{\Sigma} \sim Hht(\nu, \boldsymbol{\xi})$ results in Half-t distributions with ν degrees of freedom and scale parameter ξ_i for the standard deviations. Half-t priors for the standard deviations were recommended by Gelman(2006) as non-informative priors which do allow for small values. The higher the value of ξ_i, the larger the uncertainty about the standard deviations. Furthermore, $\nu = 2$ still leads to a uniform prior for the correlation coefficients.

3.5 Visualizing the priors for covariance matrices

To illustrate and compare the properties of the Inverse Wishart, the Separation Strategy, the Scaled Inverse Wishart and the Huang Half-t priors, we follow the STAN Manual (2016) and Alvarez, *et al.* (2014) who selected parameters to get uniform priors for the correlations and similar medians for the prior distributions of the standard deviations, see Table 2. Remark that we assume here the same prior density for all standard deviations although the last four priors can easily cope with different knowledge for the various σ_i's.

The corresponding prior distributions of the standard deviations and the correlation coef-

Table 2: Hyperparameters given in Alvarez, *et al.*(2014) and STAN Manual (2016)

Prior distribution	Hyperparameters for this prior distribution
$IW(\nu, \boldsymbol{T})$	$\nu = p + 1, \mathbf{T} = \mathbf{I}_p$
$SS(\nu, \mathbf{T}, \boldsymbol{b}, \boldsymbol{\xi})$	$\nu = p + 1, \mathbf{T} = \mathbf{I}_p, b_i = \log(0.72)/2, \xi_i = 1$
$SS_{STAN}(\eta, \boldsymbol{b}, \boldsymbol{\xi})$	$\eta = 1, b_i = 0.72, \xi_i = 1$
$SIW(\nu, \mathbf{T}, \boldsymbol{b}, \boldsymbol{\xi})$	$\nu = p + 1, \mathbf{T} = 0.8\mathbf{I}_p, b_i = 0, \xi_i = 1$
$Hht(\nu, \boldsymbol{\xi})$	$\nu = 2, \xi_i = 1.04$

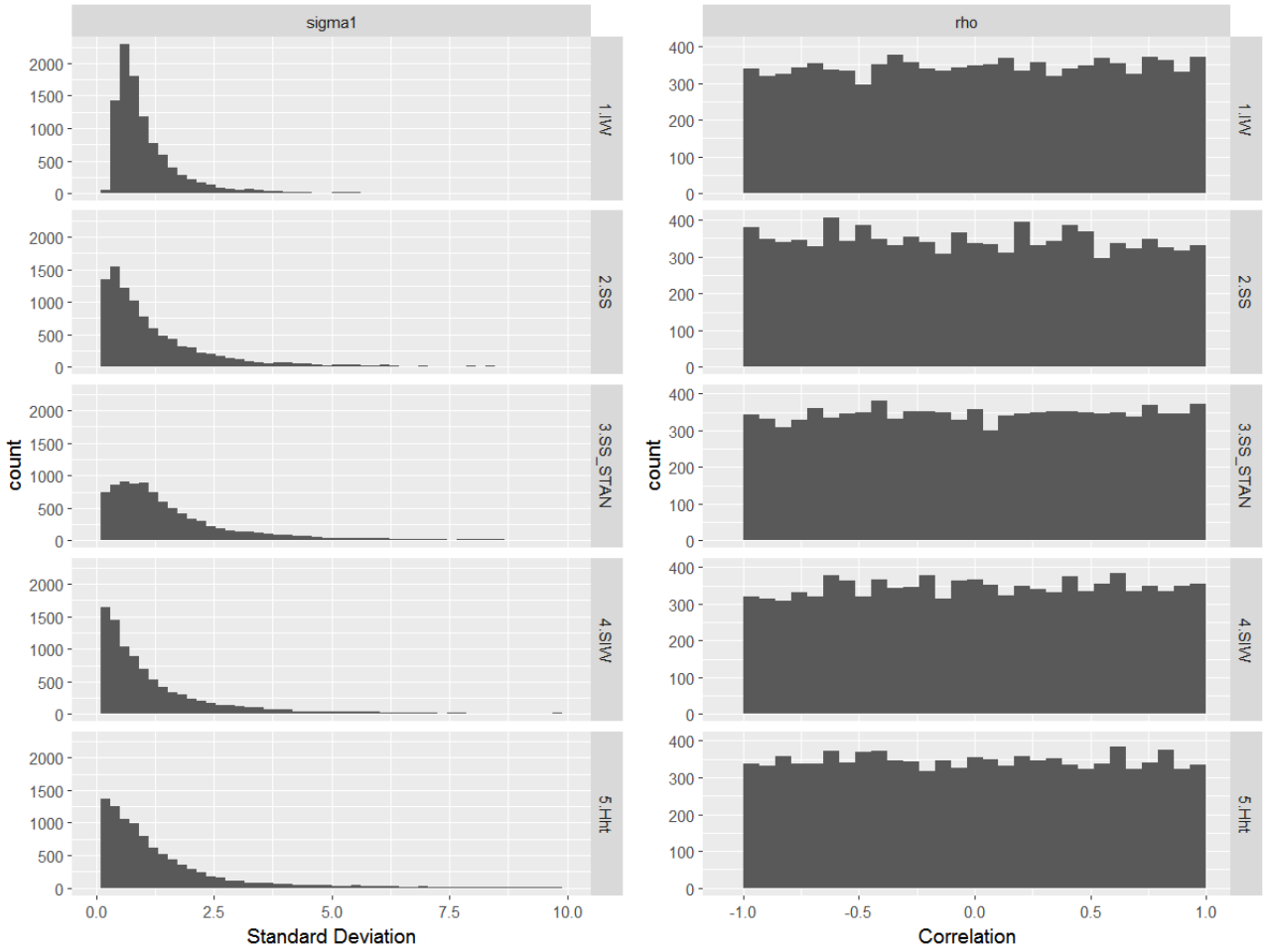

Figure 3: Prior distributions of the first standard deviation (σ_1) and the correlation (ρ_{12}) implied by the different priors on the covariance matrix in case $p = 2$ (based on 10000 draws); Prior Distributions: (i) $IW = IW(\nu, \mathbf{I}_p)$; (ii) $SS = SS(\nu, \mathbf{T}, \mathbf{b}, \mathbf{I}_p)$; (iii) $SS_{STAN}(\eta, \boldsymbol{b}, \boldsymbol{\xi})$ (iv) $SIW = SIW(\nu, \boldsymbol{\Lambda}, \boldsymbol{b}, \boldsymbol{\xi})$; (v) $Hht = Hht(\nu, \boldsymbol{\xi})$.

ficient are shown in Figure 3. The left panel shows that the last four distributions indeed all allow for very small standard deviations and cover a large range of possible values. The right panel of Figure 3 reveals that all prior distributions can indeed provide uniform priors for the correlation coefficients.

