Covariance Decompositions for Accurate Computation in Bayesian Scale-Usage Models

Chris Hans, Greg M. Allenby, Peter F. Craigmile,
Ju Hee Lee, Steven MacEachern, and Xinyi Xu

Abstract Analyses of multivariate ordinal probit models typically use data augmentation to link the observed (discrete) data to latent (continuous) data via a censoring mechanism defined by a collection of "cutpoints." Most standard models, for which effective Markov chain Monte Carlo (MCMC) sampling algorithms have been developed, use a separate (and independent) set of cutpoints for each element of the multivariate response. Motivated by the analysis of ratings data, we describe a particular class of multivariate ordinal probit models where it is desirable to use a common set of cutpoints. While this approach is attractive from a data-analytic perspective, we show that the existing efficient MCMC algorithms can no longer be accurately applied. Moreover, we show that attempts to implement these algorithms by numerically approximating required multivariate normal integrals over high-dimensional rectangular regions can result in severely degraded estimates of the posterior distribution. We propose a new data augmentation that is based on a covariance decomposition and that admits a simple and accurate MCMC algorithm. Our data augmentation requires only that univariate normal integrals be evaluated, which can be done quickly and with high accuracy. We provide theoretical results that suggest optimal decompositions within this class of data augmentations, and, based on the theory, recommend default decompositions that we demonstrate work well in practice.

Keywords: Approximate transition kernel; Convergence rate; Data augmentation; Limiting distribution; MCMC; Multivariate ordinal probit; Truncated multivariate normal

---

Chris Hans is Assistant Professor, Peter F. Craigmile is Associate Professor, Steven MacEachern is Professor and Xinyi Xu is Assistant Professor, Department of Statistics, and Greg M. Allenby is Helen C. Kurtz Chair in and Professor of Marketing, The Ohio State University, Columbus, OH 43210 (E-mail addresses: *hans@stat.osu.edu*, *pfc@stat.osu.edu*, *snm@stat.osu.edu*, *xinyi@stat.osu.edu*, and *allenby.1@osu.edu*). Ju Hee Lee is Post-doctoral Fellow, Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030 (E-mail address: *jlee14@mdanderson.org*).

1 Introduction

Statistical models involving discrete behaviors and responses often view the observed data as a censored outcome of a latent, continuous variable. Classical inference for models involving discreteness requires the calculation of integrals over the space of error terms associated with specific response options. Bayesian methods, in contrast, have relied on the method of data augmentation to simplify these calculations. Rather than dealing directly with the integrals associated with discreteness, data augmentation introduces a latent variable used as a conditioning argument in Markov chain Monte Carlo (MCMC) estimation. Early examples of applying data augmentation to choice data are the binary and multivariate probit models of Albert and Chib (1993) and Chib and Greenberg (1998), and the multinomial probit model of McCulloch and Rossi (1994) and McCulloch *et al.* (2000).

Methods of data augmentation and associated computation for Bayesian ordinal probit models have received extensive attention in the literature. Albert and Chib (1993) introduced the basic setup, with computational improvements offered by Cowles (1996), Nandram and Chen (1996) and Chen and Dey (2000), among others. The interplay of identification constraints and computation has been explored by Jeliazkov *et al.* (2008). The major computational challenge in MCMC for (multivariate) ordinal probit models relates to the "cutpoint" parameters that map the latent responses to the observed data. Efficient sampling of the cutpoints — the key to good MCMC for these models — features heavily in the literature. The gold standard approaches to sampling in these models integrate out portions of the latent responses when updating the cutpoints, allowing for more aggressive moves in the cutpoint space and resulting in Markov chains that converge faster. Use of these advanced algorithms is imperative in order to obtain accurate MCMC-based inference; this is especially true when working with high-dimensional models for large datasets.

A common formulation of the multivariate ordinal probit model uses a different, independent set of cutpoints for each element of the latent response. In Section 2.1 we describe data-analytic settings where a strongly dependent, perhaps even common set of cutpoints across elements of the latent response, is preferred. While attractive from a modeling perspective, collapsing to a common set of cutpoints has the unfortunate effect of rendering

the gold standard sampling approaches unusable under standard data augmentations, as the latent variables can no longer be accurately marginalized when sampling the cutpoints. Attempting to approximate the marginalization in MCMC results in sampling from the wrong transition kernel, which can severely alter the limiting distribution of the Markov chain.

In this paper we propose a new data augmentation for the broad class of multivariate ordinal probit models where standard sampling approaches cannot be applied. The data augmentation is based on a covariance decomposition and results in an MCMC algorithm which is easy to implement and has the correct limiting distribution. In Section 2 we introduce the particular version of the multivariate ordinal probit model we use to motivate and illustrate our new computational methodology. The particular model is designed for the analysis of scale-usage effects in ratings data (see also Rossi *et al.*, 2001; Javaras and Ripley, 2007). This model has wide application to the analysis of survey response data, where respondents are known to have a tendency to use only portions of the response scale (e.g., yea-sayers and nay-sayers). The formulation and motivation of the common cutpoint model is given in Section 2.1 along with a discussion of identification constraints.

Computational challenges under the common cutpoint model are described in Section 3. Our new covariance-based data augmentation is introduced in Section 3.1, and the new MCMC sampler is described in Section 3.2. We introduce theoretical results describing an "optimal" data augmentation/decomposition in Section 4. Our approach is similar in spirit to the conditional augmentation of van Dyk and Meng (2001), however we propose a different criterion for picking a good data augmentation. We provide a default decomposition method that is suggested by the theory, is easy to implement and that works well in practice. The decomposition-based approach to MCMC is illustrated in Section 5. In particular, we highlight the need for the new data augmentation by providing a comparison with an approach that approximates the marginalization of the latent variables when updating the cutpoints. We demonstrate that the posterior distribution under the approximation approach can be severely altered unless the approximations are extraordinarily accurate, which can correspond to compute times of several weeks instead of several hours.

2 A Scale-Usage Model for Survey Data

In this paper we develop computational methodology for models for data where N individuals provide responses to M questions, and the responses are given on a K-level ordinal scale. The observed data are represented by an $N \times M$ matrix $\mathbf{X}$, where $X_{ij} \in \{1, \ldots, K\}$ is the response provided by individual i to question j. The responses are modeled with a multivariate ordinal probit model by introducing an $N \times M$ matrix of latent variables $\mathbf{Y}$ with rows $\mathbf{Y}_i$ distributed conditionally independently according to

$$\mathbf{Y}_i \mid \boldsymbol{\mu}, \sigma_i^2, \tau_i, \boldsymbol{\Sigma} \overset{\text{ind}}{\sim} \mathrm{N}(\boldsymbol{\mu} + \tau_i \mathbf{1}, \sigma_i^2 \boldsymbol{\Sigma}), \quad i = 1, \ldots, N, \tag{1}$$

where $\mathbf{1}$ is a vector of ones. Given Y_{ij} and a vector of cutpoints $\mathbf{c}$, the observed data satisfy

$$X_{ij} \mid \mathbf{c}, Y_{ij} = \{k \ : \ c_{k-1} < Y_{ij} \leq c_k, \ k = 1, \ldots, K\}, \tag{2}$$

where $-\infty = c_0 < \cdots < c_K = \infty$.

The above model, introduced by Rossi *et al.* (2001), is designed to incorporate heterogeneity in scale-usage across individuals through inclusion of the individual-specific location and scale parameters τ_i and σ_i^2. Throughout this paper we illustrate the computational methodology we propose using the customer satisfaction survey dataset examined in Rossi *et al.* (2001) and Rossi *et al.* (2005), which is available in the `R` package `bayesm` (Rossi and McCulloch, 2008). The data consist of the responses of $N = 1,811$ customers who were asked $M = 10$ questions about a particular advertising product. The response scale contains $K = 10$ ordinal values. A complete description of the data, along with a justification of the scale-usage model in this setting, can be found in Rossi *et al.* (2001) and Rossi *et al.* (2005).

While we choose to illustrate our computational methodology in this interesting model setting, we note that our methods can be applied under a variety of related models: the mean of $\mathbf{Y}_i$ may be $\boldsymbol{\mu} + \sigma_i \tau_i \mathbf{1}$, and it may have additional structure reflecting other features of the study, e.g. fixed- and random-effect covariates or additional structure to capture longitudinal effects (Chib and Jeliazkov, 2006). Individual heterogeneity may exhibit substructure with demographic or other variables driving a latent class model for the (τ_i, σ_i^2). The covariance matrix $\boldsymbol{\Sigma}$ may be constrained to encode other structure in the model. We focus on the basic

model as it is of interest in marketing applications (Rossi *et al.*, 2001) with the understanding that our new computational methods can also be applied in more complicated model settings.

Our particular formulation of the scale-usage model assumes that the individual-specific effects arise from population distributions $\tau_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma_\tau^2)$ and $\sigma_i^2 \overset{\text{iid}}{\sim} \text{IG}(a/2, (a-2)/2)$. The prior means of the τ_i are taken to be zero so that μ_j is the baseline mean for the latent response to question j; the individual scale parameters σ_i^2 are specified with prior means of one so that $\boldsymbol{\Sigma}$ represents the baseline covariance structure for $\mathbf{Y}_i$. We model the shared parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ with prior distributions $\boldsymbol{\mu} \sim \text{N}(\mathbf{0}, \mathbf{V})$ and $\boldsymbol{\Sigma} \sim \text{IW}(\delta, \boldsymbol{\Sigma}_0)$, with the inverse Wishart distribution parameterized so that $\text{E}[\boldsymbol{\Sigma}] = (\delta - M - 1)^{-1}\boldsymbol{\Sigma}_0$. When lacking substantive prior information about the covariance structure, we take $\boldsymbol{\Sigma}_0 = (\delta - M - 1)\mathbf{I}$ so that $\text{E}[\boldsymbol{\Sigma}] = \mathbf{I}$.

2.1 Cutpoints and Identification Constraints

The models we consider have a single vector of cutpoints $\mathbf{c}$ that is shared across the J questions. This choice is motivated by the applications we consider. In other versions of the multivariate ordinal probit model (Chen and Dey, 2000; Jeliazkov *et al.*, 2008), different sets of cutpoints for each response are used, replacing (2) with

$$X_{ij} \mid \mathbf{c}'_j, Y_{ij} = \{k \ : \ c'_{j,k-1} < Y_{ij} \leq c'_{j,k}, \ k = 1, \ldots, K\}, \tag{3}$$

where $\mathbf{c}'_j$ is a vector of cutpoints corresponding to the jth response. The $\mathbf{c}'_j$ are typically considered to be *a priori* independent. Chen and Dey (2000) have exploited this *a priori* independence structure in developing efficient algorithms for updating the entire collection of $\mathbf{c}_j$ in an MCMC setting by carefully transforming the parameters. In the applications we consider, responses tend to have the same meaning across questions: in customer satisfaction surveys, a response of 8 out of 10 on one question should have the same meaning on the latent scale as a response of 8 out of 10 on another. The cutpoints, then, should be very similar, if not identical, across questions. This suggests a prior that features very strong dependence between the $\mathbf{c}'_j$. The limiting case where the dependence becomes arbitrarily strong results in a parsimonious model with a single vector of cutpoints $\mathbf{c}$ that is shared across questions. Unfortunately, while this formulation of the model is attractive from a data-

analytic perspective, the efficient computational methodology of Nandram and Chen (1996) or Chen and Dey (2000) can no longer be applied. The specific computational challenges associated with the common cutpoint model are discussed in Section 3.

