Diagonal Orthant Multinomial Probit Models

James E. Johndrow
Duke University

Kristian Lum
Virginia Tech

David B. Dunson
Duke University

Abstract

Bayesian classification commonly relies on probit models, with data augmentation algorithms used for posterior computation. By imputing latent Gaussian variables, one can often trivially adapt computational approaches used in Gaussian models. However, MCMC for multinomial probit (MNP) models can be inefficient in practice due to high posterior dependence between latent variables and parameters, and to difficulties in efficiently sampling latent variables when there are more than two categories. To address these problems, we propose a new class of diagonal orthant (DO) multinomial models. The key characteristics of these models include conditional independence of the latent variables given model parameters, avoidance of arbitrary identifiability restrictions, and simple expressions for category probabilities. We show substantially improved computational efficiency and comparable predictive performance to MNP.

1 Introduction

This work is motivated by the search for an alternative to the multinomial logit (MNL) and multinomial probit (MNP) models that is more amenable to efficient Bayesian computation, while maintaining flexibility. Historically, the MNP has been preferred for Bayesian inference in polychotomous regression, since the data augmentation approach of Albert and Chib [1993] leads to straightforward Gibbs sampling. Efficient methods for Bayesian inference in the MNL are a more recent development. A series of proposed data-augmentation methods for Bayesian inference in the MNL dates at least to O'Brien and Dunson [2004], who use a student t data augmentation scheme with the latent t variables expressed as scale mixtures of Gaussians. The scale and degrees of freedom in the t are chosen to provide a near exact approximation to the logistic density. Holmes and Held [2006] represent the logistic distribution as a scale-mixture of normals where the scales are a transformation of Kolmogorov-Smirnov random variables. citet fruhwirth2009improved propose an alternative data-augmentation scheme which approximates a log-Gamma distribution with a mixture of normals, resulting in conditionally-conjugate updates for regression coefficients. Polson et al. [2012] develop a novel data-augmented representation of the likelihood in a logistic regression using Polya-Gamma latent variables. With a normal prior, the regression coefficients have conditionally-conjugate posteriors. Their method has the advantage that the latent variables can be sampled directly via an efficient rejection algorithm without the need to rely on additional auxiliary variables.

Although the work outlined above has opened MNL to Gibbs sampling, the distributions of the latent variables are either exotic or complex scale-mixtures of normals for which multivariate analogues are not simple to work with. As such, the extension to analysis of multivariate unordered categorical data, nonparametric regression, and other more complex situations is not straightforward. In contrast, the multinomial probit (MNP) model is trivially represented by a set of Gaussian latent variables, allowing access to a wide range of methods developed for Gaussian models. MNP is also a natural choice for Bayesian estimation because of the fully-conjugate updates for model parameters and latent variables. However, because the latent Gaussians are not conditionally independent given regression parameters, mixing is poor and computation does not scale well.

Three characteristics of the multinomial probit model lead to challenging computation and render it of limited use in complex problems: the need for identifying restrictions, the specification of non-diagonal covariance matrices for the residuals when it is not well-motivated, and high dependence between latent vari-

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31. Copyright 2013 by the authors.

ables. Because our proposed method addresses all three of these issues, we review each of them here and summarize the pertinent literature.

1.1 Identifying Restrictions

Many choices of identifying restrictions for the MNP have been suggested, and the choice of identifying restrictions has important implications for Bayesian inference (Burgette and Nordheim [2012]). Consider the standard multinomial probit, where here we assume a setting where covariates are constant across levels of the response (i.e. a classification application):

$$\begin{aligned} y_i = j &\Leftrightarrow u_{ij} = \bigvee_k u_{ik} \\ u_{ij} &= \boldsymbol{x}_i \boldsymbol{\beta}_j + \boldsymbol{\epsilon}_{ij} \\ \boldsymbol{\epsilon}_i &\sim N(\mathbf{0}, \boldsymbol{\Sigma}) \end{aligned}$$

where $\bigvee$ is the max function. The u_{ij}'s are referred to as latent utilities, after the common economic interpretation of them as unobserved levels of welfare in choice models. We make the distinction between a classification application as presented above and the choice model common to the economics literature in which each category has its own set of covariates (i.e. $u_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \epsilon_{ij}$). A common approach to identifying the model is to choose a base category and take differences. Suppose we select category 1 as the base. We then have the equivalent model:

$$\begin{aligned} \tilde{u}_{i1} &= 0 \\ \tilde{u}_{ij} &= \boldsymbol{x}_i(\boldsymbol{\beta}_j - \boldsymbol{\beta}_1) + \epsilon_{ij} - \epsilon_{i1} \end{aligned}$$