Figure 4 shows the correlation pattern between the correlation coefficient and the standard deviation for the priors in Table 2. It is clear that only for the Separation Strategy approaches there is no dependence between the correlation coefficients and the variance components but the scaled IW and the Hht prior diminish this dependence compared to the IW prior.

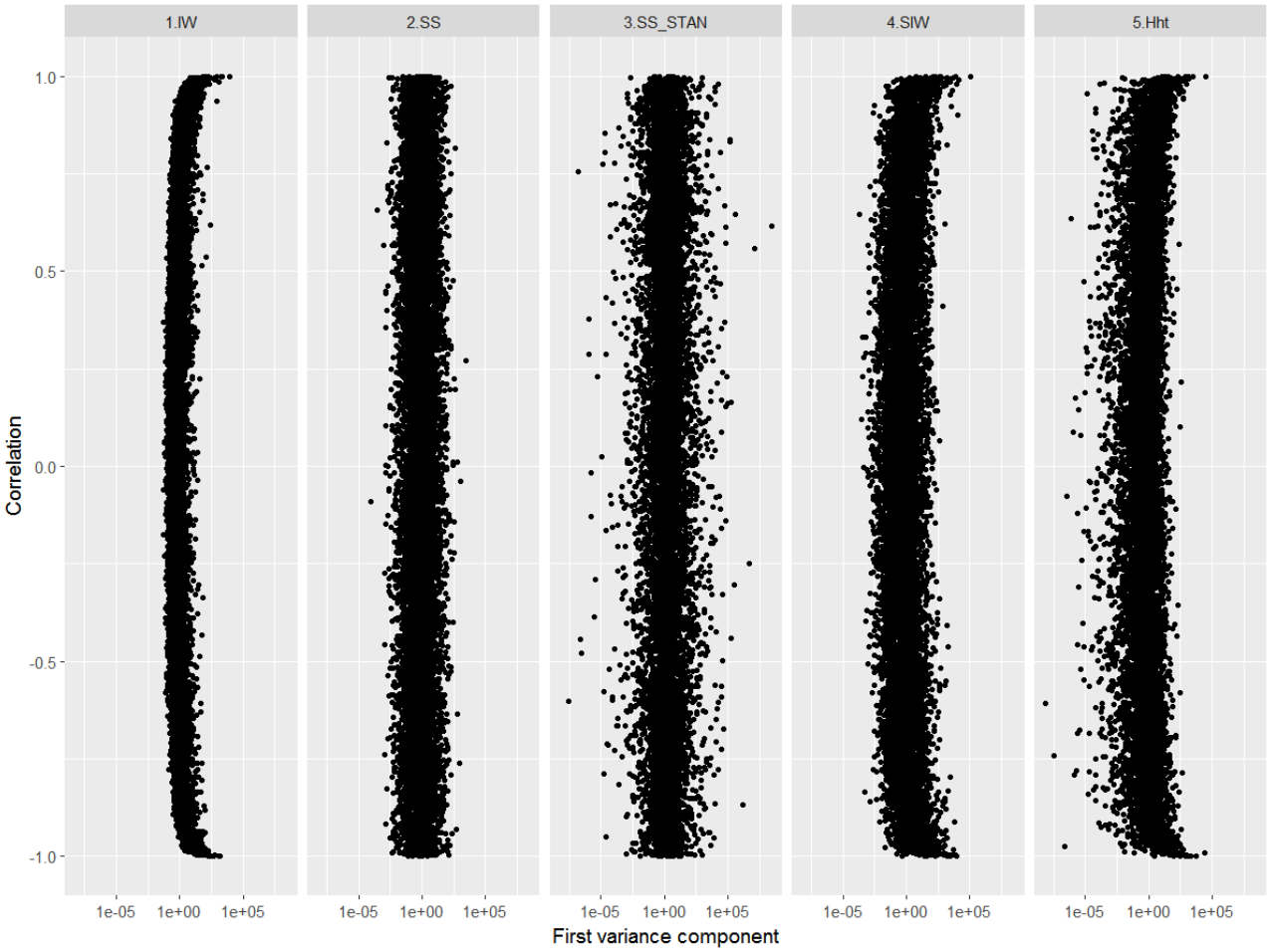

Figure 4: Scatterplot of the correlation coefficient vs. the first variance component implied by the different priors on the covariance matrix in case $p = 2$ (based on 10000 draws); Prior Distributions: (i) $IW = IW(\nu, \mathbf{I}_p)$; (ii) $SS = SS(\nu, \mathbf{T}, \mathbf{b}, \mathbf{I}_p)$; (iii) $SS_{STAN}(\eta, \boldsymbol{b}, \boldsymbol{\xi})$ (iv) $SIW = SIW(\nu, \boldsymbol{\Lambda}, \boldsymbol{b}, \boldsymbol{\xi})$; (v) $Hht = Hht(\nu, \boldsymbol{\xi})$.

3.6 Implementation of the Separation Strategy, Scaled Inverse Wishart and Huang Half-t priors in Bayesian estimation

3.6.1 Using the Separation Strategy prior in STAN

As stated before, the Separation Strategy prior can easily be implemented in STAN software which uses the Hamiltonian Monte Carlo method to draw values from the joint posterior distribution as to update all parameters simultaneously. Interested readers may turn to the STAN manual (Stan Development Team, 2016) for a detailed description and extensive examples on the features of the HMC. As the details of the HMC approach is beyond the scope of this study, we only provide the R-STAN codes for the implementation of SS priors to estimate the mixed logit model on the website (https://github.com/dakinc/estimation-comparison/). Ben-Akiva

et al. (2015), have also used R-STAN with SS prior by setting flat distributions for the standard deviations and the correlations and have explained the procedure in detail. They have found that the standard Gibbs sampling combined with a Metropolis-Hastings step outperformed the Hamiltonian Monte Carlo in terms of computation time.

The Scaled Inverse Wishart and Huang Half-t priors can easily be implemented into a regular MCMC algorithm. The only difference between using these priors and the IW priors occurs at the computation of the conditional posterior distribution of $\mathbf{\Sigma}$, as the conditional posterior of $\boldsymbol{\mu}$ and $\boldsymbol{\beta}_n$ remain the same as given in section 2.2.

3.6.2 Gibbs sampling with the Scaled Inverse Wishart prior

In the Scaled Inverse Wishart prior specification, $SIW(\nu, \mathbf{T}, \boldsymbol{b}, \boldsymbol{\xi})$, the prior for $\mathbf{\Sigma}$ is expressed as $\mathbf{\Sigma} = \mathbf{\Delta\Phi\Delta}$. A combination of a Gibbs step and a Metropolis-Hastings step is used to update $\mathbf{\Sigma}$ in this case.

- The conditional posterior distribution of $\mathbf{\Phi}$, given $\boldsymbol{\beta}_n$, $\boldsymbol{\mu}$ and $\mathbf{\Delta}$ is $IW(\nu + N,\ \mathbf{T} + \mathbf{\Delta}^{-1}\left[\sum_{n=1}^{N}(\boldsymbol{\beta}_n - \boldsymbol{\mu})(\boldsymbol{\beta}_n - \boldsymbol{\mu})'\right]\mathbf{\Delta}^{-1})$.