In general, one must be careful when performing inference under the multivariate ordinal probit model described above due to lack of identifiability of the parameters. A common approach is to force identifiable inference by imposing constraints on the parameters. For example, fixing the boundary values $c_0 = -\infty$ and $c_K = \infty$, as well as one additional cutpoint (say $c_m = 0$ for some m), and then specifying an improper, uniform prior on the remaining cutpoints (subject to the constraint that they are nondecreasing) guarantees identifiability. This approach is described by Albert and Chib (1993), Cowles (1996), Johnson and Albert (1999) and Bradlow and Zaslavsky (1999). Another approach is to fix the two end cutpoints c_1 and c_{K-1} at particular values and then specify a proper uniform prior on the remaining free cutpoints subject to the ordering constraints (Webb and Forster, 2008). Jeliazkov *et al.* (2008) compares computation under both of these approaches. Rossi *et al.* (2001) constrain their single set of cutpoints to lie on a parabola, which, along with further identification restrictions, results in an even more parsimonious model requiring only a single free parameter. Kottas *et al.* (2005) avoid cutpoint-related issues entirely by using a nonparametric mixture of normal distributions for the latent variables, allowing the cutpoints to be fixed arbitrarily. We do not provide a more detailed discussion of identification constraints as this has been investigated extensively elsewhere (e.g. Jeliazkov *et al.*, 2008). We note that different identification constraints correspond to different Bayesian models and choose to focus on our particular application-motivated, albeit generalizable, model.

For identification purposes we pursue the approach of fixing the two end cutpoints, which requires no further restrictions on the covariance matrix $\mathbf{\Sigma}$, and note that the following prior under these constraints is attractive and flexible. Fix c_1 and c_{K-1} at points $-C$ and C, respectively, for some $C > 0$. This sets a baseline location and scale for the latent data. The prior for the remaining cutpoints is specified indirectly by modeling the distances between

adjacent cutpoints:

$$\begin{aligned} \zeta_k &\overset{\text{ind}}{\sim} \Gamma(\alpha_k, 1),\ k = 1, \ldots, K-2, \\ c_k|\boldsymbol{\zeta} &= -C + 2C \left(\sum_{l=1}^{k-1} \zeta_l \Big/ \sum_{l=1}^{K-2} \zeta_l \right),\ \ k = 2, \ldots, K-2. \end{aligned}$$

Under this prior the conditional distribution of c_k given its neighbors c_{k-1} and c_{k+1} is a scaled and shifted beta distribution with density function

$$p(c_k|c_{k-1}, c_{k+1}) \propto (c_k - c_{k-1})^{\alpha_k - 1}(c_{k+1} - c_k)^{\alpha_{k+1} - 1}, \quad c_{k-1} < c_k < c_{k+1}. \tag{4}$$

If we set $\alpha_k = \alpha$ for all k, then $\alpha = 1$ reduces to the usual (constrained) uniform prior. As $\alpha \to \infty$, the prior favors evenly spaced cutpoints; as $\alpha \to 0$, the prior favors unevenly spaced cutpoints. This model is easily extended to allow strong dependence in cutpoints across questions if (3) is used in place of (2).

3 Computational Challenges and Innovations

The main computational challenges to MCMC for this model are related to sampling the latent responses $\mathbf{Y}$ and the cutpoints $\mathbf{c}$. While there has been considerable research into improving computation for the multivariate ordinal probit model, the work has focused on models where different cutpoints are used for different questions, and where the cutpoints for different questions are conditionally independent. Under these models the latent data can be partially or entirely marginalized when sampling the cutpoints, resulting in faster-mixing Markov chains. For the model with a single set of cutpoints, this marginalization cannot be done accurately under the standard data augmentation, as described below. Under the standard data augmentation, the most naïve — and current standard — approach to MCMC updates each Y_{ij} and each c_k individually from their full conditional distributions. These two full conditionals have the restricted supports

$$c_{X_{ij}-1} < Y_{ij} < c_{X_{ij}} \quad \text{and} \quad \max_{(i,j)\,:\,X_{ij}=k} Y_{ij} < c_k < \min_{(i,j)\,:\,X_{ij}=k+1} Y_{ij}. \tag{5}$$

The full conditional for Y_{ij} — a truncated normal distribution — is essentially a regression on $\mathbf{Y}_{i,-j}$ (we use the notation $\mathbf{Y}_{i,-j}$ to represent all elements of the vector $\mathbf{Y}_i$ except Y_{ij}).

The full conditional for c_k is a renormalized version of (4) where the support is restricted as in (5). As noted by Cowles (1996), the width of the restricted support of c_k can be extremely small, especially when N or M is large, which causes the Markov chain to mix very slowly because each c_k can only move a small amount at each step in the sampler. Letting $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_4) \equiv (\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\sigma}, \boldsymbol{\Sigma})$, the standard sampler, which uses one-at-a-time updating, is as follows.

Algorithm 1 : Standard Sampler

Step 1: Update $\boldsymbol{\theta}_l$ *from* $p(\boldsymbol{\theta}_l \mid \boldsymbol{\theta}_{-l}, \mathbf{Y}, \mathbf{c}, \mathbf{X})$ *for* $l = 1, \ldots, 4$.

Step 2: Update c_k *from* $p(c_k \mid c_{-k}, \mathbf{Y}, \boldsymbol{\theta}, \mathbf{X})$ *for* $k = 2, \ldots, K-2$.

Step 3: Update Y_{ij} *from* $p(Y_{ij} \mid \mathbf{Y}_{-i,-j}, \boldsymbol{\theta}, \mathbf{c}, \mathbf{X})$ *for* $j = 1, \ldots, M$ *and* $i = 1, \ldots, N$.

Typical output from this sampler is displayed in Figure 1. The chains are for the seven free cutpoints for the model described in Section 2 when applied to the customer satisfaction survey data. The slowly mixing black chains are from the Standard Sampler (the gray chains are from the sampler we later propose in Section 3.2). The chains from the Standard Sampler have extremely high autocorrelation, which can be seen even more clearly in Figure 2. The thick, solid lines in this plot (labeled as $\rho = 0$) are the estimated autocorrelation functions for the cutpoints under the Standard Sampler. The acfs were estimated based on all but the first 50,000 samples from a run of million iterations. The rest of the figure is explained in Section 5.2. The obvious poor mixing and slow convergence makes inference based on output from the Standard Sampler practically impossible.

Two early approaches for improving MCMC convergence and mixing are to block variates or to marginalize them (Liu *et al.*, 1994; MacEachern, 1994). In this context, one might marginalize the latent variable $\mathbf{Y}$ when updating $\mathbf{c}$. This is known to work well for the univariate ordinal probit model (Cowles, 1996), and is the foundation of the approaches introduced by Nandram and Chen (1996) and Chen and Dey (2000). Such a sampler is structured as follows.

Algorithm 2 : Blocked Sampler

Step 1: Update $\boldsymbol{\theta}_l$ *from* $p(\boldsymbol{\theta}_l \mid \boldsymbol{\theta}_{-l}, \mathbf{Y}, \mathbf{c}, \mathbf{X})$ *for* $l = 1, \ldots, 4$.

Step 2: Update $\mathbf{c}$ *from* $p(\mathbf{c} \mid \boldsymbol{\theta}, \mathbf{X})$ *(perhaps via the Metropolis-Hastings algorithm).*
Step 3: Update $\mathbf{Y}$ *from* $p(\mathbf{Y} \mid \mathbf{c}, \boldsymbol{\theta}, \mathbf{X}) = \prod_{i=1}^{N} p(\mathbf{Y}_i \mid \mathbf{c}, \boldsymbol{\theta}, \mathbf{X})$.

This strategy of updating is also known as a partially collapsed Gibbs sampler (van Dyk and Park, 2008) where the update for $(\mathbf{c}, \mathbf{Y})$ is blocked in Steps 2 and 3. The acceptance probability in Step 2 requires evaluation of the integral

$$\int p(\mathbf{X} \mid \mathbf{Y}, \mathbf{c})\, p(\mathbf{Y} \mid \boldsymbol{\theta})\, d\mathbf{Y} = \prod_{i=1}^{N} \int_{R_i(\mathbf{c})} \mathrm{N}(\mathbf{Y}_i \mid \boldsymbol{\mu} + \tau_i \mathbf{1}\ ,\ \sigma_i^2 \boldsymbol{\Sigma})\, d\mathbf{Y}_i \tag{6}$$

under both the proposed and current values of the cutpoints, where $R_i(\mathbf{c})$ are the rectangular regions implied by the observed data $\mathbf{X}_i$ for a given set of cutpoints $\mathbf{c}$. Step 3 requires sampling $\mathbf{Y}_i$ independently from multivariate normal distributions restricted to the rectangular regions $R_i(\mathbf{c})$. Use of these two steps would remove the hard (conditional) restrictions on $\mathbf{c}$ in (5), allowing the cutpoints to move more freely at each step in the sampler.

There are two major challenges to face when implementing the Blocked Sampler in the common cutpoint model. First, sampling directly from the truncated normal distributions in Step 3 is nontrivial when M is even moderately large. Most commonly-used approaches rely on full-conditional (Gibbs) sampling (though see Damien and Walker, 2001; Liechty and Lu, 2010, for auxiliary variables approaches) where the components of $\mathbf{Y}_i$ are updated sequentially from the conditional distributions $Y_{ij} \mid \mathbf{Y}_{i,-j}, \mathbf{c}, \boldsymbol{\theta}, \mathbf{X}$ or where similar full conditional generations are made after a transformation. Unfortunately, such approaches cannot be used to perform Step 3 when $\mathbf{Y}$ has been marginalized for the generation of $\mathbf{c}$ in Step 2: because $\mathbf{c}$ was updated without conditioning on $\mathbf{Y}$, each $\mathbf{Y}_i$ must be generated *de novo* from its full conditional distribution $\mathbf{Y}_i \mid \mathbf{c}, \boldsymbol{\theta}, \mathbf{X}$ in order for the sampler to have the correct limiting distribution.

The second challenge to face is the evaluation of the integrals in (6). This is a difficult problem even in low-dimensional settings (see Hsu, 1992; Genz, 1992; Hajivassiliou *et al.*, 1996, for approximation methods). For survey data the number of questions determines the dimension of the integrals, and the number of respondents determines the number of such integrals that must be evaluated at each iterate of the sampler. The integrals must be evaluated under both the current and proposed sets of cutpoints, and the ratio of these two

integrals is a core component of the Metropolis–Hastings acceptance probability. For example, if the cutpoints are jointly updated via Metropolis–Hastings in the customer satisfaction dataset, 3,622 ten-dimensional integrals would have to be approximated at each iteration of the MCMC. The combination of a difficult integral with great replication leads to the possibility of substantial approximation error, both in terms of bias and variance. Currently, there is no theory that describes the impact of substituting the "wrong" acceptance probabilities on the limiting distribution of the Markov chain; our empirical investigation of such approximations for this model in Section 5.3 demonstrates that the limiting distribution can indeed be severely altered. With these difficulties in mind, we propose a new approach to computation for this model.

3.1 Data augmentation by covariance decomposition

Instead of attempting to obtain better approximations to $\mathbf{Y}_i \mid \mathbf{c}, \boldsymbol{\theta}, \mathbf{X}$ or to the integrals in (6), we devise a different sampler which allows us to bypass these challenges. Here we employ a decomposition of the covariance matrix $\boldsymbol{\Sigma}$ which facilitates a new sampler with the useful properties that (i) the limiting distribution of the Markov chain is indeed the posterior distribution of interest, (ii) the cutpoints can be updated without imposing the hard constraints in (5), and (iii) sampling from multivariate truncated normal distributions is not required.

Rather than working directly with the latent $\mathbf{Y}_i$ which have covariances $\sigma_i^2\boldsymbol{\Sigma}$, we create an additional orthogonalizing latent variable that will allow us to turn the multivariate integral into a product of univariate integrals. Specifically, we decompose the covariance matrix $\boldsymbol{\Sigma}$ into two parts,

$$\boldsymbol{\Sigma} \;=\; \mathbf{D} + \mathbf{R}, \tag{7}$$

where $\mathbf{D} = \text{diag}(d_1, \cdots, d_M)$ is a positive definite matrix and $\mathbf{R}$ is a non-negative definite matrix. This allows us to represent the likelihood as

$$\begin{aligned} \mathbf{Y}_i \mid \boldsymbol{\mu}, \tau_i, \sigma_i^2, \mathbf{Z}_i, \mathbf{D} &\sim \text{N}\big(\boldsymbol{\mu} + \tau_i\mathbf{1} + \mathbf{Z}_i, \sigma_i^2\mathbf{D}\big)\,, \\ \mathbf{Z}_i \mid \sigma_i^2, \boldsymbol{\Sigma}, \mathbf{D} &\sim \text{N}\big(\mathbf{0}, \sigma_i^2(\boldsymbol{\Sigma} - \mathbf{D})\big)\,, \end{aligned} \tag{8}$$

by introducing a new collection of latent variables $\mathbf{Z}_i$. The original data augmentation is recovered by integrating out the new latent variables $\mathbf{Z}_i$.