The $\tilde{u}$'s are a linear transformation $\boldsymbol{M}$ of the original latent utilities, so $\tilde{\boldsymbol{u}}_{i,2:J} \sim N(\boldsymbol{x}_i(\boldsymbol{\beta}_{2:J} - \boldsymbol{\beta}_1), \boldsymbol{M}^T\boldsymbol{\Sigma}\boldsymbol{M})$. Early approaches to Bayesian computation in the MNP placed a prior on $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{M}^T\boldsymbol{\Sigma}\boldsymbol{M}$. Even this parametrization is not fully identified, and requires that one of the variances be fixed (usually to one) to fully identify the model. McCulloch and Rossi [1994] ignored this issue and adopted a parameter-expanded Gibbs sampling approach by placing an inverse-Wishart prior on $\tilde{\boldsymbol{\Sigma}}$. McCulloch et al. [2000] developed a prior specification on covariance matrices with the upper left element restricted to one. Imai and Van Dyk [2005] use a working parameter that is resampled at each step of the Gibbs sampler to transform between identified and unidentified parameters, with improved mixing, though the resulting full conditionals are complex and the sampling requires twice the number of steps as the original McCullough and Rossi sampler. Zhang et al. [2006] used a parameter-expanded Metropolis-Hastings algorithm to obtain samples from a correlation matrix, resulting in an identified model that restricts the scales of the utilities to be the same. Zhang et al. [2008] extended their algorithm to multivariate multinomial probit models.

Whatever prior specification and computational approach is used, the selection of an arbitrary base category is an important modeling decision that may have serious implications for mixing (Burgette and Hahn [2010]). The class probabilities in the MNP are linked to the model parameters by integrating over subsets of $\mathbb{R}^J$. The choice of one category as base results in these regions having different geometries, causing an asymmetry in the effect of translations of the elliptical multivariate normal density on the resulting class probabilities. To address this issue, Burgette and Nordheim [2012] and Burgette and Hahn [2010] rely on an alternative identifying restriction, circumventing the need to select an arbitrary base category. Although mixing is improved and the model has an appealing symmetry in the representation of categorical probabilities as a function of latent variables, it does not address the issue of high dependence between latent variables, and the proposed Gibbs sampler is quite complex (though computationally no slower than simpler algorithms).

1.2 Dependence in Idiosyncratic Errors

The multinomial probit model arose initially in the econometrics literature (see e.g. Hausman and Wise [1978]), where dependence between the errors in the linear utility models is often considered desirable. In econometrics and marketing, interest often centers on predicting the effects of changes in product prices on product shares. If the errors have a diagonal covariance matrix, the model has the "independence of irrelevant alternatives" (IIA) property (see Train [2003]), which is often considered too restrictive for characterizing the substitution patterns between products in a market. Because early applications for the MNP were largely of the marketing and econometrics flavor, it has become standard to specify a model with dependence in the errors without much attention to whether it is scientifically motivated. We find little compelling reason for this assumption for most applications outside of economics and marketing. If the application does not motivate dependence in errors, the resulting model will certainly have higher variance than a model with independent errors, and the additional dependence between latent variables will negatively impact mixing. In this case, the applications we have in mind are not specifically economics/marketing applications, and as such we will generally assume that the errors are conditionally independent given regression parameters.

1.3 Dependence Between Latent Variables

Except in the special case of marketing applications where there are covariates that differ with the level of the response variable (such as the product price), an identified MNP must have $J-1$ sets of regression coefficients, with one set restricted to be zero. If we assume $\boldsymbol{\beta}_1 = 0$ and an identity covariance matrix, we can actually sample all J latent variables for each observation i. Note this is equivalent to setting one class to have utility that is identically zero and sampling the remaining utility differences from the transformed distribution described in section 2.1; however, the intractability of the high dependence between utilities is much clearer using this alternative parametrization.

To implement data augmentation MCMC, we must sample the latent utilities conditional on regression coefficients from a multivariate normal distribution restricted to the space where $u_{i,y_i} = \bigvee_k u_{ik}$. Albert and Chib [1993] suggest rejection sampling; but this tends to be extremely inefficient, particularly as J grows. All subsequent algorithms have sampled $u_{ij} \mid u_{i,-j}$ in sequence. With an identity covariance matrix for latent utilities conditional on regression parameters, we can reduce this to a two-step process:

1. Set $b_i = \bigvee_{k \neq y_i} u_{ik}$. Sample $\tilde{u}_{[i,y_i]} \sim N_{[b,\infty)}(x_i\beta_{y_i}, 1)$.
2. Sample $u_{[i,j]} \sim N_{(-\infty,u_{[i,y_i]}]}(x_i\beta_{[i,j]}, 1)$ independently for all $j \neq y_i$.

Even in simple cases, mixing tends to be poor because the truncation region for each latent Gaussian depends on the current value in the Markov chain for the other latent Gaussians, resulting in high dependence. We reiterate that this issue is not simply a consequence of the choice of a non-diagonal covariance matrix; it is true for any choice of covariance matrix for the ϵ's. While recent work on developing a Hamiltonian MC scheme for jointly sampling from truncated multivariate normal distributions appears promising as an alternative to the standard Gibbs sampling approach (Pakman and Paninski [2012]), the need to update each latent variable conditional on the others has historically been a substantial contibutor to the inefficiency of Bayesian computation for MNP.