- The log-conditional posterior distribution of σ_i is given by O'Malley and Zaslavsky (2008) as

$$\log f(\sigma_i) = constant - (N+1)\log(\delta_i) - [\mathbf{\Phi}^{-1}]_{ii}S_{ii}/(2\sigma_i^2) - \frac{1}{\sigma_i}\sum_{j\neq i}[\mathbf{\Phi}^{-1}]_{ij}S_{ij}/(\sigma_j) - (\log(\delta_i) - b_i)^2/(2\xi_i^2)$$

  where S_{ij} is the ij^{th} element of $\left[\sum_{n=1}^{N}(\boldsymbol{\beta}_n - \boldsymbol{\mu})(\boldsymbol{\beta}_n - \boldsymbol{\mu})'\right]$ and $[\mathbf{\Phi}^{-1}]_{ij}$ is the ij^{th} element of $\mathbf{\Phi}^{-1}$. In order to update the value σ_i, a Metropolis-Hastings random walk procedure is used with a proposal distribution to simulate the target posterior distribution given above. The proposal distribution, which has been used in the literature O'Malley and Zaslavsky (2008), is the logarithm of a t-distribution with 3 degrees of freedom, with the current value of σ_i^{t-1} as location parameter and with a scale parameter which is tuned to make the acceptance rate around 0.44 (the details on this optimal rate are given in Gelman, *et al.*, 1996).

- By using the draws of $\mathbf{\Phi}$ and σ_i, the new draw of $\mathbf{\Sigma}$ is obtained by $\mathbf{\Sigma} = \mathbf{\Delta}\mathbf{\Phi}\mathbf{\Delta}$.

3.6.3 Gibbs sampling with the Huang Half-t prior

In case of a Huang Half-t prior $Hht(\nu, \boldsymbol{\xi})$ for $\mathbf{\Sigma}$, the following Gibbs steps are used.

- The conditional posterior distribution of $\mathbf{\Sigma}$, given $\boldsymbol{\beta}_n$, $\boldsymbol{\mu}$ and λ_i, is $IW(\nu + N + p - 1,\ 2\nu\mathbf{\Delta} + \sum_{n=1}^{N}(\boldsymbol{\beta}_n - \boldsymbol{\mu})(\boldsymbol{\beta}_n - \boldsymbol{\mu})')$, where $\mathbf{\Delta}$ a diagonal matrix with elements λ_i.
- As the prior distribution of λ_i is $Gamma\left(\frac{1}{2}, \frac{1}{\xi_i^2}\right)$, the conditional posterior distribution of λ_i, given $\boldsymbol{\beta}_n$, $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ is $Gamma\left(\frac{\nu+p}{2}, \frac{1}{\xi_i^2} + \nu(\mathbf{\Sigma}^{-1})_{ii}\right)$ where $(\mathbf{\Sigma}^{-1})_{ii}$ denotes the ii^{th} element of the inverse of $\mathbf{\Sigma}$.

The R-codes for the MCMC algorithm with IW, SIW and Hht priors can be obtained from https://github.com/dakinc/estimation-comparison.

4 Simulation Study

In this section, we compare the results of Bayesian estimation with various prior distributions for the covariance matrix and of maximum simulated likelihood estimation with different numbers of Halton draws. To be able to generalize the results, we consider several scenarios following Arora and Huber (2001), Toubia *et al.* (2004) and Yu *et al.* (2010). They consider a low and a high level for both the response accuracy and the consumer heterogeneity to cover a large range of situations. We simulate choices for 200 respondents based on 18 choice sets consisting of 3 alternatives. Each alternative is described by 3 attributes with 3 levels ($3^3/3/18$). These choice sets are generated by minimizing the local $\mathcal{D}$-error for the multinomial logit model (MNL). Effects coding is used to represent the attribute levels.

The individual-level parameters, $\boldsymbol{\beta}_n$, which are used to simulate the choices, are assumed to come from a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance matrix $\mathbf{\Sigma}$. The values for $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ that are used in the simulation study are listed in Table 3a for the scenarios without correlation and in Table 3b for the scenarios with correlation. Remark that p, the number of coefficients in the model, is 6.

Table 3a: Scenarios without correlation in the covariance matrix

True mean vector ($\boldsymbol{\mu}$)	**True heterogeneity** ($\boldsymbol{\Sigma}$)	
	Low Heterogeneity	High Heterogeneity
Low response accuracy $0.5 \times (-1, 0, -1, 0, -1, 0)$	$0.25\mathbf{I}_p$	$1.0\mathbf{I}_p$
High response accuracy $2.0 \times (-1, 0, -1, 0, -1, 0)$	$1.0\mathbf{I}_p$	$4.0\mathbf{I}_p$

Table 3b: Covariance matrices with correlations

True mean vector ($\boldsymbol{\mu}$)	**True heterogeneity** ($\boldsymbol{\Sigma}$)
$2.0 \times (-1, 0, -1, 0, -1, 0)$	$\mathbf{I}_p + \alpha \times \begin{bmatrix} 0.0 & 0.0 & 1.0 & 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 1.0 \\ 1.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 1.0 \\ 1.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 & 1.0 & 0.0 & 0.0 \end{bmatrix}$ $\alpha = 0.1$ for medium correlation $\alpha = 0.5$ for medium correlation $\alpha = 0.9$ for high correlation

As in Arora and Huber (2001), we define the values of the variance relative to the magnitude of the mean vector. For high respondent heterogeneity, the variance is twice the population mean and in case of low respondent heterogeneity, it is set to half the population mean. Remark that the MNL local optimal design were also constructed using the true parameter values.

For the maximum simulated likelihood approach, we use the R-package **gmnl** with 100, 400 and 1000 Halton draws. We use the R-code and R-STAN code that were described in section 3.6, to get the estimates based on the Bayesian approach. For the Bayesian estimation, we use the standard prior for the mean: a multivariate normal distribution with zero mean vector and large variances, $100\mathbf{I}_p$.

For each design scenario, we generate an optimal design, simulate choices 10 times and estimate the mixed logit model for each repetition by using the maximum simulated likelihood and the Bayesian approaches. We compute the estimates for the mean vector, the full covariance matrix and the individual parameters. To measure the accuracy of the estimates obtained with these different estimation methods, we compute the relative root mean square errors (RRMSE)

which measure the error of the estimates relative to the true parameters. This measure is given by

$$\text{RRMSE}_{\boldsymbol{\theta}} = \sqrt{\frac{1}{p_\theta}\left(\frac{\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}}{\boldsymbol{\theta}}\right)'\left(\frac{\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}}{\boldsymbol{\theta}}\right)}, \tag{9}$$

with $\boldsymbol{\theta}$ the non-zero true parameter values that we use to simulate the choices, $\hat{\boldsymbol{\theta}}$ contains the corresponding estimates and p_θ is the total number of non-zero parameters to be estimated. We evaluate the estimation accuracy of $\boldsymbol{\mu}$, the stacked unique elements of $\boldsymbol{\Sigma}$ and the individual preference parameters $\boldsymbol{\beta}_n$. We also computed the root mean square errors (RMSE) for these estimates where we could also take the zero elements of the parameters into account and obtain similar patterns as with the RRMSE values. As RMSE values are scale-dependent, we prefer to report the RRMSE values which make it easier to compare the performances across the different scenarios. We also calculated the root mean square errors separately for the standard deviations and for the correlations, but these results are not reported as they do not give extra insights.