Before describing how we use the decomposition to facilitate computation, we note that this decomposition has been used for different purposes in related model settings. Stern (1992, 1997) uses this decomposition to construct an efficient Monte Carlo estimator of the likelihood probabilities (6) under a related model. Chib and Jeliazkov (2006) use this decomposition and data augmentation to facilitate a matrix inversion in an MCMC algorithm for a highly-structured multivariate probit model. Their use of the decomposition speeds computation but does not affect properties of the MCMC algorithm. In contrast, our use of the decomposition, as described below, is unique in that we explicitly use the orthogonalizing latent variables $\mathbf{Z}_i$ to allow a fraction of the data augmentation to be integrated out when moving the cutpoints. The specific decomposition impacts the mixing of the Markov chain, as discussed in Section 4 and illustrated in Section 5.

The key to moving the cutpoints in the Blocked Sampler (Algorithm 2) is being able to evaluate integrals like (6), which under our new data augmentation becomes

$$\int p(\mathbf{X} \mid \mathbf{Y}, \mathbf{c})\, p(\mathbf{Y} \mid \mathbf{Z}, \boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\mu}, \mathbf{D})\, d\mathbf{Y}. \tag{9}$$

The Y_{ij} are conditionally independent given the Z_{ij} and so we can express this probability by

$$\begin{aligned} &\prod_{i=1}^{N}\prod_{j=1}^{M} \int_{c_{x_{ij}-1}}^{c_{x_{ij}}} (\sigma_i\sqrt{d_j})^{-1}\phi\left(\frac{y_{ij}-\mu_j-\tau_i-z_{ij}}{\sigma_i\sqrt{d_j}}\right) dy_{ij} \\ = &\prod_{i=1}^{N}\prod_{j=1}^{M} \left[\Phi\left(\frac{c_{x_{ij}}-\mu_j-\tau_i-z_{ij}}{\sigma_i\sqrt{d_j}}\right) - \Phi\left(\frac{c_{x_{ij}-1}-\mu_j-\tau_i-z_{ij}}{\sigma_i\sqrt{d_j}}\right)\right], \end{aligned} \tag{10}$$

where ϕ and Φ are, respectively, the pdf and cdf of the standard normal distribution. Evaluation of the above one-dimensional normal integrals (or their logarithms to improve accuracy over sometimes very large products) is quick and accurate. This leads to an accurately-computed acceptance probability for the Metropolis–Hastings step in which we move the cutpoints, solving the approximation problem associated with Step 2 of the Blocked Sampler.

The difficulty associated with Step 3 of the Blocked Sampler is sampling from the conditional distributions $\mathbf{Y}_i \mid \mathbf{c}, \boldsymbol{\theta}, \mathbf{X}$. Under our decomposition we can focus on the conditional

distributions $\mathbf{Y}_i \mid \mathbf{Z}_i, \mathbf{c}, \boldsymbol{\theta}, \mathbf{X}$, which decompose into independent truncated normal distributions $Y_{ij} \mid Z_{ij}, \mathbf{c}, \boldsymbol{\theta}, X_{ij}$ that are easily sampled (details in Section 3.2). Our decomposition also requires that $\mathbf{Z}$ be updated in the new Gibbs sampler we use to fit the scale-usage model (or any variant on it); we show in Section 3.2 that $\mathbf{Z}$ can be easily updated using draws from multivariate normal distributions.

Different decompositions of $\boldsymbol{\Sigma}$ into $\mathbf{R}$ and $\mathbf{D}$ will lead to different Markov chains, and these chains will have different convergence rates. The space of all allowable decompositions can be thought of as a class of data augmentations. Any particular decomposition then corresponds to a conditional augmentation in the sense of van Dyk and Meng (2001), who provide a framework for choosing good augmentations within a class. In Section 4 we propose a new approach for choosing good conditional augmentations and derive a theoretical description of an "optimal" decomposition for our model. For the moment we assume we have a method of decomposing $\boldsymbol{\Sigma}$.

3.2 Covariance Decomposition MCMC

We use our new data augmentation to construct the following improved MCMC algorithm. Each step can be implemented quickly and accurately.

Algorithm 3 : Decomposition Sampler

Step 1: Update $\boldsymbol{\theta}_l$ *from* $\boldsymbol{\theta}_l \mid \boldsymbol{\theta}_{-l}, \mathbf{Y}, \mathbf{c}, \mathbf{X}$ *for* $l = 1, \ldots, 4$.

Step 2: Decompose $\boldsymbol{\Sigma}$ *into* $\mathbf{R}$ *and* $\mathbf{D}$*, and update* $\mathbf{Z}_i$ *independently from* $\mathbf{Z}_i \mid \mathbf{Y}_i, \boldsymbol{\theta}, \mathbf{c}, \mathbf{X}_i$ *for* $i = 1, \ldots, N$.

Step 3: Update c_k *given* c_{-k} *via Metropolis–Hastings with* $p(c_k \mid c_{-k}, \mathbf{Z}, \boldsymbol{\theta}, \mathbf{X})$ *as the target density functions for* $k = 2, \ldots, K-2$.

Step 4: Update Y_{ij} *independently from* $Y_{ij} \mid Z_{ij}, \mathbf{c}, \boldsymbol{\theta}, X_{ij}$ *for* $i = 1, \ldots, N$ *and* $j = 1, \ldots, M$.

Steps 1 and 2 are blocked for $\boldsymbol{\theta}$ and $\mathbf{Z}$, as are Steps 3 and 4 for $\mathbf{c}$ and $\mathbf{Y}$. Blocking typically results in a better convergence rate of the Gibbs sampler when compared to standard one-at-a-time updating (MacEachern and Müller, 1998; van Dyk and Park, 2008).

The conditional distributions in Step 1 for the elements of $\boldsymbol{\theta}$ (i.e. $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and the individual-specific parameters τ_i and σ_i^2) have standard forms and are described in Appendix B. Meth-

ods for decomposing $\mathbf{\Sigma}$ in Step 2 are discussed in Section 4.2. Given a decomposition of $\mathbf{\Sigma}$ into $\mathbf{R}$ and $\mathbf{D}$, the full conditional distribution of $\mathbf{Z}$ factors into the independent conditional distributions $\mathbf{Z}_i \mid \mathbf{Y}_i, \boldsymbol{\theta}, \mathbf{c}, \mathbf{X}_i$ in Step 2 which have the form

$$\mathrm{N}\big((\mathbf{R}^{-1}+\mathbf{D}^{-1})^{-1}\mathbf{D}^{-1}(\mathbf{Y}_i-\boldsymbol{\mu}-\tau_i\mathbf{1}), \sigma_i^2(\mathbf{R}^{-1}+\mathbf{D}^{-1})^{-1}\big)\,, \tag{11}$$

and which do not depend on the cutpoints $\mathbf{c}$. Generating the $\mathbf{Z}_i$ is straightforward if we decompose $\mathbf{\Sigma}$ so that $\mathbf{R} = \mathbf{\Sigma} - \mathbf{D}$ is positive definite. In our definition of the decomposition we allowed for decompositions that result in singular $\mathbf{R}$ matrices. Such singular decompositions are important, as we will show in Sections 4.1 and 5.2 that good decompositions correspond to singular $\mathbf{R}$.

If we decompose $\mathbf{\Sigma}$ so that $\mathbf{R}$ is singular, the conditional distributions for $\mathbf{Z}_i$ are of reduced rank and care must be taken when sampling. In the singular case where $\mathrm{rank}(\mathbf{R}) = M-1$, orthogonally diagonalize the matrix $\mathbf{D}^{-1/2}\mathbf{R}\mathbf{D}^{-1/2}$ as $\mathbf{H}\mathbf{\Lambda}\mathbf{H}'$, where $\lambda_M = 0$ due to the singularity of $\mathbf{R}$. To sample $\mathbf{Z}_i$, first compute $\mathbf{w}_i = \mathbf{D}^{-1/2}(\mathbf{Y}_i - \boldsymbol{\mu} - \tau_i\mathbf{1})/\sigma_i$ and then sample

$$u_{ij} \sim \mathrm{N}\left(\frac{\sum_{l=1}^{M} H_{lj}w_{il}}{1+\lambda_j^{-1}}\,,\ \frac{1}{1+\lambda_j^{-1}}\right)$$

independently for $j = 1, \ldots, M-1$, setting $u_{iM} = 0$. A sample from the reduced-rank normal distribution is $\mathbf{Z}_i = \sigma_i\mathbf{D}^{1/2}\mathbf{H}\mathbf{u}_i$.

Because we condition on $\mathbf{Z}$, the cutpoints can be updated in Step 3 without conditioning on $\mathbf{Y}$. We cycle through the conditional distributions $p(c_k|\mathbf{c}_{-k}, \mathbf{Z}, \mathbf{X}, \boldsymbol{\theta}) \propto p(c_k|c_{k-1}, c_{k+1})w_k(c_k)$, where

$$w_k(c_k) = \prod_{l=k}^{k+1}\prod_{(i,j)\in A_l}\left\{\Phi\left(\frac{c_l-\mu_j-\tau_i-z_{ij}}{\sigma_i\sqrt{D_{jj}}}\right) - \Phi\left(\frac{c_{l-1}-\mu_j-\tau_i-z_{ij}}{\sigma_i\sqrt{D_{jj}}}\right)\right\},$$

which comes from integral (9) with $A_l = \{(i,j)\ :\ X_{ij} = l\}$. The one-dimensional integrals $\Phi(\cdot)$ can be evaluated numerically with high accuracy. Metropolis-Hastings steps are used to update each c_k: the probability of accepting a proposed value c_k^* generated from a proposal distribution with density function g is

$$\alpha_k = \min\left\{1, \left(\frac{c_k^*-c_{k-1}}{c_k-c_{k-1}}\right)^{\alpha_k-1}\left(\frac{c_{k+1}-c_k^*}{c_{k+1}-c_k}\right)^{\alpha_{k+1}-1}\frac{w(c_k^*)g(c_k)}{w(c_k)g(c_k^*)}\right\}.$$

We take the proposal distribution to be a truncated normal distribution "centered" at the current value of c_k: $g(c_k^*|c_k) \propto s^{-1}\phi((c_k^* - c_k)/s)I(c_{k-1} < c_k^* \leq c_{k+1})$, where $I(\cdot)$ is an indicator function.

The cutpoints are updated one-at-a-time in Step 3 of the Decomposition Sampler described above. This choice was made in part due to ease of computation. A more attractive sampler would perform a joint update of the cutpoints, as this is known to help improve mixing in related models. While we do not explore such an approach here, we note that a joint update is feasible now that we are able to condition on the orthogonalizing latent variables $\mathbf{Z}$. For example, a modified version of the Metropolis-Hastings approach described by Chen and Dey (2000) would provide an automatic method for generating a joint update. It is important to note, however, that under the single cutpoint model this approach could not be accurately applied without conditioning on $\mathbf{Z}$.

Finally, Step 4 requires sampling from $p(\mathbf{Y} \mid \mathbf{Z}, \mathbf{c}, \boldsymbol{\theta}, \mathbf{X}) = \prod_{i=1}^{N} \prod_{j=1}^{M} p(Y_{ij} \mid Z_{ij}, \mathbf{c}, \boldsymbol{\theta}, X_{ij})$, where the full conditional distribution of Y_{ij} is $\mathrm{N}(\mu_j + \tau_i + z_{ij} \ , \ \sigma_i^2 D_{jj})$ restricted to the interval $(c_{x_{ij}-1}, c_{x_{ij}}]$. A draw from this univariate truncated normal distribution can be obtained efficiently using the method of Geweke (1991).