In this paper we propose Diagonal Orthant Multinomial models (DO models), a novel class of models for unordered categorical response data that admits latent variable representations (including a Gaussian latent variable representation for the DO-probit model), does not require the selection of a base category, and in which the latent variables are conditionally independent given regression parameters and thus may be updated in a block, greatly improving mixing and computational scaling. The remainder of the paper is organized as follows. In section 2, we introduce DO models, and show that a unique solution to the likelihood equations can be obtained from independent binary regressions. We also illustrate the relationship of DO-logistic to MNL and DO-probit to MNP. We give several interpretations for regression coefficients in DO models, and explain why the model parameters are often easier to interpret than in the MNP. In section 3, we outline a simple algorithm for Bayesian computation in the DO-probit and discuss extensions to the basic regression setting. In section 4, we compare the DO-probit, MNP, and MNL in simulation studies. In section 5, we apply both methods to a real dataset and show that they are virtually indistinguishable in prediction. In section 6 we conclude and discuss potential future directions for this work.

2 The Diagonal Orthant Multinomial Model

The Diagonal Orthant Multinomial (DO) class of models represent an unordered categorical variable as a set of binary variables. Let y be unordered categorical with J levels and suppose $\gamma_{[1:J]}$ are independent binary variables. Define $y = j \Leftrightarrow \{\gamma_j = 1\} \cup \{\gamma_k = 0 \ \forall \ k \neq j\}$. Binary variables have a well-known latent variable representation. Let $z_j \sim f(\mu_j, \sigma)$ where f is a location-scale density with location parameter μ_j and common scale σ, and set $\gamma_j = 1 \Leftrightarrow z_j > 0$. However, for our purposes, we must ensure that only one γ_j is one, and thus we restrict the z's to belong to the set:

$$\Omega = \bigcup_{j=1}^{J} \{z \in \mathbb{R}^J : z_j > 0, z_k < 0, k \neq j\}$$

By the Radon-Nikodym theorem, the joint distribution of the z's is:

$$f(z) = \frac{\mathbf{1}(\boldsymbol{z} \in \Omega) \prod_{j=1}^{J} f(z_j - \mu_j)}{\displaystyle\int_{\mathbb{R}^J} \mathbf{1}(\boldsymbol{z} \in \Omega) \prod_{j=1}^{J} f(z_j - \mu_j) d\boldsymbol{z}}$$

If we let $f = \phi(\cdot)$, where $\phi(\cdot)$ is the univariate normal pdf, the result is a probit analogue that we refer to as DO-Probit.

The joint probability density of z's in DO-probit is that of a J-variate normal distribution with identity covariance that is restricted to the regions of $\mathbb{R}^J$ with one sign positive and the others negative. This is easy to visualize in $\mathbb{R}^2$, where the density is a bivariate normal with correlation zero and unit variance restricted to the second and fourth quadrants.

The density in the two dimensional case is shown in Figure 1. In higher dimensions, the restriction will define orthants over which the density is nonzero. The marginal distribution of any two latent variables will always be restricted to orthants that are diagonally apposed rather than adjacent, hence the designation Diagonal Orthant Multinomial model.

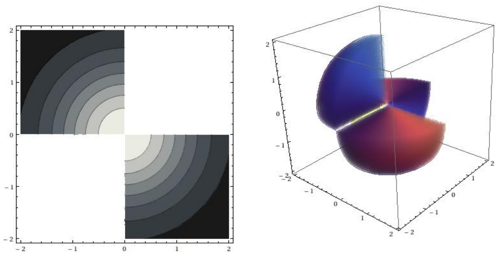

Figure 1: The joint density of the z's in the DO-probit for the 2-category case with zero means (left panel). The right panel shows approximate semispherical regions covering 95 percent of the total probability in the 3-category case with zero means for all categories.

The probability measure for z induces a probability measure on y. The categorical probabilities are easily calculated as:

$$\Pr(y=j) = \frac{(1-F(-\mu_j))\prod_{k\neq j} F(-\mu_k)}{\sum_{j=1}^{J}(1-F(-\mu_j))\prod_{k\neq j} F(-\mu_k)}$$
$$= \psi_j(\mu_1,\ldots,\mu_J)$$

where $F(\cdot)$ is the CDF corresponding to f. Clearly, if $f=\phi$, then $F=\Phi(\cdot)$ is the standard normal CDF. While strictly speaking, the DO model with a full set of J category-specific intercepts and $J\times p$ regression coefficients is not identified, in the next section we suggest a simple identifying restriction that, critically, allows the parameters for the DO-model to be estimated via independent binary regressions, providing substantial computational advantages.