Table 4: Priors on the covariance matrix for Bayesian estimation of the mixed logit model

Type of prior distribution	Hyperparameters for this prior distribution
$IW1(\nu, \mathbf{T})$	$\nu = p, \mathbf{T} = \mathbf{I}_p$
$IW2(\nu, \mathbf{T})$	$\nu = p, \mathbf{T} = \nu\mathbf{I}_p$
$IW3(\nu, \mathbf{T})$	$\nu = p+1, \mathbf{T} = \nu\mathbf{I}_p$
$IW4(\nu, \mathbf{T})$	$\nu = p+3, \mathbf{T} = \nu\mathbf{I}_p$
$IW5(\nu, \mathbf{T})$	$\nu = p+4, \mathbf{T} = \nu\mathbf{I}_p$
$IW6(\nu, \mathbf{T})$	$\nu = 0.5p(p+1), \mathbf{T} = 0.1\nu\mathbf{I}_p$
$SS1_{BMM}(\nu, \mathbf{T}, \mathbf{b}, \boldsymbol{\xi})$	$\nu = p+1, \mathbf{T} = \nu\mathbf{I}_p, b_i = 0, \xi_i = 1$
$SS2_{BMM}(\nu, \mathbf{T}, \mathbf{b}, \boldsymbol{\xi})$	$\nu = p+4, \mathbf{T} = \nu\mathbf{I}_p, b_i = 0, \xi_i = 0.5$
$SS3_{STAN}(\eta, \mathbf{b}, \boldsymbol{\xi})$	$\eta = 1, b_i = 0, \xi_i = 2.5$
$SS4_{STAN}(\eta, \mathbf{b}, \boldsymbol{\xi})$	$\eta = 50, b_i = 0, \xi_i = 2.5$
$SIW1(\nu, \mathbf{T}, \mathbf{b}, \boldsymbol{\xi})$	$\nu = p+1, \mathbf{T} = \nu\mathbf{I}_p, b_i = 0, \xi_i = 0.1$
$SIW2(\nu, \mathbf{T}, \mathbf{b}, \boldsymbol{\xi})$	$\nu = p+4, \mathbf{T} = 0.5\nu\mathbf{I}_p, b_i = 0, \xi_i = 0.1$
$Hht1(\nu, \boldsymbol{\xi})$	$\nu = 2, \xi_i = 1$
$Hht2(\nu, \boldsymbol{\xi})$	$\nu = 2, \xi_i = 0.5$
$Hht3(\nu, \boldsymbol{\xi})$	$\nu = p+4, \xi_i = 1$

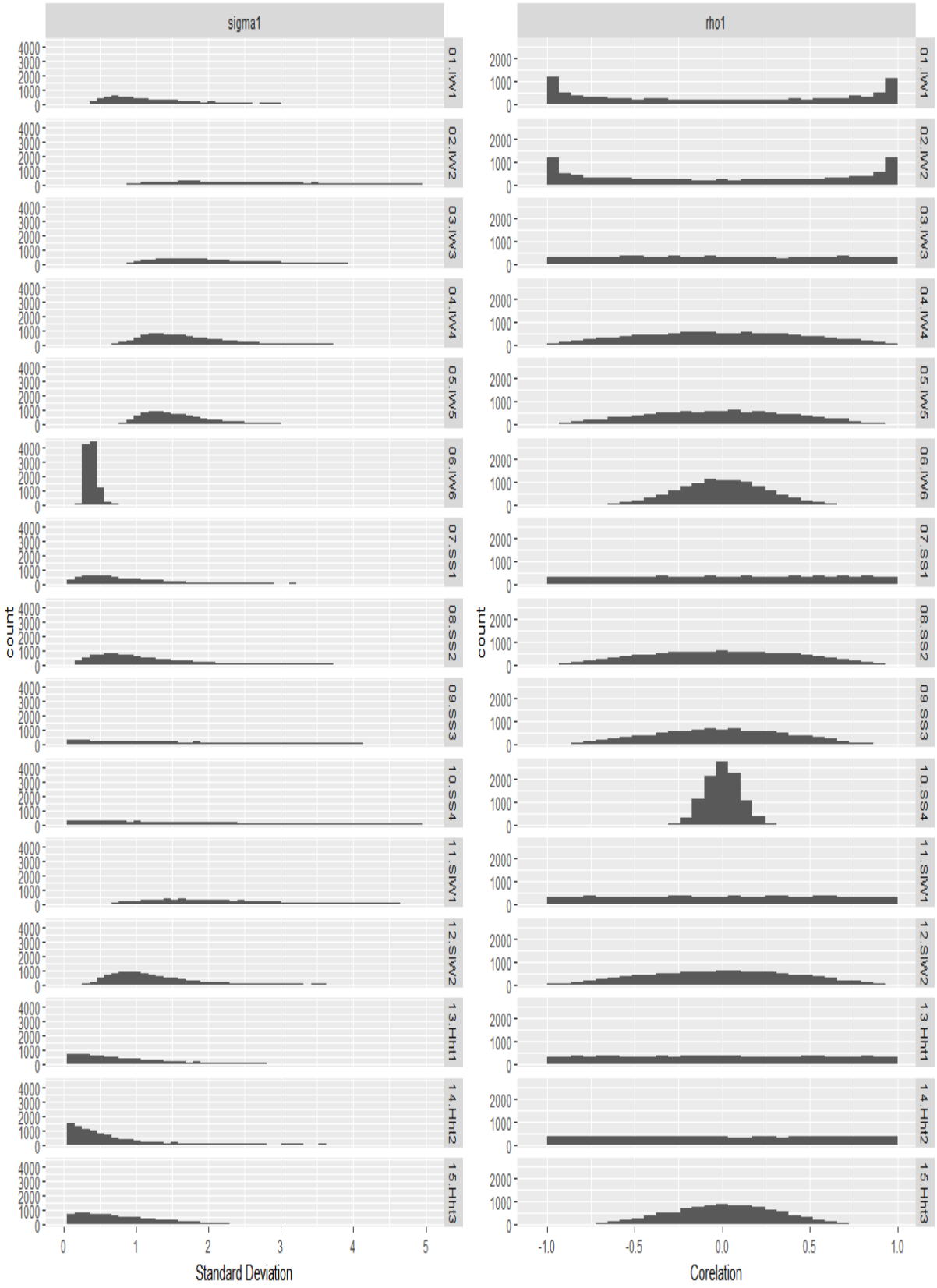

Figure 5: Prior distributions of the first standard deviation (σ_1) and the first correlation coefficient (ρ_{12}) implied by the different priors on the covariance matrix in case $p = 2$ given in the same order as in Table 4 (based on 10000 draws).