4 Approaches to Covariance Decomposition

Different covariance decompositions correspond to different data augmentations, which lead to Markov chains with different convergence rates. van Dyk and Meng (2001) study the problem of selecting a good data augmentation when the class of augmentations is indexed by a working parameter. They propose an optimality criterion for choosing a particular value of the working parameter — a conditional augmentation — that is motivated by convergence properties of related EM algorithms.

Here, we propose a different approach to choosing a conditional augmentation. To study the problem of choosing a decomposition theoretically, we pass from the full hierarchical model with unknown τ_i and σ_i^2 and with constraints and cutpoints to a simpler proxy model which focuses only on unconstrained $\mathbf{Y}$ and $\mathbf{Z}$. Our approach is to optimize the convergence dynamics of the two-step Gibbs sampler that can be used to fit this proxy model (8), which

has steps $[\mathbf{Z}|\mathbf{Y}]$ and $[\mathbf{Y}|\mathbf{Z}]$. Focusing on the transition from the $(k-1)^{st}$ iterate to the k^{th} iterate, the conditional distribution for the i^{th} subject, $\mathbf{Y}_i^{(k)} \mid \mathbf{Y}_i^{(k-1)}$, is

$$\mathrm{N}\Big((\mathbf{R}^{-1}+\mathbf{D}^{-1})^{-1}\mathbf{D}^{-1}(\mathbf{Y}_i^{(k-1)}-\boldsymbol{\mu}-\tau_i\mathbf{1})+\boldsymbol{\mu}+\tau_i\mathbf{1}, \sigma_i^2\mathbf{D}+\sigma_i^2(\mathbf{R}^{-1}+\mathbf{D}^{-1})^{-1}\Big). \qquad (12)$$

We recognize this as a vector autoregressive model for which Roberts and Sahu (1997) have developed convergence results. Their Theorem 1 states that the convergence rate is geometric, with the largest eigenvalue of the propagator matrix determining the rate. Thus to maximize the convergence rate of the Markov chain (12) for $\mathbf{Y}_i$, we wish to find a diagonal, positive definite matrix $\mathbf{D}$ such that the largest eigenvalue of $(\mathbf{R}^{-1}+\mathbf{D}^{-1})^{-1}\mathbf{D}^{-1}$ is minimal.

4.1 Eigenvalue theory for the optimal covariance decomposition

We first provide some general conditions for the optimal $\mathbf{D}$ in the matrix decomposition (7). "Optimal" is used here with respect to the simplified proxy model (12). In particular, we show that (i) the optimal decomposition must be singular (i.e., $\mathbf{R}$ must be a singular matrix) and that (ii) in attempting to find the optimal decomposition we can equivalently work with $\boldsymbol{\Sigma}$ in correlation matrix form. Throughout this section we denote the eigenvalues of an $M \times M$ matrix $\mathbf{P}$ by $\boldsymbol{\lambda}(\mathbf{P}) = (\lambda_1(\mathbf{P}), \cdots, \lambda_M(\mathbf{P}))$ with $\lambda_1(\mathbf{P}) \geq \cdots \geq \lambda_M(\mathbf{P})$. We begin by defining the decomposition space.

Definition 1 *The decomposition space of the $M \times M$ positive definite covariance matrix $\boldsymbol{\Sigma}$ is $DS(\boldsymbol{\Sigma}) = \{\mathbf{A} \mid \mathbf{A} = diag(a_1, \cdots, a_M) > 0$ and $\mathbf{R} = \boldsymbol{\Sigma} - \mathbf{A} \geq 0\}$.*

In this definition, $>$ denotes positive definiteness of the matrix $\mathbf{A}$ and $\geq$ denotes non-negative definiteness of the matrix $\mathbf{R}$.

The first proposition characterizes the optimal decomposition. As later examples show, this optimal decomposition need not be unique. Proofs of the results appear in Appendix A.

Proposition 1 *To achieve the maximum possible convergence rate of the Markov chain (12) for $\mathbf{Y}_i$, the optimal $\mathbf{D}$ in the matrix decomposition (7) must lie in the decomposition space of $\boldsymbol{\Sigma}$, and*

$$\mathbf{D} \;=\; \arg \min_{\mathbf{A} \in DS(\boldsymbol{\Sigma})} \lambda_1(\mathbf{A}^{-1/2}\boldsymbol{\Sigma}\mathbf{A}^{-1/2}). \qquad (13)$$

Proposition 1 shows that the problem of finding the optimal matrix decomposition is equivalent to the problem of finding an optimal $\mathbf{D} \in \mathrm{DS}(\mathbf{\Sigma})$ that minimizes the largest eigenvalue of $\mathbf{D}^{-1/2}\mathbf{\Sigma}\mathbf{D}^{-1/2}$. The next result shows that, under the optimal decomposition, $\mathbf{R} = \mathbf{\Sigma} - \mathbf{D}$ must be singular.

Corollary 1 *Let* $\mathbf{T}$ *be a diagonal and positive definite matrix. Consider the class of decompositions generated by* $t\mathbf{T}$*, where* $t\mathbf{T} \in DS(\mathbf{\Sigma})$*. The resulting* t *lies in some interval* $(0, t_{\max}]$*. Within this class,* $\mathbf{D} = t_{\max}\mathbf{T}$ *provides the best decomposition of* $\mathbf{\Sigma}$*.*

To search for the optimal decomposition, the next proposition allows us to focus on the correlation matrix rather than the covariance matrix. The result follows from the observation that the Markov chains based on the covariance matrix $\mathbf{\Sigma}$ and the corresponding correlation matrix, $\mathbf{C}$, can be perfectly coupled.

Proposition 2 *Let* $\mathbf{V} = diag(\mathbf{\Sigma})$ *denote a matrix of the diagonal elements of* $\mathbf{\Sigma}$*. The optimal convergence rate for the Markov chain based on* $\mathbf{Y}$ *is equal to that for the Markov chain based on* $\mathbf{V}^{-1/2}\mathbf{Y}$*, which has covariance matrix* $\mathbf{C} = \mathbf{V}^{-1/2}\mathbf{\Sigma}\mathbf{V}^{-1/2}$*. Furthermore, if* $D_{\mathbf{\Sigma}} = \arg \min_{\mathbf{A} \in DS(\mathbf{\Sigma})} \lambda_1 \left(\mathbf{A}^{-1/2}\mathbf{\Sigma}\mathbf{A}^{-1/2}\right)$*, then*

$$\mathbf{D}_{\mathbf{C}} = \arg \min_{\mathbf{A} \in DS(\mathbf{C})} \lambda_1 \left(\mathbf{A}^{-1/2}\mathbf{C}\mathbf{A}^{-1/2}\right) = \mathbf{D}_{\mathbf{\Sigma}}\mathbf{V}^{-1}.$$

In other words, if we can find the optimal decomposition for the chain with covariance matrix $\mathbf{C}$, we can then transform to find the optimal decomposition for the chain with covariance matrix $\mathbf{\Sigma}$. We use this approach in our default method of decomposition described below.

4.2 Default Decomposition

An explicit form for the optimal $\mathbf{D}$ can be found in special cases where $\mathbf{\Sigma}$ or $\mathbf{C}$ have particular structure. Such cases include independence structure, certain exchangeable correlation structures (including circular and reversible correlation structures), and block diagonal structures where each block itself has a special structure. We derive the optimal $\mathbf{D}$ for these cases in Appendix C (in the online supplement to the paper). For a general covariance matrix,

it is difficult to obtain a solution to the eigenvalue problem, and so if we want to find the optimal decomposition we must resort to numerical searches.

We have investigated several guided search strategies for the proxy model (12) that perturb elements of $\mathbf{D}$ deterministically or stochastically, with the search terminating according to a specified convergence criterion (e.g., the decrease in largest eigenvalue of $\mathbf{D}^{-1/2}\boldsymbol{\Sigma}\mathbf{D}^{-1/2}$ at a particular iteration of the search is less than a certain tolerance value). While these searches appear to work reasonably well, implementing a search at each iteration of the MCMC algorithm for the real model is computationally expensive, and "optimal" decompositions may not necessarily provide noticeable improvements over decompositions that are "nearly optimal." In practice it is more important to avoid decompositions that result in obviously and demonstrably poor data augmentations.

With this in mind, we propose a default approach to covariance decomposition that is motivated by the theory and that works well in practice. After $\boldsymbol{\Sigma}$ is sampled at each iterate, we employ the decomposition defined by $\mathbf{D} = \rho\lambda_M(\mathbf{C})\mathbf{V}$, where $\lambda_M(\mathbf{C})$ is the smallest eigenvalue of the sampled covariance matrix $\mathbf{C} = \mathbf{V}^{-1/2}\boldsymbol{\Sigma}\mathbf{V}^{-1/2}$ and $\rho \in [0, 1]$. Taking $\rho = 1$ results in singular $\mathbf{R}$. Values of ρ in $(0, 1)$ result in nonsingular $\mathbf{R}$, while $\rho = 0$ corresponds to the case were $\mathbf{Z}$ is removed from the model. We prefer working with the correlation matrix, which places all variates on a common scale, as it tends to stabilize the decomposition by not allowing any one direction to dominate. We explore the class of decompositions $\mathbf{D} = \rho\lambda_M(\mathbf{C})\mathbf{V}$ indexed by ρ in Section 5.2. Our results indicate that the choice of ρ substantially impacts the convergence properties of the chain, with chains based on singular decompositions ($\rho = 1$) performing best, in accordance with the theory developed above. While this default approach to decomposition may not be "optimal", we find that for the datasets we have looked at this approach produces Markov chains with good behavior.

5 Illustrations and Comparisons

Here we use the customer satisfaction dataset to illustrate some properties of our new Decomposition Sampler and provide comparisons with other existing approaches to sampling. All of our examples were run on a Mac Pro with eight cores (two 2.66 GHz Quad-Core Intel

Xeon 3500 series processors) with 8 GB of RAM.

5.1 Comparing the Standard and Decomposition Samplers

We first compare our new Decomposition Sampler (Algorithm 3) with the Standard Sampler (Algorithm 1) in order to demonstrate improvements in convergence and mixing of the Markov chain. Both samplers were run for 100,000 iterations using the same starting values and hyperparameter settings of $\mathbf{V} = 16 \times I_{10}$, $\delta = 15$, $\mathbf{\Sigma}_0 = (\delta - M - 1) \times I_{10} = 4 \times I_{10}$, $\sigma_\tau^2 = 16$, $a = 5$, $\alpha_k = 1$ and $C = 10$ (details in Section 2). For the decomposition sampler, the covariance matrix $\mathbf{\Sigma}$ was decomposed at each iteration as described in Section 4.2 with $\mathbf{D} = \lambda_M(\mathbf{C})\mathbf{V}$, corresponding to $\rho = 1$ and yielding a singular $\mathbf{R}$ matrix. Also for the decomposition sampler, the "standard deviation" of the truncated normal proposal distribution for the cut points was taken to be $s = 0.1$, which was appropriate for this dataset.

Traceplots for the seven free cutpoint parameters $c_2, \ldots, c_8$ for both samplers are shown in Figure 1. The black sample paths are from the Standard Sampler and the gray sample paths are from the Decomposition Sampler. The new Decomposition Sampler has a much shorter burn-in period, mixes better and converges faster than does the standard sampler. The improvement in mixing is confirmed by Figure 2, which compares the autocorrelation functions for the cutpoints under the Standard Sampler ($\rho = 0$) and the Decomposition Sampler ($\rho = 1$). The plots are based on all but the first 50,000 samples from a run of one million iterations. While we have focused here on output for the cutpoints — the parameters that are known to exhibit poor mixing in these models — we note that the chains for the other parameters of interest, $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$, mix well under our sampler.