2.1 Interpretation of Regression Coefficients and Relationship to Multinomial Logit

In a regression context, DO models define an alternative link function in a GLM for unordered categorical/multinomial data, where:

$$y_i \sim \text{Categorical}(\boldsymbol{p}_i)$$
$$\boldsymbol{p}_i = \Big(\psi_1(x_i,\boldsymbol{\beta}_{[1:J]}),\psi_2(x_i,\boldsymbol{\beta}_{[1:J]}),\ldots,\psi_J(x_i,\boldsymbol{\beta}_{[1:J]})\Big)$$

The conditional class probabilities in the DO models have a useful correspondence with the multinomial logit. The class probabilities in the regression problem, $\Pr(y_i=j\mid x_i,\boldsymbol{\beta}_{[1:J]})=\psi_j(x_i,\boldsymbol{\beta}_{[1:J]})$, are given by:

$$\frac{(1-F(-x_i\beta_j))\prod_{l\neq j}F(-x_i\beta_l)}{\sum_{s=1}^{J}(1-F(-x_i\beta_s))\prod_{l\neq j}F(-x_i\beta_l)}$$

The ratio of probabilities for two classes j and k is therefore:

$$\frac{\frac{\left[(1-F(-x_i\beta_j))\prod_{l\neq j}F(-x_i\beta_l)\right]}{\left[\sum_{s=1}^{J}(1-F(-x_i\beta_s))\prod_{l\neq s}F(-x_i\beta_l)\right]}}{\frac{\left[(1-F(-x_i\beta_k))\prod_{l\neq k}F(-x_i\beta_l))\right]}{\left[\sum_{s=1}^{J}(1-F(-x_i\beta_s))\prod_{l\neq s}F(x_i\beta_l)\right]}}$$

The denominators and all but one of the terms in the products in the numerators cancel, leaving:

$$\frac{\psi_j(x_i,\boldsymbol{\beta}_{[1:J]})}{\psi_k(x_i,\boldsymbol{\beta}_{[1:J]})} = \frac{(1-F(-x_i\beta_j))F(-x_i\beta_k)}{(1-F(-x_i\beta_k))F(x_i\beta_j)}$$
$$= \frac{(1-F(-x_i\beta_j))\big/F(-x_i\beta_j)}{(1-F(-x_i\beta_k))\big/F(-x_i\beta_k)}$$

which is a function of only β_j and β_k. Recall that for the multinomial logit model, the ratio of class probabilities for classes j and k also depends only on β_j and β_k and is given by:

$$\log\left(\frac{\Pr(y_i=j)}{\Pr(y_i=k)}\right) = x_i\beta_j - x_i\beta_k$$

In DO-probit, $F(\cdot)=\Phi(\cdot)$, so the log relative class probability, $\log\left(\frac{\Pr(y_i=j)}{\Pr(y_i=k)}\right)$, is:

$$\log\left(\frac{\Phi(-x_i\beta_j)}{1-\Phi(-x_i\beta_j)}\right) - \log\left(\frac{\Phi(-x_i\beta_k)}{1-\Phi(-x_i\beta_k)}\right)$$

While not as convenient as the linear expression arising from the multinomial logit, this quantity is nonetheless easily calculated and provides a direct relationship between the coefficients in the multinomial logit and those in the DO-probit. This is a substantial advantage over the multinomial probit model, in which the class probabilities do not have a closed form.

The more interesting case arises when we choose a logistic distribution for the z_j's in the DO model, giving the DO-logistic model. Here, $F(t)=\frac{1}{1+e^{-t}}$ is the logistic CDF, and the log probability ratio for two categories is:

$$\log\left(\frac{p_i^{(j)}}{p_i^{(k)}}\right) = \log\left(\frac{\frac{1}{1+e^{x_i\beta_k}}}{1-\frac{1}{1+e^{x_i\beta_k}}}\right) - \log\left(\frac{\frac{1}{1+e^{x_i\beta_j}}}{1-\frac{1}{1+e^{x_i\beta_j}}}\right)$$
$$= \log\left(\frac{\frac{1}{1+e^{x_i\beta_k}}\frac{e^{x_i\beta_j}}{1+e^{x_i\beta_j}}}{\frac{1}{1+e^{x_i\beta_j}}\frac{e^{x_i\beta_k}}{1+e^{x_i\beta_k}}}\right) \tag{1}$$
$$= x_i\beta_j - x_i\beta_k$$

where $p_i^{(j)} = \Pr(y_i = j)$. This is identical to the log probability ratios in the multinomial logit. Evidently, the DO-logistic is an alternative form of the MNL. In the next section, we suggest an identifying restriction for the DO model that is much more convenient than the use of an arbitrary base category in MNL models and treats all of the categories identically. We also note that using the approach in O'Brien and Dunson [2004] one can easily do computation for the DO-logistic model by introducing a scale parameter for the latent variables that is mixed over a Gamma density. This immediately suggests an alternative Gibbs sampling algorithm for an MNL-like model that involves only one additional sampling step relative to the DO-probit.