We use various priors for the covariance matrix which are listed in Table 4. We consider six Inverse Wishart distributions that have often been used in the discrete choice literature, four Separation Strategy approaches, two Scaled Inverse Wishart distribution and three Huang Half-t distributions.

To illustrate the differences between these priors, we plot in Figure 5 the corresponding prior densities of the first standard deviation and the first correlation coefficient. It is clear from these plots that some of the priors do not allow for small values of the standard deviations and are expected to perform badly when the true variances are quite small. On the other hand, some priors can be expected to perform better in the high heterogeneity case.

It is also clear from the graphs that one can model variances and correlation coefficients separately by using SS prior. While allowing for large values for the variances, correlations can be kept close to zero, by using for instance SS with LKJ(50). SIW priors can scale down the variance components to some degree while keeping a flat prior for the correlation coefficients. However, Hht priors allow for even smaller values for the variances while allowing for larger correlation values.

4.1 Results in case there is no correlation

Figure 6 shows the $\text{RRMSE}_{\boldsymbol{\Sigma}}$ values obtained for all scenarios in Table 3a with on top the true density of the first coefficient in $\boldsymbol{\beta}_n$ to visualize the different scenarios. The boxplots corresponding to Bayesian estimation are shown first in the same order as in Table 4. The last three boxplots in each figure correspond to MSL estimation with different numbers of Halton draws.

As can be seen from Figures 6, the number of draws has most impact on the MSL results if the heterogeneity is high as a larger parameter space has to be covered. It is clear that in this case, the small default number of draws in most software packages are insufficient (see Table 1). In almost all cases, the Bayesian approach, even with misspecified priors, leads to more precise estimates than the MSL approach, even when using 1000 Halton draws which is much more than the default settings in most packages. Next, we look at how Bayesian methods perform with different priors on covariance matrix in detail.

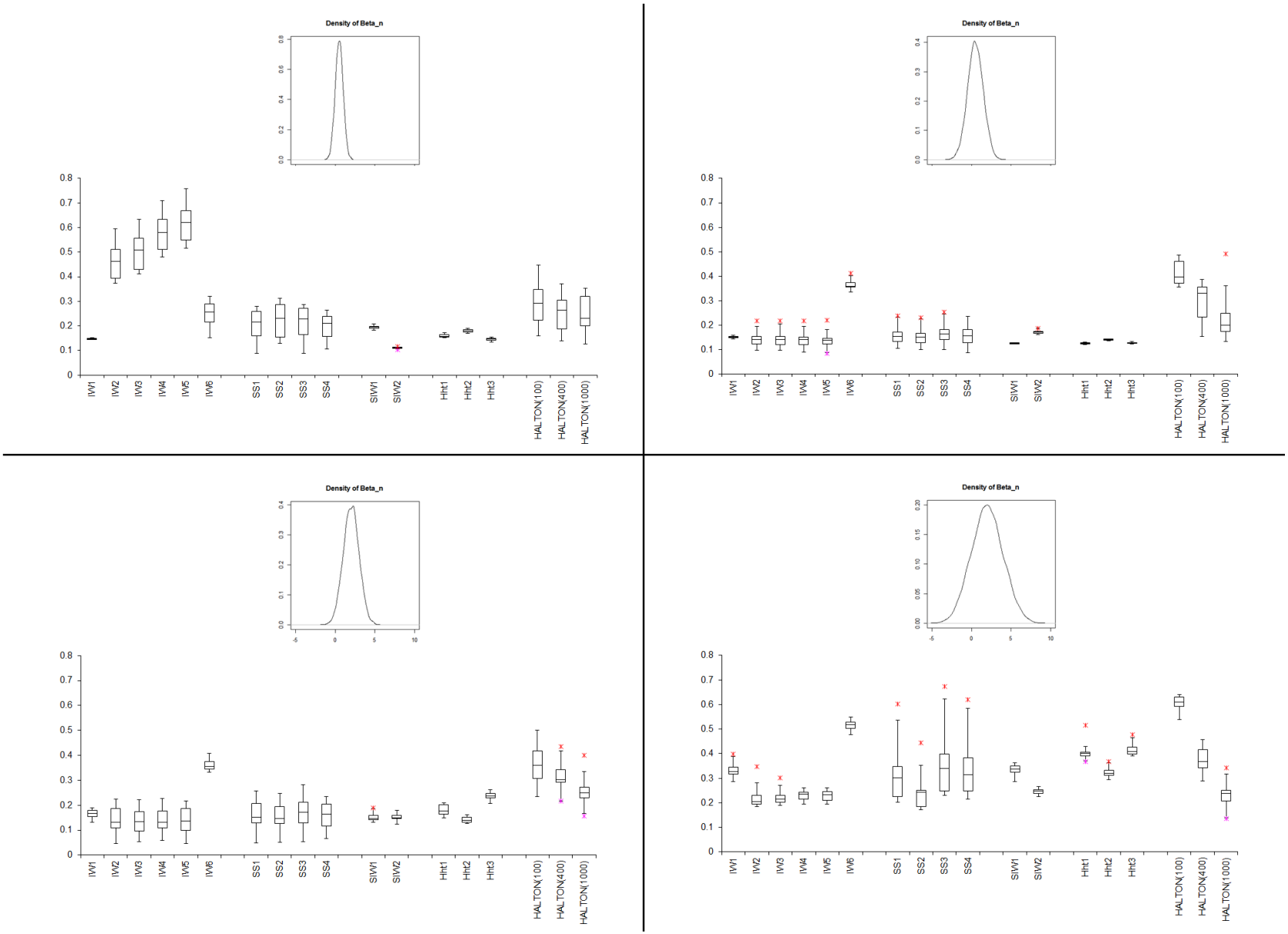

Figure 6: $\text{RRMSE}_{\boldsymbol{\Sigma}}$ values using the prior distributions of Table 4 in Bayesian estimation and using 100, 400 and 1000 Halton draws in MSL estimation. The plots at the top correspond to low accuracy and the plots at the bottom to high accuracy, at the left hand side there is low heteroscedasticiy and at the right there is high heteroscedasticity. The true parameter values are given in Table 3a (no correlation) and the corresponding densities of true individual parameters for each scenarios are plotted.

In general, we get the results that can be expected: IW2-IW5 do not perform well if the true variance is quite small, as these priors do not allow for values close to zero. IW6 has a very limited range for the standard deviations so it performs well if the range is well specified but is bad otherwise. The Separation Strategy, the Scaled IW and the Huang Half-t allow for small values as well as larger values and therefore perform quite well in most cases.