5.2 Impact of the choice of decomposition

We described our default approach to covariance decomposition in Section 4.2: set $\mathbf{D} = \rho\lambda_M(\mathbf{C})\mathbf{V}$, with $\rho = 1$ (a singular decomposition) recommended. This recommendation was based on the theoretical results developed in Section 4.1 for the proxy model (12). Within the class of decompositions indexed by ρ, we expect decompositions with ρ close to zero to perform poorly. In this case, the conditional variance of Y_{ij}, $\sigma_i^2 D_{jj}$, will also be close to zero

and so the matrix of latent variables $\mathbf{Y}$ will be very highly pinned down at each step in the MCMC algorithm, resulting in high autocorrelation. When $\rho = 1$ (or nearly so), the matrix $\mathbf{R}$ will be singular (or nearly so), and we expect the Markov chain to mix well based on the theoretical results.

To demonstrate this empirically with the scale-usage model, we ran the Decomposition Sampler five times with values of $\rho \in \{0.05, 0.10, 0.25, 0.5, 1\}$. The hyperparameters were set as above, and each of the five chains was run for 1,000,000 iterations, with the first 50,000 iterations discarded as burn in. Autocorrelation functions for the cutpoints $c_2, \ldots, c_8$ for the five different values of ρ are shown in Figure 2. As expected, the autocorrelation function decays most rapidly for values of ρ near one, with the chain with $\rho = 0.05$ exhibiting the slowest decay. Because singular decompositions provide no particular computational challenges, we recommend their use as a default.

We also note that in some cases, regardless of the choice of decomposition, our data augmentation sampler may perform poorly. If $\mathbf{\Sigma}$ is nearly singular, the decomposition will necessarily have some D_{jj} close to zero even if we push all the way to a singular $\mathbf{R}$. The resulting Markov chain will behave similarly to the $\rho \approx 0$ case described above: the conditional updates $[\mathbf{Y} \mid \mathbf{Z}]$ and $[\mathbf{Z} \mid \mathbf{Y}]$ will not allow the chain to mix freely and the data augmentation sampler will not perform much better than the standard sampler.

5.3 Comparison to approximation methods

Here we document the impact that approximating the multivariate integrals in (6) can have on the limiting distribution of the Markov chain, highlighting the importance of avoiding such integrals in our Decomposition Sampler. In comparing our Decomposition Sampler with any approximation approach, it is essential to use the identical model (including identical prior distributions), to use as much of the same code as possible, and to run the code on the same platform. This reduces the likelihood of coding errors and stabilizes trace comparisons. It also controls for a variety of (presumably) minor numerical issues such as machine roundoff and variations in methods used to generate particular variates. It is also necessary to have a low-dimensional summary of the posterior distribution so that any effect of the choice of algorithm can be easily visualized.

To make the comparisons, we focus on the scale-usage model of Rossi *et al.* (2001, 2005) which has an excellent implementation in the `R` package `bayesm` as the function `rscaleUsage`. This model is very similar to the model we described in Section 2; the most important difference is the treatment of the cutpoints. Rossi *et al.* (2001) constrain the cutpoints to lie on a parabola and then apply further restrictions to ensure identification. As a result, a single free parameter e determines the entire set of cutpoints. As the cutpoints are closely tied to the approximation issues, we focus on the single parameter e to make our comparisons. The Gibbs sampler implemented in the `rscaleUsage` function has the form of the Blocked Sampler (Algorithm 2), where $\mathbf{c}$ is replaced with the single parameter e. The integrals (6) required in Step 2 of the Gibbs sampler are estimated using the GHK importance sampling method (Keane, 1994; Hajivassiliou *et al.*, 1996), which requires specification of the number of replications, n_{ghk}, to be used for importance sampling. The approximation becomes better as n_{ghk} increases, however the integrals must be done for each of the $N = 1,811$ respondents at each iteration of the sampler and so there is a practical trade-off between accuracy and computation time.

We first fit the model using the default settings in the `rscaleUsage` function (except that we fixed the matrix hyperparameter $\mathbf{\Lambda}$ to have diagonal elements 10 and 1, with 0 in the off-diagonal elements). The default setting for the GHK method is $n_{\text{ghk}} = 100$. We ran four independent chains each for 26,000 iterations and discarded the first 1,000 iterations of each chain as burn in, providing a total of 100,000 samples. Each chain took approximately 12 hours to run. The estimated marginal posterior distribution of e is given by the blue dots in Figure 3. The `rscaleUsage` function discretizes the support of e in the implementation of the MCMC algorithm and so the posterior distribution in Figure 3 is estimated on a grid; the lines are added for visual reference. The vertical lines through the points in the figure represent approximate 95% confidence intervals that were constructed using the batch means method based on $m = 50$ batches of length $k = 2{,}000$ iterations; the intervals were robust to reasonable changes in m and k.

To assess the impact of the GHK approximation on the limiting distribution we re-ran the sampler two more times with $n_{\text{ghk}} = 1,000$ (five independent chains yielding 100,000 samples) and $n_{\text{ghk}} = 10{,}000$ (four independent chains yielding 50,000 samples). Each of the

five chains with $n_{\text{ghk}} = 1{,}000$ took approximately 68 hours to run, and each of the four chains with $n_{\text{ghk}} = 10{,}000$ took approximately 308 hours — almost 13 days — to run. The estimated posterior distributions of e for these two runs are given by the green and red dots in Figure 3. It is clear that the different values of n_{ghk} result in different estimates of the posterior distribution. Smaller values of n_{ghk} correspond to more approximation error and result in a "flattening" of the estimated posterior distribution. Larger values of n_{ghk} correspond to less approximation error and should result in more accurate estimates.

To assess the accuracy of the estimate of the posterior distribution when $n_{\text{ghk}} = 10{,}000$, we created a modified version of the `rscaleUsage` sampler that uses our decomposition of $\mathbf{\Sigma}$. Our decomposition approach avoids approximating the integrals in (6), and so comparing the posterior distribution estimated under the $n_{\text{ghk}} = 10{,}000$ sampler to the posterior distribution estimated under the decomposition sampler should indicate whether further increases in n_{ghk} would result in further changes to the estimated posterior distribution. To implement our sampler, we modified the `rscaleUsage` function in three specific ways. First, we added a step to decompose $\mathbf{\Sigma}$ into $\mathbf{R}$ and $\mathbf{D}$ and sample $\mathbf{Z}$ from its full conditional distribution, as described in Section 3.2. Second, we modified the update of e in the Gibbs sampler by conditioning on $\mathbf{Z}$ and replacing the approximation of the integrals in (6) with evaluation of the integrals in (10). Third, we modified the step where $\mathbf{Y}$ is updated by sampling from the full conditional distributions $Y_{ij} \mid Z_{ij}, \mathbf{c}, \mu_j, \sigma_i^2, D_{jj}, \tau_i, X_{ij}$ as described in Section 3.2, where $\mathbf{c}$ is the set of cutpoints implied by the current value of e.

We ran four independent chains of our modified sampler for 26,000 iterations each, discarding the first 1,000 iterations of each chain as burn in, providing 100,000 samples from the posterior distribution. On average, the run time for a single chain was slightly less than 12 hours, which is comparable to the sampler that used $n_{\text{ghk}} = 100$. The estimated posterior distribution of e is given by the black points in Figure 3. We expect this estimated posterior distribution to be close to the true posterior distribution because the modified sampler does not include any approximate likelihood evaluations. Indeed, we see that moving toward a more exact evaluation of the likelihood by increasing n_{ghk} in the `rscaleUsage` sampler from 100 to 1,000 to 10,000 moves the estimated posterior distribution closer to the posterior estimated by our decomposition sampler.

6 Discussion

We have introduced a new data-augmentation scheme that facilitates MCMC for a widely-used class of multivariate ordinal probit models, and, in particular, Bayesian models incorporating scale-usage heterogeneity. Current MCMC methods for fitting such models either converge too slowly or rely on approximations that distort the limiting distribution of the Markov chain. Our new data augmentation, which is based on a covariance decomposition, preserves the correct limiting distribution and facilitates Markov chain mixing. An examination of a method that relies on approximations showed that the approximations have a surprisingly large effect on the limiting distribution of the Markov chain. In the illustration in Section 5, the approximations resulted in a severe "flattening" of the posterior distribution of the parameter e, the parameter determining the location and spacing of the cutpoints. In general, we expect such flattening to occur when approximations to the Metropolis–Hastings acceptance probability are used, as stochastic simulation experiments have reproduced this behavior in simple cases. We suspect that the large effect observed in our example is due to accumulation of error over the large number of individual approximations that make up the likelihood calculation required in the Metropolis–Hastings acceptance probability. As dataset sizes continue to increase in applications within and beyond the field of marketing, understanding the impact of such approximations will become increasingly important, underscoring the value of Markov chains that are carefully designed to have the desired limiting distribution.

A Proofs of the results

Proof of Proposition 1 In the matrix decomposition (7), we require that matrix $\mathbf{D}$ be diagonal and positive definite and that the matrix $\mathbf{R}$ be non-negative definite. Throughout, where $\mathbf{R}$ is singular, we use the convention of replacing the expression for a singular $\mathbf{R}$ with its limit under a sequence of decompositions. Moreover, as shown in Roberts and Sahu (1997), the convergence rate of the Markov Chain (12) for $\mathbf{Y}_i$ is given by the largest eigenvalue of the matrix $(\mathbf{R}^{-1} + \mathbf{D}^{-1})^{-1}\mathbf{D}^{-1}$. Recalling that $\mathbf{R} = \mathbf{\Sigma} - \mathbf{D}$, the optimal $\mathbf{D}$

matrix satisfies

$$\mathbf{D} \;=\; \arg \min_{\mathbf{A}\in \mathrm{DS}(\boldsymbol{\Sigma})} \lambda_1 \left((\mathbf{R}^{-1} + \mathbf{A}^{-1})^{-1} \mathbf{A}^{-1} \right) \;=\; \arg \min_{\mathbf{A}\in \mathrm{DS}(\boldsymbol{\Sigma})} \lambda_1 \left(\left(\mathbf{A}\mathbf{R}^{-1} + \mathbf{I} \right)^{-1} \right).$$

Since for any invertible matrix $\mathbf{P}$, eigenvalues of $\mathbf{P}^{-1}$ are reciprocals of those of $\mathbf{P}$, and eigenvalues of $\mathbf{P} \pm \mathbf{I}$ are those of $\mathbf{P}$ increased/decreased by 1, we have

$$\begin{aligned} \mathbf{D} \;=\; \arg \max_{\mathbf{A}\in \mathrm{DS}(\boldsymbol{\Sigma})} \lambda_M \left(\mathbf{A}\mathbf{R}^{-1} \right) \;&=\; \arg \min_{\mathbf{A}\in \mathrm{DS}(\boldsymbol{\Sigma})} \lambda_1 \left(\mathbf{R}\mathbf{A}^{-1} \right) \\ &=\; \arg \min_{\mathbf{A}\in \mathrm{DS}(\boldsymbol{\Sigma})} \lambda_1 \left(\boldsymbol{\Sigma}\mathbf{A}^{-1} - I \right) \\ &=\; \arg \min_{\mathbf{A}\in \mathrm{DS}(\boldsymbol{\Sigma})} \lambda_1 \left(\boldsymbol{\Sigma}\mathbf{A}^{-1} \right). \end{aligned}$$

Finally, since for any two $n \times n$ matrices $\mathbf{P}$ and $\mathbf{Q}$, the eigenvalues of $\mathbf{PQ}$ are the same as those of $\mathbf{QP}$, the optimal matrix $\mathbf{D}$ satisfies $\mathbf{D} = \arg\min_{\mathbf{A}\in \mathrm{DS}(\boldsymbol{\Sigma})} \lambda_1 \left(\mathbf{A}^{-1/2} \boldsymbol{\Sigma} \mathbf{A}^{-1/2} \right)$, which is the same as condition (13). ‡

Proof of Corollary 1 For any $t \in (0, t_{\max}]$,

$$\lambda_1 \left((t\mathbf{T})^{-1/2} \boldsymbol{\Sigma} (t\mathbf{T})^{-1/2} \right) = \lambda_1 \left(t^{-1} \mathbf{T}^{-1/2} \boldsymbol{\Sigma} \mathbf{T}^{-1/2} \right) = t^{-1} \lambda_1 \left(\mathbf{T}^{-1/2} \boldsymbol{\Sigma} \mathbf{T}^{-1/2} \right).$$