2.2 Identification

Closer inspection of (1) reveals something rather striking about the representation of category probabilities in DO models. Consider (1) in an intercepts-only model:

$$\log\left(\frac{p_i^{(j)}}{p_i^{(k)}}\right) = \log\left(\frac{\frac{1}{1+e^{\mu_k}}\frac{e^{\mu_j}}{1+e^{\mu_j}}}{\frac{1}{1+e^{\mu_j}}\frac{e^{\mu_k}}{1+e^{\mu_k}}}\right)$$

As presented thus far, DO is an under-identified generalized linear model, and thus there will be multiple sets of parameters $\mu_1, \ldots, \mu_J$ that maximize the likelihood (inifinitely many, in fact). However, all of the solutions to the likelihood equations $\hat{\mu}_1, \ldots, \hat{\mu}_J$ must satisfy:

$$\log\left(\frac{\hat{p}_j}{\hat{p}_k}\right) = \hat{\mu}_j - \hat{\mu}_k$$

for any $j, k \in \{1, \ldots, J\}$, a simple consequence of MLE invariance. Yet (1) suggests that we might identify the model and obtain a very useful approach to estimation simply by recognizing the connection with binary logistic regression. Consider an intercept-only binary logistic regression with response $\tilde{y}_i^{(j)} = \mathbf{1}(y_i = j)$ and let $p_j^{(B)} = \Pr(\tilde{y}_j = 1)$. We have that:

$$\log\left(\frac{\hat{p}_j^{(B)}}{1-\hat{p}_j^{(B)}}\right) = \hat{\mu}_j^{(B)}$$

and so for two independent logistic regressions of $\tilde{y}_j$ and $\tilde{y}_k$ on an intercept, we get:

$$\log\left(\frac{\hat{p}_j^{(B)}(1-\hat{p}_k^{(B)})}{\hat{p}_k^{(B)}(1-\hat{p}_j^{(B)})}\right) = \hat{\mu}_j^{(B)} - \hat{\mu}_k^{(B)}$$

But since

$$\frac{\hat{p}_j^{(B)}}{1-\hat{p}_j^{(B)}} = \frac{\frac{e^{\hat{\mu}_j^{(B)}}}{1+e^{\hat{\mu}_j^{(B)}}}}{\frac{1}{1+e^{\hat{\mu}_j^{(B)}}}}$$

we have that:

$$\log\left(\frac{\frac{1}{1+e^{\hat{\mu}_k^{(B)}}}\frac{e^{\hat{\mu}_j^{(B)}}}{1+e^{\hat{\mu}_j^{(B)}}}}{\frac{1}{1+e^{\hat{\mu}_j^{(B)}}}\frac{e^{\hat{\mu}_k^{(B)}}}{1+e^{\hat{\mu}_k^{(B)}}}}\right) = \hat{\mu}_j^{(B)} - \hat{\mu}_k^{(B)}$$

Therefore the collection $\hat{\mu}_1^{(B)}, \ldots, \hat{\mu}_J^{(B)}$ - that is, the MLEs from independent binary regressions on $\tilde{y}_1, \ldots, \tilde{y}_J$ - is a valid solution to the likelihood equations for the DO model. Since the MLEs $\hat{\mu}_j^{(B)}$ for any j are unique, this solution is also unique, and is in fact identical to the solution that results from imposing the restriction $\sum_{j=1}^J \hat{p}_j = 1$ in the DO logistic model. A similar argument goes through for the general DO model. This further hints at the strategy we employ for Bayesian computation, which is identical to that used for J independent binary regressions. Of course, this was by construction; DO models were conceived of and designed expressly to allow for Bayesian computation using a set of J independent latent variables.

Note that while this is conceptually similar to the derivation of the multinomial logit model from independent binary regressions, in that case the response for each binary regression is defined by $\tilde{y}_i^{(j)} = 1$ if $y_i = j$ and $\tilde{y}_i^{(j)} = 0$ if $y_i = b$, where b is the base category, and thus the binary regressions are on a subset of the observations and defined against an arbitrary base category. For DO models, we perform binary regressions on all of the observations and do not require a base category. We also have the advantage of an additional interpretation for the estimated parameters as relating to the marginal probability of each category.

2.3 Relationship of DO-probit to Multinomial Probit

Both MNP and DO-probit link probability vectors to latent Gaussian variables by integrating multivariate normal densities over particular regions. For reasons discussed earlier, we consider only cases in which the latent Gaussian variables are conditionally independent given parameters. To simplify the exposition, we consider an intercepts-only MNP and a corresponding DO-probit model with J category-specific mean parameters and the identifying restriction presented in the previous section. Suppose we have an MNP with identity covariance and restrict the mean of the first utility to be zero. If we take $\mu_2 = -0.75$, the resulting category probabilities for the trivial two-category model are $(0.7115, 0.2885)$. The equivalent DO-probit model will have parameters $\mu_1 = \Phi^{-1}(0.7115) = 0.5578$ and $\mu_2 = \Phi^{-1}(0.2885) = -0.5578$. The two models give identical category probabilities, but they arrive at them differently. Figure 2 shows the contours of the bivariate normal distribution for the latent utilities in the MNP (left panel) and the truncated bivariate normal distribution in the DO-probit (right panel). The MNP integrates above and below the line $y = x$ (shown on the figure) to calculate probabilities, whereas DO-probit integrates over the second and fourth quadrants.

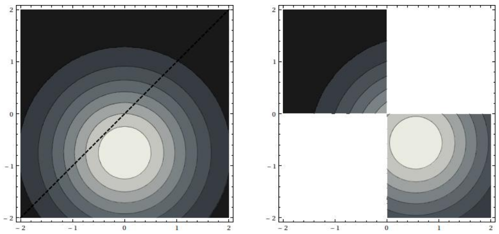

Figure 2: Contours and integration regions for equivalent two-category MNP and DO-probit models with category probabilities $(0.7115, 0.2885)$.