To get more insight, we look in more detail to the results of the two most extreme scenarios, in the top left panel (low accuracy and low heterogeneity) and the bottom right panel (high accuracy and high heterogeneity). For the first scenario, we expect the priors that have high densities for values close to zero for both the standard deviations and the correlations, such as IW6, SS4, SIW2 and Hht3, to outperform the other priors and this is confirmed by the boxplots in the top left panel. Furthermore, it turns out that all the selected SIW and Hht priors give very reliable results too.

For the latter scenario, we expect the priors with long-tailed distribution for standard deviations and with the density for correlations close to zero, such as IW2-IW5, SS2 and SIW2, to outperform the other priors. It is seen from Figure 6, these IW priors perform well when the true variances are quite large. However, the estimates obtained from SIW2 have less variability in contrast to IW priors and SS2.

We also computed the boxplots of $\text{RRMSE}_{\boldsymbol{\mu}}$ and $\text{RRMSE}_{\boldsymbol{\beta}}$ values which can be found in Appendix A. The boxplots of the $\text{RRMSE}_{\boldsymbol{\mu}}$ values show similar patterns as those of the $\text{RRMSE}_{\boldsymbol{\Sigma}}$ values that we have discussed before, but are less dependent on the prior for $\boldsymbol{\Sigma}$. The $\text{RRMSE}_{\boldsymbol{\beta}}$ values do not seem to be affected by the different priors for $\boldsymbol{\Sigma}$ at all.

4.2 Results in case there is correlation

In order to see the effect of the different priors on the correlation structure, we generate data with different levels of correlation as summarized in Table 3b. Here, the true covariance matrix consists of standard deviations of 1 and an equal number of non-zero correlations and zero correlations. We only show the results for $\alpha = 0.9$, as similar remarks can be made for other values of α.

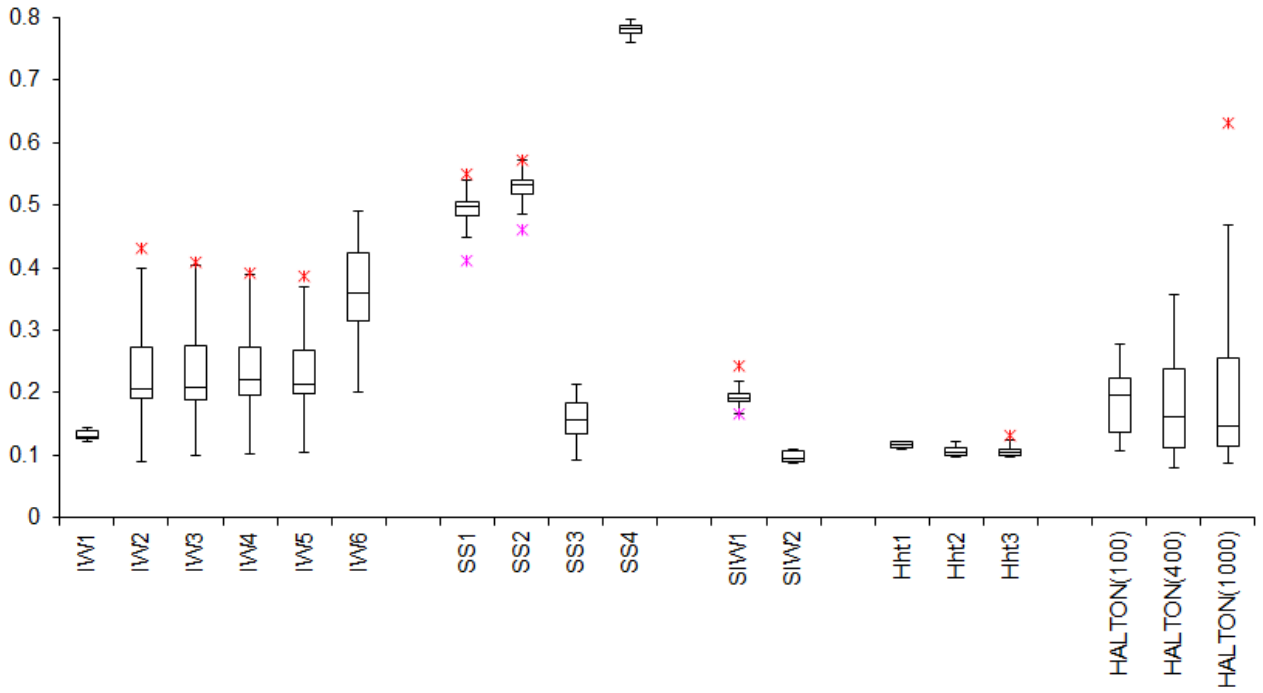

Figure 7: $\text{RRMSE}_{\boldsymbol{\Sigma}}$ values using the prior distributions of Table 3b in Bayesian estimation and using 100, 400 and 1000 Halton draws in MSL estimation in the scenario with correlation (see Table 3b).

The conclusions about the MSL estimation drawn in the no correlation cases are also valid here, though the MSL estimates now exhibit more variability than in the previous case. When we look at the performance of the Bayesian method with different priors on the covariance matrix, the priors with high density around one for the standard deviations and with a large range of likely values for the correlations are expected to perform better. Based on Figure 5, we can for instance expect IW1 and SIW2 to perform well and IW6 and SS4 to perform badly. Figure 7 confirms that this is indeed the case. Also, the three Hht priors perform quite well although the last one does not allow for large correlations.

5 Discussion and conclusions

In this paper, we compared the Bayesian approach and maximum simulated likelihood (MSL) to estimate the mixed logit model. The precision of MSL and Bayesian estimates highly depends on the number of draws in MSL and on the prior distributions in Bayesian estimation.

There are many studies in the literature that compare different types of Quasi-Monte Carlo random(QMC) draws with Pseudo-Monte Carlo (PMC) random draws. Though these studies strongly recommend the QMC draws over PMC random draws, many software packages still use the PMC draws by default. Even more problematic, the default number of draws of many software packages are way too low to get reliable estimates.

The software packages for Bayesian estimation use the Inverse Wishart distribution as the default prior for the covariance matrix because of the conjugacy property. However, this prior has some undesirable properties. The default IW prior settings do not allow for small variances. Another issue of IW priors is that there is a strong dependency among the components of the covariance matrix, such that larger variances are associated with extreme correlation values and small variances are associated with small correlations.

We compared the results obtained with IW priors with some alternative priors that have been proposed in the literature: the Separation Strategy (SS), the Scaled Inverse Wishart (SIW) and the Huang Half-t (Hht) priors. The Separation strategy approach is more flexible as it models variances and correlations components separately. This prior enables to have very different prior distributions for each component of the covariance matrix, however, deciding on a proper prior distribution becomes quite complex. Additionally, this prior cannot be used in a standard Gibbs sampling combined with Metropolis-Hastings steps and therefore requires non-standard software and longer computation times.