This shows that the largest eigenvalue of $(t\mathbf{T})^{-1/2} \boldsymbol{\Sigma} (t\mathbf{T})^{-1/2}$ is monotonically decreasing in $t \in (0, t_{\max}]$. Thus, within this class $\mathbf{D} = t_{\max}\mathbf{T}$ has the minimum largest eigenvalue and provides the optimal decomposition of $\boldsymbol{\Sigma}$. ‡

B Other conditional distributions

The conditional distributions for which details were not provided in Section 3.2 are as follows. For $i = 1, \ldots, N$ the conditional distributions for the individual-specific location parameters are $\tau_i \mid \mathbf{Y}_i, \sigma_i^2, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathrm{N}(m_i, u_i^2)$, where $u_i^2 = (\sigma_i^{-2} \mathbf{1}^T \boldsymbol{\Sigma}^{-1} \mathbf{1} + \sigma_\tau^{-2})^{-1}$ and $m_i = u_i^2 \sigma_i^{-2} (\mathbf{Y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} \mathbf{1}$. The conditional distributions for the individual-specific scale parameters are $\sigma_i^2 \mid \mathbf{Y}_i, \boldsymbol{\mu}, \tau_i, \boldsymbol{\Sigma} \sim \mathrm{IG}((a+M)/2\ ,\ ((\mathbf{Y}_i - \boldsymbol{\mu} - \tau_i \mathbf{1})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_i - \boldsymbol{\mu} - \tau_i \mathbf{1}) + b)/2)$. The conditional distribution for the common location parameter is $\boldsymbol{\mu} \mid \mathbf{Y}, \boldsymbol{\tau}, \boldsymbol{\sigma}, \boldsymbol{\Sigma} \sim \mathrm{N}(\boldsymbol{\gamma}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi} = (\mathrm{tr}(\mathbf{S}^{-2}) \boldsymbol{\Sigma}^{-1} + \mathbf{V}^{-1})^{-1}$, $\mathbf{S}^{-2} = \mathrm{diag}(\sigma_i^{-2})$ and $\boldsymbol{\gamma} = \boldsymbol{\Psi} \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \boldsymbol{\tau} \mathbf{1}_M^T)^T \mathbf{S}^{-2} \mathbf{1}_N$. The shared covariance matrix is updated from the conditional distribution $\boldsymbol{\Sigma} \mid \mathbf{Y}, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\sigma} \sim \mathrm{IW}(N + \delta, \mathbf{S}_N + \boldsymbol{\Sigma}_0)$, where $\mathbf{S}_N = \sum_{i=1}^N (\mathbf{Y}_i - \boldsymbol{\mu} - \tau_i \mathbf{1})(\mathbf{Y}_i - \boldsymbol{\mu} - \tau_i \mathbf{1})^T / \sigma_i^2$.

Supplemental Materials

Additional Appendices: Two additional appendices can be found in the online supplement to the paper. In these appendices we derive the optimal decomposition in a variety of special cases.

Acknowledgment

We thank the editor and two anonymous referees for comments that helped improve the clarity of the paper. The authors were supported by U.S. National Science Foundation grants SES-0437251, DMS-0604963, DMS-0706948, DMS-09-07070 and DMS-1007682, and by NSA grant H98230-10-1-0202. Any opinions, findings and conclusions, or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of the NSF or the NSA.

References

Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669–679.

Bradlow, E. T. and Zaslavsky, A. M. (1999). A hierarchical latent variable model for ordinal data from a customer satisfaction survey with "no answer" responses. *Journal of the American Statistical Association* **94**, 43–52.

Chen, M.-H. and Dey, D. (2000). Bayesian analysis for correlated ordinal data models. In D. K. Dey, S. K. Ghosh, and B. K. Mallick, eds., *Generalized Linear Models: A Bayesian Perspective*, 135–162. Marcel Dekker, New York.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* **85**, 347–361.

Chib, S. and Jeliazkov, I. (2006). Inference in semiparametric dynamic models for binary longitudinal data. *Journal of the American Statistical Association* **101**, 685–700.

Cowles, M. K. (1996). Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing* **6**, 101–111.

Damien, P. and Walker, S. (2001). Sampling truncated normal, beta, and gamma densities. *Journal of Computational and Graphical Statistics* **10**, 206–215.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics* **1**, 141–149.

Geweke, J. (1991). Efficient simulation from the multivariate normal and Student-t distributions subject to linear constraints and the evaluation of constraint probabilities. *Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface* **23**, 571–578.

Hajivassiliou, V., McFadden, D., and Ruud, P. (1996). Simulation of multivariate normal rectangle probabilities and their derivatives: theoretical and computational results. *Journal of Econometrics* **72**, 85–134.

Hsu, J. C. (1992). The factor analytic approach to simultaneous inference in the general linear model. *Journal of Computational and Graphical Statistics* **1**, 151–168.

Javaras, K. and Ripley, B. (2007). An "unfolding" latent variable model for Likert attitude data: Drawing inferences adjusted for response style. *Journal of the American Statistical Association* **102**, 454–463.

Jeliazkov, I., Graves, J., and Kutzbach, M. (2008). Fitting and comparison of models for multivariate ordinal outcomes. In S. Chib, W. Griffiths, G. Koop, and D. Terrell, eds., *Bayesian Econometrics*, vol. 23 of *Advances in Econometrics*, 115–156. Emerald Group Publishing Ltd.

Johnson, V. E. and Albert, J. H. (1999). *Ordinal Data Modeling.* Springer–Verlag, New York.

Keane, M. (1994). A computationally practical simulation estimator for panel data. *Econometrica* **62**, 95–116.

Kottas, A., Müller, P., and Quintana, F. (2005). Nonparametric Bayesian modeling for multivariate ordinal data. *Journal of Computational and Graphical Statistics* **14**, 610–625.

Liechty, M. W. and Lu, J. (2010). Multivariate normal slice sampling. *Journal of Computational and Graphical Statistics* **19**, 281–294.

Liu, J. S., Wong, W. H., and Kong, A. (1994). Covariance structure of the Gibs sampler with applications to comparisons of estimators and augmentation schemes. *Biometrika* **81**, 27–40.

MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics* B **23**, 727–741.

MacEachern, S. N. and Müller, P. (1998). Estimating mixtures of Dirichlet process models. *Journal of Computational Statistics* **7**, 223–238.

McCulloch, R. and Rossi, P. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics* **64**, 207–240.

McCulloch, R. E., Polson, N. G., and Rossi, P. E. (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics* **99**, 173–193.

Nandram, B. and Chen, M.-H. (1996). Reparameterizing the generalized linear model to accelerate Gibbs sampler convergence. *Journal of Statistical Computation and Simulation* **54**, 129–144.

Roberts, G. O. and Sahu, S. K. (1997). Updating schemes, correlation structure, blocking and parameterisation for the Gibbs sampler. *Journal of the Royal Statistical Society,* Series B **59**, 291–317.

Rossi, P. and McCulloch, R. (2008). *bayesm: Bayesian Inference for Marketing/Micro-economics.* R package version 2.2-2.

Rossi, P. E., Allenby, G. M., and McCulloch, R. (2005). *Bayesian Statistics and Marketing.* John Wiley & Sons, Ltd.

Rossi, P. E., Gilula, Z., and Allenby, G. M. (2001). Overcoming scale usage heterogeneity: A Bayesian hierarchical approach. *Journal of the American Statistical Association* **96**, 20–31.

Stern, S. (1992). A method for smoothing simulated moments of discrete probabilities in multinomial probit models. *Econometrica* **60**, 943–952.

Stern, S. (1997). Simulation-based estimation. *Journal of Economic Literature* **35**, 2006–2039.

van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics* **10**, 1–50.

van Dyk, D. A. and Park, T. (2008). Partially collapsed Gibbs samplers: Theory and methods. *Journal of the American Statistical Association* **103**, 790–796.

Webb, E. L. and Forster, J. J. (2008). Bayesian model determination for multivariate ordinal and binary data. *Computational Statistics and Data Analysis* **52**, 2632–2649.

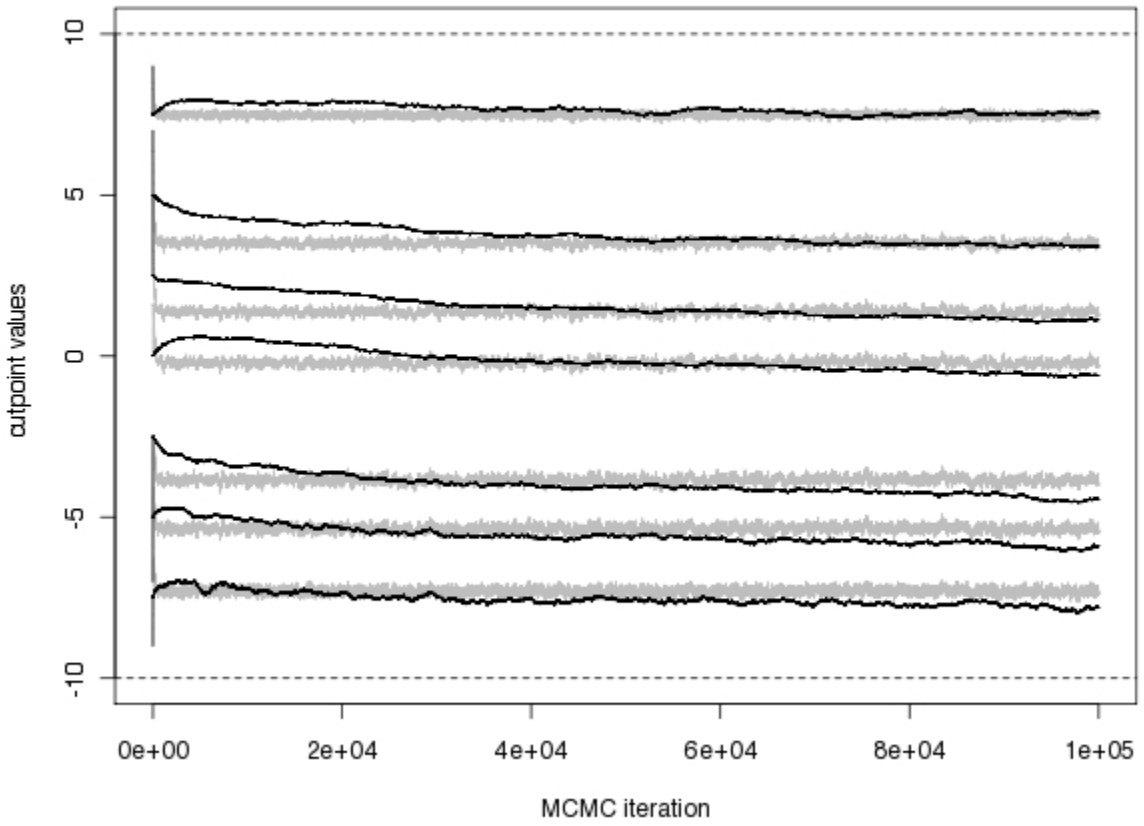

Figure 1: MCMC output for the cutpoints $c_2, \ldots, c_8$ for the illustration in Section 5. The black chains are from the Standard Sampler and the gray chains are from the Decomposition Sampler. The samplers were initialized at the same values. The chains were subsampled every ten iterations for plotting purposes; the figure is practically indistinguishable from the figure based on the full sample.