Of course, we can define a 1-1 function between MNPs with J categories, an identity covariance matrix, and $\mu_1 = 0$ and DO-probit with the marginal MLE restriction outlined in the previous section. The MNP probabilities for such a model are defined by the intractable integrals:

$$p_j = \Pr(y = j) = \int_{-\infty}^{\infty} \phi(z_j - \mu_j) \times \left[\int_{-\infty}^{z_j} \phi(z_J - \mu_J) dz_J \ldots \int_{-\infty}^{z_j} \phi(z_1) dz_1 \right] dz_j$$

for $j = 1, \ldots, J$. The integral has no analytic form. As such, marginalizing over the latent variables in MNP is computationally costly, and thus the latent variables are usually conditioned on in MCMC algorithms rather than integrating them out. This leads to high dependence and inefficient computation. In contrast, one can marginalize out the latent variables in DO-probit and obtain analytic expressions for the category probabilities, a significant benefit of our approach.

One can approximate the MNP integrals by quadrature or simulation, giving a set of probabilities $p_1, \ldots, p_J$. We can then find the equivalent parameters of a DO-probit by simply inverting the probabilities, i.e. $\mu_j = \Phi^{-1}(p_j)$ for all j. Note that while category probabilities are increasing in μ for both models, the MNP does not have the simple interpretation of the regression parameters as relating to the marginal category probabilities, and cannot be estimated from independent binary regressions as can the DO-probit.

3 Computation

Bayesian computation for the DO-probit is very straightforward. With a $N(0, cI)$, $c \in \mathbb{R}^+$ prior on the regression coefficients, the entire algorithm consists of Gibbs steps:

1. Sample $z_{i,y_i} \mid \beta_{y_i} \sim N_+(x_i^T \beta_{y_i}, 1)$ and $z_{i,k} \sim N_-(x_i^T \beta_k, 1)$ for $k \neq y_i$, where N_+ and N_- represents a normal distribution truncated below and above by zero, respectively.

2. Sample $\beta_k \mid z \sim N(\tilde{\mu}, \tilde{S})$ with $\tilde{S} = (\boldsymbol{x}^T\boldsymbol{x} + 1/cI)^{-1}$ and $\tilde{\mu} = \boldsymbol{x}^T \boldsymbol{z}_{\cdot,k}$ for $k \in 1, \ldots, K$.

Moreover, because of the latent Gaussian data augmentation scheme used to estimate the model, there are many alternatives to normal priors on regression coefficients. In classification problems with p covariates, the dimension of the parameter space is $J \times p$ (or $J \times (p-1)$ in the MNL and MNP), where J is the number of possible values of y. Also, the larger the J, the less information is provided by observing $y_i = j$. Thus, even with modest p and fairly large n, one should consider priors on regression coefficients with robust shrinkage properties, particularly when prediction is an important goal. The local-global family of shrinkage priors have the scale-mixture representation:

$$\beta_{jl} \sim N(0, \tau^2 \phi_{jl}^2)$$

where we have adapted the notation to the regression classification context with $j \in \{1, \ldots, J\}$ and $l \in \{1, \ldots, p\}$. Here, τ is a global shrinkage parameter and ϕ_{jl} are local shrinkage parameters corresponding to β_{jl}. There is a substantial literature on priors for τ and the ϕ_{jl}'s that favor aggressively shrinking most of the coefficients toward zero while retaining the sparse signals. For example, choosing independent $C_+(0,1)$ priors on τ and the ϕ_{jl}'s gives the horseshoe prior (Carvalho et al. [2010]), where $C_+(0,1)$ is a standard half-Cauchy distribution. Local-global priors can be employed in our model, with the only additional computational burden relative to the continuous response case being the imputation of the latent variables.

It is equally straightforward to apply other priors on regression parameters. Bayesian variable selection and model averaging can be performed by specifying point-mass mixture priors on β_{jl} and employing a stochastic search variable selection algorithm (see Hoeting et al. [1999] for an overview of these methods). A nonparametric prior on regression parameters can be obtained by specifying Gaussian process (GP) priors on the latent variables. The simplest approach is to specify J independent GP's corresponding to the J sets of latent variables (one for each possible value of y). In

MNP, this leads to difficult computation, and thus it is often necessary to use posterior approximations to make computation tractable (see Girolami and Rogers [2006]). The superior computational properties of DO-probit suggest that exact posterior computation using MCMC may be feasible with GP priors for problems of meaningful scale.

4 Simulation Studies

We conducted a series of simulation studies to assess the performance and properties of DO-probit and to compare it with MNP. We first simulated data from a MNP with $n \times p$ design matrix $\boldsymbol{x}$ with identity covariance matrix. We used $n = 2000$, $p = 2$, and $J = 5$ levels of the response. The $(J-1)$ category-specific intercepts were sampled from $N(0, .5)$ and the $(J-1)p$ coefficients were sampled from $N(0, 1)$. We then fit either MNP or DO-probit by Gibbs sampling. The chains were run for 10,000 iterations each, in every case starting from $\boldsymbol{\beta} = \mathbf{0}$. We repeated the simulation and subsequent fitting 10 times. Note that the parameter expansion algorithms designed to improve mixing are irrelevant for fitting this MNP model since we do not need to sample a covariance matrix. We choose category 1 as the base category for fitting.