Both SIW and Hht priors are more flexible than IW priors, but the dependency between the variances and the correlation components is not completely eliminated. When using a SIW prior, one needs to select extra hyperparameters. The Huang Half-t priors are more straightforward to use as less hyperparameters have to be chosen.

Based on our results, we strongly recommend that the reliability of the parameter estimates is always checked. With MSL estimation, this can easily be done by increasing the number of draws until convergence. In a Bayesian approach, we suggest to try different and more flexible priors to check the robustness of the results.

Appendix A: The boxplots of RRMSE_{μ} and RRMSE_{β} values for all scenarios in Table 3a:

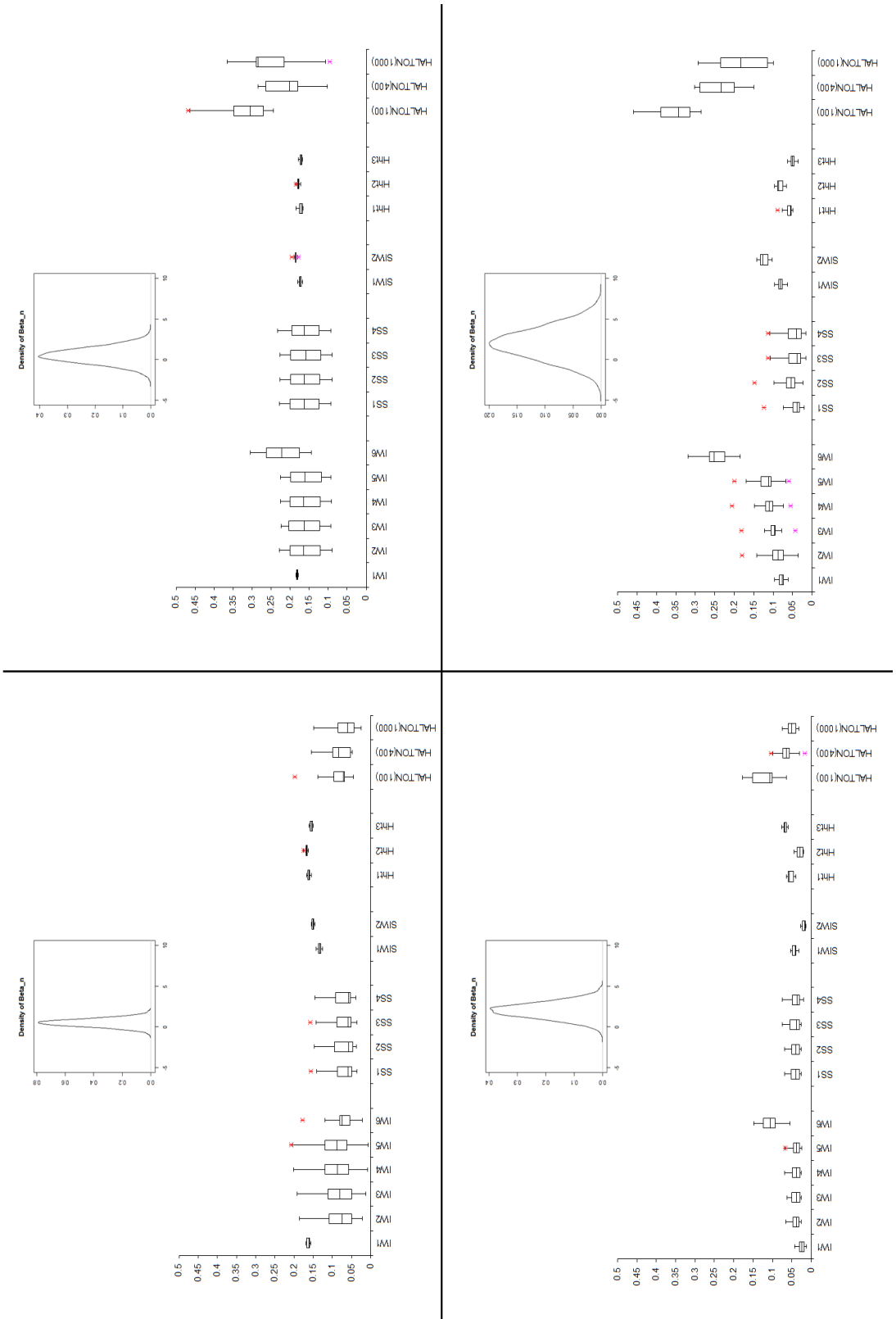

Figure 8: RRMSE_{μ} values using the prior distributions of Table 4 in Bayesian estimation and using 100, 400 and 1000 Halton draws in MSL estimation. The plots at the top correspond to low accuracy and the plots at the bottom to high accuracy, at the left hand side there is low heteroscedasticiy and at the right there is high heteroscedasticity. The true parameter values are given in Table 3a (no correlation).