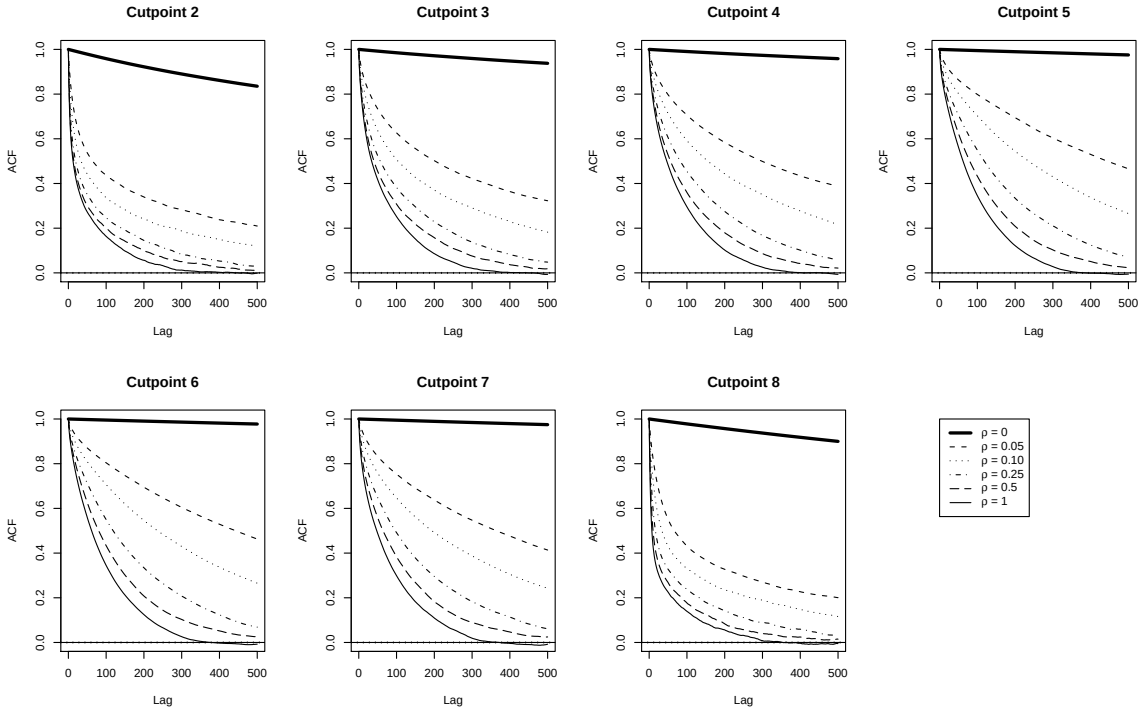

Figure 2: Autocorrelation functions for the cutpoints $c_2, \ldots, c_8$ for various values of ρ. The thick black line ($\rho = 0$) corresponds to the Standard Sampler (Algorithm 1); all other lines correspond to output from the Decomposition Sampler (Algorithm 3).

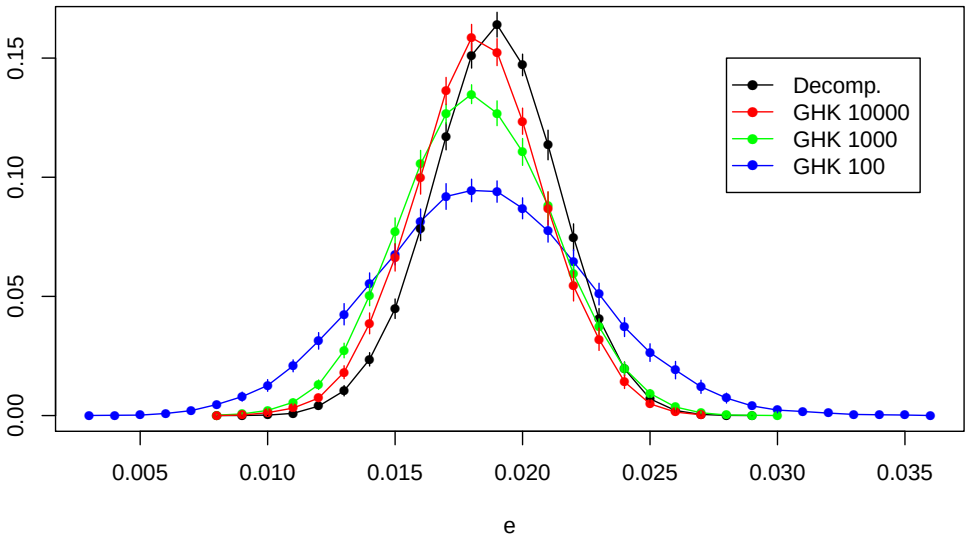

Figure 3: Estimated posterior distribution for the parameter e controlling the cutpoints in the Rossi *et al.* (2001) model. "Decomp." refers to the modified `rscaleUsage` function that uses the new covariance decomposition data augmentation. The several lines labeled "GHK" correspond to the unmodified `rscaleUsage` function using the specified values of n_{ghk}.

Covariance Decompositions for Accurate Computation in Bayesian Scale-Usage Models

Online Supplemental Materials

Chris Hans[1], Greg M. Allenby[2], Peter F. Craigmile[1], Ju Hee Lee[1], Steven MacEachern[1], and Xinyi Xu[1]

[1]Department of Statistics, The Ohio State University, Columbus, OH 43210.

[2]Department of Marketing, The Ohio State University, Columbus, OH 43210.

C Optimal Decomposition in Special Cases

C.1 Independence structure

Suppose that the distribution of $\mathbf{Y}_i \mid \boldsymbol{\mu}, \tau_i, \sigma_i^2$ is normal with diagonal covariance matrix $\boldsymbol{\Sigma}$. Then the optimal $\mathbf{D}$ is $\mathbf{D} = \boldsymbol{\Sigma}$, and all the eigenvalues of $\mathbf{D}^{-1/2}\boldsymbol{\Sigma}\mathbf{D}^{-1/2}$ are 1. Thus, $\lambda_1(\mathbf{D}^{-1/2}\boldsymbol{\Sigma}\mathbf{D}^{-1/2}) = 1$ and $\mathbf{R} = \boldsymbol{\Sigma} - \mathbf{D} = \mathbf{0}$. Convergence is immediate, and the Markov chain yields independent draws from the limiting distribution.

Suppl. Example 1 *(One-way ANOVA, independence prior) The case of* $\boldsymbol{\Sigma} = (\sigma_a^2 + \sigma_e^2)\mathbf{I}$ *with* σ_a^2 *and* σ_e^2 *known corresponds to a model with an i.i.d.* $N(0, \sigma_a^2)$ *treatment effect plus an i.i.d.* $N(0, \sigma_e^2)$ *error. The optimal* $\mathbf{D}$ *maximizing the convergence rate is* $\boldsymbol{\Sigma}$.

C.2 Exchangeable correlation structure

Suppose that, among the coordinates of $\mathbf{Y}_i \mid \boldsymbol{\mu}, \tau_i, \sigma_i^2$, a set of variables is exchangeable in the sense that the correlation matrix $\mathbf{C}$ remains unchanged under any permutation of these variables. In this case, the following proposition shows that an optimal $\mathbf{D}$ based on the correlation matrix $\mathbf{C} = \mathbf{V}^{-1/2}\boldsymbol{\Sigma}\mathbf{V}^{-1/2}$ has equal diagonal elements at the exchangeable coordinates.

Suppl. Proposition 1 *Suppose that* $\mathbf{Y}_i \mid \boldsymbol{\mu}, \tau_i, \sigma_i^2$ *follows a multivariate normal distribution, and there exist two variables* (Y_{ij}, Y_{ik}) *that are exchangeable. Then there exists an optimal matrix* $\mathbf{D} = diag(d_1, \cdots, d_M)$ *based on the correlation matrix* $\mathbf{C}$ *with* $d_j = d_k$.

Proof of Suppl. Proposition 1 Assume that $\mathbf{D}_0 = \text{diag}(d_0^1, \cdots, d_0^M) \in \text{DS}(\mathbf{C})$ is an optimal matrix that minimizes the largest eigenvalue value of $\mathbf{D}_0^{-1/2}\mathbf{C}\mathbf{D}_0^{-1/2}$ and $d_0^i \neq d_0^j$. We can construct a new diagonal matrix $\mathbf{D}_1 = \text{diag}(d_1^1, \cdots, d_1^M)$ where $d_1^k = d_0^k$ for $k \neq i$ or j, $d_1^i = d_0^j$ and $d_i^j = d_0^i$, i.e., $\mathbf{D}_1$ swaps the positions of d_0^i and d_0^j. This corresponds to a relabeling of coordinates i and j. Thus, $\mathbf{D}_1 \in \text{DS}(\mathbf{C})$, and the eigenvalues of $\mathbf{D}_1^{-1/2}\mathbf{C}\mathbf{D}_1^{-1/2}$ are the same as those of $\mathbf{D}_0^{-1/2}\mathbf{C}\mathbf{D}_0^{-1/2}$, although the eigenvectors may differ. Now let $\mathbf{D}^* = \frac{1}{2}(\mathbf{D}_0 + \mathbf{D}_1)$, then $\mathbf{D}^*$ is still in $\text{DS}(\mathbf{C})$, and by a well-known result in linear algebra (see, e.g., Bhatia, 1996),

$$\begin{aligned} \lambda_1((\mathbf{D}^*)^{-1/2}\mathbf{C}(\mathbf{D}^*)^{-1/2}) &\leq \frac{1}{2}\left[\lambda_1(\mathbf{D}_0^{-1/2}\mathbf{C}\mathbf{D}_0^{-1/2}) + \lambda_1(\mathbf{D}_1^{-1/2}\mathbf{C}\mathbf{D}_1^{-1/2})\right] \\ &= \lambda_1(\mathbf{D}_0^{-1/2}\mathbf{C}\mathbf{D}_0^{-1/2}). \end{aligned}$$

Therefore, $\lambda_1((\mathbf{D}^*)^{-1/2}\mathbf{C}(\mathbf{D}^*)^{-1/2})$ is at most as large as $\lambda_1(\mathbf{D}_0^{-1/2}\mathbf{C}\mathbf{D}_0^{-1/2})$, and so $\mathbf{D}^*$ is an optimal matrix based on the correlation matrix. ‡

The key idea in Suppl. Proposition 1 is that, for a given i, the Y_{ij}'s can be reordered without changing the correlation matrix, then the corresponding diagonal elements of a matrix $\mathbf{D}$ can be reordered in the same way without affecting the convergence rate of the Markov chain. Convexity suggests that averaging $\mathbf{D}$ and its reordered version can only hasten convergence. This argument can be applied or extended in various cases to obtain optimal decompositions. We illustrate this procedure for the following well known models.

Case 1: Exchangeable correlation structure Consider the exchangeable correlation structure, where all coordinates of $\mathbf{Y}_i \mid \boldsymbol{\mu}, \tau_i, \sigma_i^2$ are exchangeable, i.e., the correlation matrix $\mathbf{C}$ is of the form $\mathbf{C} = a\mathbf{I} + b\mathbf{J}$ with $a = 1 - b$. The following corollary shows that in this situation, an optimal $\mathbf{D}$ for $\boldsymbol{\Sigma}$ is proportional to $\text{diag}(\boldsymbol{\Sigma})$.

Suppl. Corollary 1 *If the correlation matrix* $\mathbf{C}$ *is of the form* $a\mathbf{I} + b\mathbf{J}$ *where* $a = 1 - b$, *then the matrix* $\mathbf{D} = d\ diag(\boldsymbol{\Sigma})$ *is optimal, where* $d = 1 - b$ *if* $b \geq 0$ *and* $d = 1 + (M-1)b$ *if* $b < 0$.