Figure 3 shows histograms of lag-10, lag-25, and lag-100 autocorrelations for the Markov chains for $\boldsymbol{\beta}$ from MNP and DO-probit across the 10 simulations. Autocorrelations are much lower at all three lag lengths. This is a critical aspect of the performance of Bayesian stochastic algorithms. Higher autocorrelations require much longer run times to achieve the same effective sample sizes. In more complex models, the autocorrelations for MNP can be prohibitively high. In the following section, we show that lower autocorrelations are a feature of the DO-probit that persists in real applications with more complex priors on the coefficients.

Table 1 shows the mean in-sample misclassification rate (using the posterior mode as the prediction) from each of the ten simulations. The results are quite comparable for the two models, suggesting that DO-probit and MNP are exchangeable in regard to their performance as classifiers.

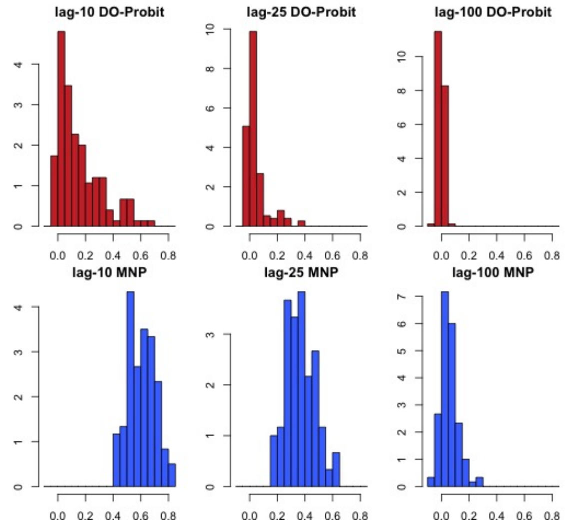

Figure 3: Distribution of Lag-10, -25, and -100 autocorrelations across all simulations and parameters for DO-probit (top) and MNP (bottom)

Simulation	MNP	DO-probit
1	0.30	0.30
2	0.36	0.37
3	0.51	0.51
4	0.32	0.32
5	0.42	0.42
6	0.49	0.49
7	0.49	0.48
8	0.36	0.37
9	0.42	0.42
10	0.45	0.45

Table 1: Comparison of misclassification rates for each of 10 simulated data sets.

5 Applications

We consider the glass identification data from the UCI machine learning website (Frank and Asuncion [2010]). There are 214 observations in the data and the response (class of the glass sample) is unordered categorical with seven possible levels, of which six are observed. There are nine continuous covariates. Thus a DO-probit classifier has sixty parameters (6 × 9 regression coefficients and 6 intercepts). Because n is not large relative to p in this case, we use a Horseshoe shrinkage prior on the β's (see Carvalho et al. [2010] and the discussion in section 3).

A boxplot of posterior samples for β is shown in the left panel of figure 4, and a corresponding plot for the MNP with identity covariance matrix is presented in the right panel. Red colored boxes indicate parameters that are considered nonzero on the basis of the criteria suggested in Carvalho et al. (let $\kappa_{jk} = 1/(1-\tau^2\phi^2_{jk})$, and consider β_{jk} nonzero if $\hat{\kappa}_{jk}$, the posterior mean of κ_{jk}, is > 0.5). Recent work has shown that this inclusion criterion has optimal properties under a 0-1 loss function (Datta and Ghosh [2012]). Note that of the 60 coefficients in the DO-probit, 56 are effectively shrunk to zero, whereas of the 50 coefficients in the MNP, 39 of them are effectively shrunk to zero. In addition, of the 9 covariates, 6 of them have all 6 coefficients shrunk to zero in the DO-probit, which is the equivalent of excluding that covariate entirely. In the MNP, 5 of these 6 covariates also have all 5 coefficients effectively shrunk to zero.

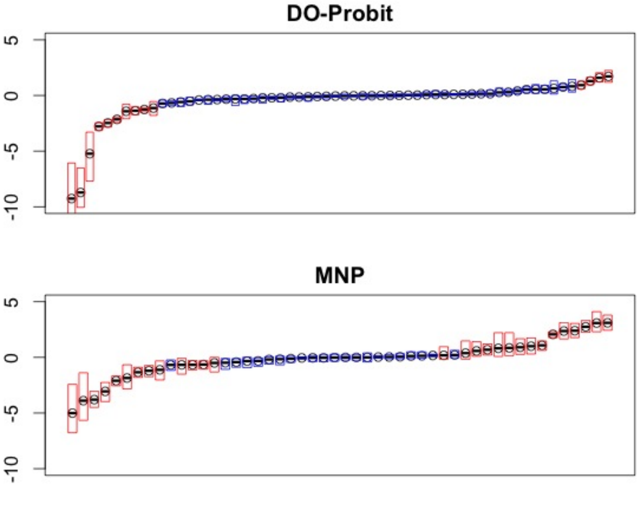

Figure 4: Boxplots of posterior samples for $\boldsymbol{\beta}_{1:p}$ for MNP and DO-probit. Red boxes indicate that the coefficient is considered nonzero by the criterion described above.