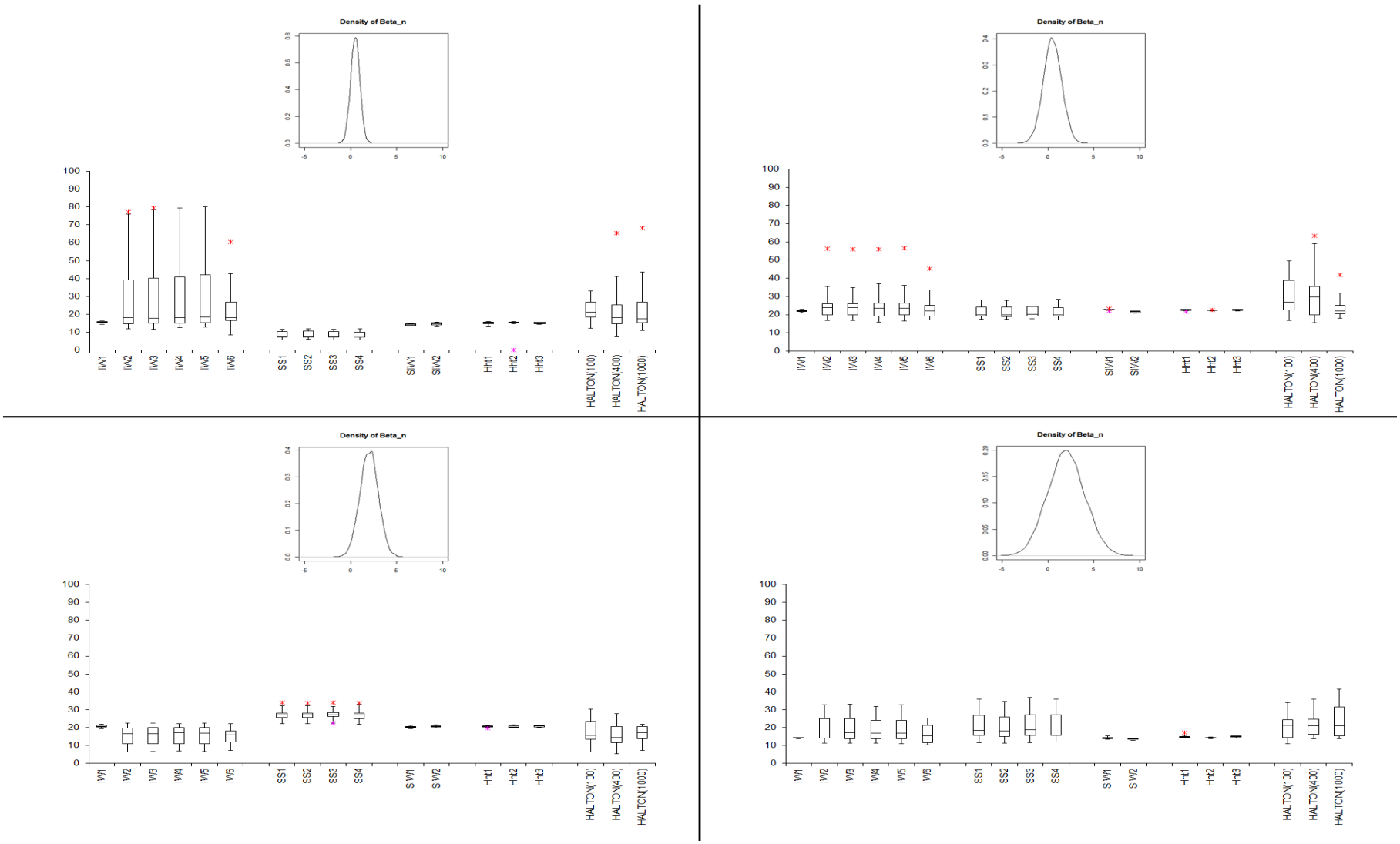

Figure 9: $\text{RRMSE}_{\boldsymbol{\beta}}$ values using the prior distributions of Table 4 in Bayesian estimation and using 100, 400 and 1000 Halton draws in MSL estimation. The plots at the top correspond to low accuracy and the plots at the bottom to high accuracy, at the left hand side there is low heteroscedasticiy and at the right there is high heteroscedasticity. The true parameter values are given in Table 3a (no correlation).

References

Allenby, G. (1997). An introduction to hierarchical Bayesian modeling. In Tutorial notes, Advanced Research Techniques Forum, American Marketing Association.

Alvarez, I., Niemi, J., & Simpson, M. (2014). Bayesian inference for a covariance matrix. Proceedings of the 26th Annual Conference on Applied Statistics in Agriculture, 71-82.

Arora, N., & Huber, J. (2001). Improving parameter estimates and model prediction by aggregate customization in choice experiments. Journal of Consumer Research, 28(2), 273-283.

Balcombe, K., Chalak, A., & Fraser, I. (2009). Model selection for the mixed logit with Bayesian estimation. Journal of Environmental Economics and Management, 57(2), 226-237.

Barnard, J., McCulloch, R., & Meng, X. L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. Statistica Sinica, 1281-1311.

Barthelmé, S. (2012, March 7). Why an inverse-Wishart prior may not be such a good idea. Retrieved from https://dahtah.wordpress.com/2012/03/07/why-an-inverse-wishart-prior-may-not-be-such-a-good-idea/

Ben-Akiva, M., McFadden, D., & Train, K. (2015). Foundations of stated preference elicitation consumer behavior and choice-based conjoint analysis. Retrieved from https://eml.berkeley.edu/ train/foundations.pdf

Bhat, C. R. (2001). Quasi-random maximum simulated likelihood estimation of the

mixed multinomial logit model. Transportation Research Part B: Methodological, 35(7), 677-693.

Bliemer, M. C., Rose, J. M., & Hess, S. (2008). Approximation of Bayesian efficiency in experimental choice designs. Journal of Choice Modelling, 1(1), 98-126.

Elshiewy, O., Zenetti, G., & Boztug, Y. (2017). Differences Between Classical and Bayesian Estimates for Mixed Logit Models: A Replication Study. Journal of Applied Econometrics, 32(2), 470-476.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian analysis, 1(3), 515-534.

Gelman, A., & Hill, J. (2007). Data analysis using regression and multilevelhierarchical models (Vol. 1). New York, NY, USA: Cambridge University Press.

Gelman, A., Roberts, G. O., & Gilks, W. R. (1996). Efficient Metropolis jumping rules. Bayesian statistics, 5(599-608), 42.

Haan, P., Kemptner, D., & Uhlendorff, A. (2015). Bayesian procedures as a numerical tool for the estimation of an intertemporal discrete choice model. Empirical Economics, 49(3), 1123-1141.

Huang, A., & Wand, M. P. (2013). Simple marginally noninformative prior distributions for covariance matrices. Bayesian Analysis, 8(2), 439-452.

Huber, J., & Train, K. (2001). On the similarity of classical and Bayesian estimates of individual mean partworths. Marketing Letters, 12(3), 259-269.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. Journal of multivariate analysis, 100(9), 1989-2001.

O'Malley, A. J., & Zaslavsky, A. M. (2008). Domain-level covariance analysis for multilevel survey data with structured nonresponse. Journal of the American Statistical Association, 103(484), 1405-1418. (published as working paper 2005)

Regier, D. A., Ryan, M., Phimister, E., & Marra, C. A. (2009). Bayesian and classical estimation of mixed logit: an application to genetic testing. Journal of health economics, 28(3), 598-610.

Revelt, D., & Train, K. (1998). Mixed logit with repeated choices: households' choices of appliance efficiency level. Review of economics and statistics,80(4), 647-657.

Rossi, P. E., McCulloch, R. E., & Allenby, G. M. (1996). The value of purchase history data in target marketing. Marketing Science, 15(4), 321-340.

Rossi, P. E., Allenby, G. M., & McCulloch, R. E. (2005). Bayesian statistics and marketing (pp. 1-368). New York: Wiley.

Sándor, Z., & Train, K. (2004). Quasi-random simulation of discrete choice models. Transportation Research Part B: Methodological, 38(4), 313-327.

Simpson, M. (2012, August 20). R code to visualize the Inverse Wishart, Scaled IW and Separation Strategy: Prior distributions for covariance matrices: the scaled inverse-Wishart prior. Retrieved from http://www.themattsimpson.com/2012/08/20/prior-distributions

-for-covariance-matrices-the-scaled-inverse-wishart-prior/

Stan Development Team (2016). Stan Modeling Language Users Guide and Reference Manual, Version 2.15.0. http://mc-stan.org

Toubia, O., Hauser, J. R., & Simester, D. I. (2004). Polyhedral methods for adaptive choice-based conjoint analysis. Journal of Marketing Research, 41(1), 116-131.

Train, K. E., 2003. Discrete choice methods with simulation. Cambridge University Press.

Yu, J., Goos, P., & Vandebroek, M. (2010). Comparing different sampling schemes for approximating the integrals involved in the efficient design of stated choice experiments. Transportation Research Part B: Methodological, 44(10), 1268-1289.