Proof of Suppl. Corollary 1 By Proposition 2 in Section 4.1 of the main text of the paper, it suffices to show that an optimal $\mathbf{D}$ based on the correlation matrix is $d\mathbf{I}$. Since, when the correlation matrix is $a\mathbf{I}+b\mathbf{J}$, any pair within $Y_{i1}, \cdots, Y_{iM} \mid \mu, \tau_i, \sigma_i^2$ is exchangeable, by Suppl. Proposition 1 an optimal $\mathbf{D}$ is proportional to the identity matrix. Moreover, it is easy to see that for any fixed t, the largest eigenvalue of $(t\mathbf{I})^{-1/2}\mathbf{\Sigma}(t\mathbf{I})^{-1/2}$ is $\lambda_1(\mathbf{C})/t$. Thus, to retain $\mathbf{R}$ as a non-negative definite matrix, the largest t that can be used is $t_{\max} = \lambda_M(\mathbf{C})$, which is $1-b$ if $b \geq 0$ and is $1+(M-1)b$ if $b < 0$. ‡

Suppl. Example 2 *(One-way ANOVA, hierarchical prior) With the one-way ANOVA model now assume that the prior distribution for the treatment is hierarchical. The center of the distribution of the treatment effects, μ, follows the $N(0, \sigma_\mu^2)$ distribution. The M treatment effects are jointly $N(\mu\mathbf{1}, \mathbf{I})$, conditional on their center. This implies that the treatment effects are jointly $N(\mathbf{0}, \mathbf{I}+\sigma_\mu^2\mathbf{J})$, and thus the correlation matrix is $\mathbf{C} = a\mathbf{I}+b\mathbf{J}$, where $a = 1-b$ and $b = \sigma_\mu^2/(1+\sigma_e^2+\sigma_\mu^2)$. Appealing to Corollary 1, the optimal $\mathbf{D}$ is $(1+\sigma_e^2)/(1+\sigma_e^2+\sigma_\mu^2)\mathbf{I}$.*

Case 2: Circular correlation structure Consider the circular correlation structure, where $Y_{ik}, \cdots, Y_{iM}, Y_{i1}, \cdots, Y_{i,k-1} \mid \boldsymbol{\mu}, \tau_i, \sigma_i^2$ is the same as the distribution of $Y_{i1}, \cdots, Y_{iM} \mid \boldsymbol{\mu}, \tau_i, \sigma_i^2$ for any $k = 1, \ldots, M$. Thus, the covariance matrix remains the same under a circular transformation of the coordinates. In this case, we can easily extend the symmetric argument in Suppl. Proposition 1 to show that an optimal $\mathbf{D}$ is a multiplier of the identity matrix.

Suppl. Corollary 2 *Suppose that $Y_{i1}, \cdots, Y_{iM} \mid \boldsymbol{\mu}, \tau_i, \sigma_i^2$ follows a multivariate normal distribution and its covariance matrix is invariant under circular transformations. Then an optimal $\mathbf{D}$ in the decomposition $\mathbf{\Sigma} = \mathbf{D} + \mathbf{R}$ is $\mathbf{D} = t\mathbf{I}$, where $t = \lambda_M(\mathbf{\Sigma})$.*

Proof of Suppl. Corollary 2 Following similar steps as in the proof of Proposition 1, we can see that there exists an optimal $\mathbf{D}$ that satisfies $(d_k, \cdots, d_M, d_1, \cdots, d_{k-1}) = (d_1, \cdots, d_M)$ for each $k = 1, \ldots, M$. That is, $d_1 = \cdots = d_M = t$ for some $t > 0$. To retain $\mathbf{R}$ as a nonnegative definite matrix, the largest t that we can choose is $t_{\max} = \lambda_M(\mathbf{C})$. Therefore, this optimal $\mathbf{D}$ is equal to $\lambda_M(\mathbf{\Sigma})\mathbf{I}$. ‡

We demonstrate this theory in the case of the circular AR(1) process. Further details

of the process, eigenvalue decomposition, and another example (the circular MA(1) process) are provided in Appendix D below.

Suppl. Example 3 *(Circular AR(1) process) For $M > 2$, the stationary circular autoregressive process of order 1 has a covariance matrix $\mathbf{\Sigma}$ that is circular, with $\Sigma_{jk} = \gamma_{i-j \bmod M}$ and $\gamma_h = \sigma^2(\phi^{|h|} + \phi^{M-|h|})/((1-\phi^2)(1-\phi^M))$, $h = 0, \ldots, M-1$, for autocorrelation parameter $-1 < \phi < 1$ and innovation variance $\sigma^2 > 0$. The smallest eigenvalue of $\mathbf{\Sigma}$ when $\phi \geq 0$ is $\lambda_M(\mathbf{\Sigma}) = \sigma^2/(1 - 2\phi\cos(2\pi\lfloor M/2 \rfloor/M) + \phi^2)$ (which simplifies to $\lambda_M(\mathbf{\Sigma}) = \sigma^2/(1+\phi)^2$ when M is even). When $\phi < 0$, the smallest eigenvalue is $\lambda_M(\mathbf{\Sigma}) = \sigma^2/(1-\phi)^2$. Therefore, by Corollary 2 an optimal $\mathbf{D}$ based on $\mathbf{\Sigma}$ is $\lambda_M(\mathbf{\Sigma})\mathbf{I}$.*

Case 3: Reversible correlation structure Consider the reversible correlation structure, where the correlation matrix remains unchanged when the order of the variables is reversed. In this case, we characterize a property of the optimal $\mathbf{D}$. The proof follows the argument in Suppl. Proposition 1.

Suppl. Corollary 3 *Suppose that the correlation matrix has the form $C_{ij} = C_{M+1-i\, M+1-j}$, for $j = 1, \ldots, M$. Then the optimal $\mathbf{D}$ matrix based on $\mathbf{C}$ has, with $diag(\mathbf{D}) = (d_1, \ldots, d_M)$, $d_i = d_{M+1-i}$ for $i = 1, \ldots, M$.*

Suppl. Example 4 *(AR(1) process). An autoregressive process of order 1 has a correlation matrix which is reversible. Direct application of Suppl. Corollary 3 implies that the optimal $\mathbf{D}$ is symmetric, with $d_j = d_{M+1-j}$ for $j = 1, \ldots, M$.*

C.3 Block diagonal structure

Suppose that the observations $Y_{i1}, \cdots, Y_{iM}$ $(i = 1, \ldots, N)$ can be divided into several parts where variables in different parts are independent conditional on $\boldsymbol{\mu}, \tau_i, \sigma_i^2$, i.e., the covariance matrix $\mathbf{\Sigma}$ is a block diagonal matrix. Then we can divide the matrix $\mathbf{D}$ into several corresponding blocks. The optimization problems in different blocks are independent, and the overall convergence rate is determined by the "worst" block. The next result shows that one can optimize each block separately and then paste the pieces together to get a big $\mathbf{D}$ matrix.

Suppl. Proposition 2 *Suppose the covariance matrix* $\mathbf{\Sigma}$ *is block diagonal with* k *blocks, and* $\mathbf{D} \in DS(\mathbf{\Sigma})$ *is diagonal. Then the matrix* $\mathbf{D}^{-1/2}\mathbf{\Sigma}\mathbf{D}^{-1/2}$ *is also block diagonal with* k *blocks, and its eigenvalues are of the form* $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_k$*, where* $\boldsymbol{\lambda}_i$ *is the vector of eigenvalues from the* i^{th} *block. Solving each block and collecting the results together guarantees an optimal solution.*

Suppl. Example 5 *(Two-way ANOVA) The prior distribution for the* J *treatment effects is* $N(\mu\mathbf{1}, \sigma_\alpha^2\mathbf{I})$*, conditional on a known value* μ*. The* n_j *replicate measurements on treatment* j *are conditionally independent, with mean equal to the treatment mean and variance* σ_e^2*. The covariance matrix of* $\mathbf{Y}_i$ *is* $\mathbf{\Sigma} = diag(\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_J)$*, with* $\mathbf{\Sigma}_i = \sigma_e^2\mathbf{I} + \sigma_\alpha^2\mathbf{J}$*. By Suppl. Proposition 2 and Suppl. Corollary 1, an optimal* $\mathbf{D} = diag(\mathbf{D}_1, \ldots, \mathbf{D}_J)$ *where* $\mathbf{D}_j = (1 + \sigma_e^2)/(1 + \sigma_e^2 + \sigma_\alpha^2)\mathbf{I}$.

D Eigenvalues of a circulant matrix

For a positive integer M, a circulant $M \times M$ covariance matrix $\mathbf{\Sigma}$ is defined by the relation $\Sigma_{jk} = \gamma_{j-k \bmod M}$, for M constants $\gamma_0, \ldots, \gamma_{M-1}$, such that $\mathbf{\Sigma}$ is positive definite. In this case there is a closed from expression for the eigenvalues of $\mathbf{\Sigma}$. The unordered eigenvalues are calculated using the discrete Fourier transform (DFT) of $\{\gamma_k : k = 0, \ldots, M-1\}$ (Gray, 2006), where $i = \sqrt{-1}$:

$$\psi_j = \sum_{k=0}^{M-1} \gamma_k e^{-i2\pi(j-1)k/M}, \qquad j = 1, \ldots, M.$$

An example of a process with a circular covariance matrix is the circular autoregressive process of order one. For an integer $M \geq 2$, let $\{U_t : t = 0, \ldots, M-1\}$ be a set of uncorrelated mean zero random variables with variance σ^2, such that $0 < \sigma^2 < \infty$. Then the circular AR(1) process is defined by the recursion, $\eta_t = \phi\eta_{t-1 \bmod M} + U_t$, $t = 0, \ldots, M-1$. For $|\phi| < 1$ this process is stationary and for each t we can express η_t as $\eta_t = (1 - \phi^M)^{-1} \sum_{k=0}^{M-1} \phi^k U_{t-k \bmod M}$, which leads to that fact that $E(\eta_t) = 0$ for all t and for $|h| < M$,

$$\gamma_h = \text{cov}(\eta_t, \eta_{t+h}) = \frac{\sigma^2(\phi^{|h|} + \phi^{M-|h|})}{(1-\phi^2)(1-\phi^M)}.$$

The DFT of $\{\gamma_k\}$ is

$$\psi_j \;=\; \sum_{k=0}^{M-1} \gamma_k e^{-i2\pi(j-1)k/M} \;=\; \frac{\sigma^2}{1-2\phi\cos(2\pi(j-1)/M)+\phi^2}, \;\; j=1,\ldots,M.$$

For $\phi \geq 0$, the smallest eigenvalue occurs at $j = \lfloor M/2 \rfloor + 1$, with value $\sigma^2/(1-2\phi\cos(2\pi\lfloor M/2\rfloor/M)+\phi^2)$, which simplifies to $\sigma^2/(1+\phi^2)$ when M is even. For $\phi < 0$, the smallest eigenvalue occurs at $j = 1$ with value $\sigma^2/(1-\phi^2)$.

Another example is the circular moving average (MA) process of order one. For some integer $M \geq 3$ and $\theta \neq 0$, the circular MA(1) process is defined by $\eta_t = U_t + \theta U_{t-1 \bmod M}, \;\; t = 0,\ldots,M-1$, where $\{U_t\}$ was defined as for the circular AR(1) process. We restrict to the case that the process is invertible; i.e., when $|\theta| < 1$ (e.g., Brockwell and Davis, 2002). Then this process has mean zero with a covariance structure described, for $|h| < M$ by,

$$\gamma_h = \mathrm{cov}(\eta_t, \eta_{t+h}) = \begin{cases} \sigma^2(1+\theta^2), & h = 0 \\ \sigma^2\theta, & h = \pm 1, \pm(M-1) \\ 0, & \text{otherwise.} \end{cases}$$

The $M \times M$ covariance matrix again is circular with $\Sigma_{jk} = \gamma_{j-k \bmod M}$, and the unordered eigenvalues are $\psi_j = \sigma^2\left(1 + 2\theta\cos(2\pi(j-1)/M) + \theta^2\right)$ for $j = 1,\ldots,M$. When θ is positive, the minimum eigenvalue occurs at $j = \lfloor M/2 \rfloor + 1$, with value $\sigma^2\left(1 + 2\theta\cos(2\pi\lfloor M/2\rfloor/M) + \theta^2\right)$, which simplifies to a value of $\sigma^2\left(1-\theta\right)^2$ when M is even. When θ is negative, the minimum eigenvalue occurs at $j = 1$, with value $\sigma^2 = \sigma^2(1+2\theta+\theta^2) = \sigma^2(1+\theta)^2$.

References

Bhatia, R. (1996). *Matrix Analysis.* Springer, New York.

Brockwell, P. J. and Davis, R. A. (2002). *Introduction to Time Series and Forecasting.* Springer, New York.

Gray, R. M. (2006). *Toeplitz and Circulant Matrices: A review.* Now Publishers, Norwell, Massachusetts.