Figure 5 shows boxplots of the lag 1 through lag 50 posterior autocorrelations in the Markov chains for $\boldsymbol{\beta}$ from the DO-probit and MNP models. The boxes show the interquartile range and the black dots the medians. This confirms that much lower autocorrelations in the DO-probit persist in real applications with larger number of parameters and more complex hierarchical modeling structures. Note that the residual autocorrelation for DO-probit is due mainly to the Metropolis-Hastings steps used to sample the scale parameters for the horseshoe prior, and that parameter expansion or slice sampling could improve the mixing, see Scott [2010]. Table 2 shows quantiles of the effective sample size for DO-probit and MNP estimated on the glass data. The median effective sample size for DO-probit is 3.85 times that for MNP, even in this relatively simple modeling context. The run time per iteration for DO-probit was 1.06 times that of MNP, a difference that is entirely due to the larger number of parameters for DO-probit.

twenty

	10%	25%	50%	75%	90%
DO probit	101	250	1204	3176	6085
MNP	78	128	216	464	835

Table 2: Quantiles of effective sample sizes for β parameters from DO-probit and MNP using 15,000 MCMC iterations with 1000 iteration burn-in.

We assessed out-of-sample prediction on the dataset by randomly holding out 10 percent of observations and estimating the model on the remaining 90 percent, a process that was repeated twenty times. We formed out of sample predictions by taking the posterior mode of the predicted class for each observation in the test set. The full set of 60 (50) coefficients were used for prediction in the DO-probit (MNP) models, which is standard practice for shrinkage priors. Table 3 shows the best, median, and worst misclassification rates for the two models. The results show that the two classifiers are equivalent. Note that for over half of the test datasets, the misclassification rates for the two models were identical.

	best	median	worst
Probit	0.22	0.38	0.58
DO-Probit	0.22	0.37	0.60

Table 3: Summary of misclassification rates for 20 random holdouts of glass identification data.

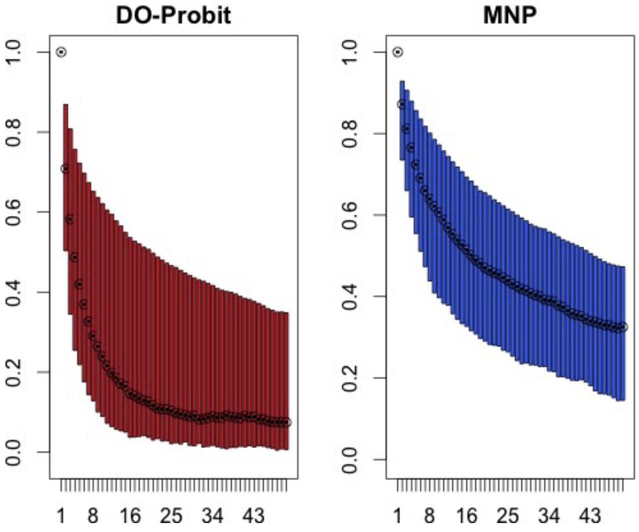

Figure 5: Boxplots of lag-1 through lag-50 autocorrelations for MCMC samples of $\boldsymbol{\beta}_{1:p}$ from MNP and DO-probit.

6 Discussion

The DO-probit model provides an attractive alternative to the multinomial logit and multinomial probit models that overcomes the major limitations of each while retaining many of their attractive properties. Like MNP, DO-probit admits a Gaussian latent variable representation, allowing for simple conjugate updates for regression parameters and application of numerous methods designed for multivariate Gaussian data, a feature that MNL lacks. However, our model does not suffer from the poor mixing and high autocorrelations in MCMC samples that make MNP practically infeasible for use in high-dimensional applications. Unlike MNP and MNL, our model does not require a base category for identification. However, class probabilities for DO-probit marginal of latent variables are easily calculated, and relative class probabilities are functions only of the regression parameters corresponding to the compared classes, an attractive feature shared with MNL.

The DO-probit link provides a number of possible avenues for future work. The Gaussian latent variable representation of the model allows for extension to nonparametric regression via specifying a Gaussian process prior on the latent variables. Another interesting possibility would be to explore a novel class of discrete choice models by allowing dependence between latent variables, the analogue of a non-diagonal covariance matrix in the MNP. Multivariate unordered categorical response data with covariates is a particularly challenging context in which to develop computationally tractable Bayesian methods. One could potentially allow for dependence in multivariate cases by specifying a prior on structured covariance matrices for the latent Gaussian variables. Another possible application would be to time series of polychotomous variables by specifying AR models on the latent variables. In complex modeling situations such as these, fully Bayesian estimation using DO-probit may be straightforward, whereas the computational challenges of the MNP make these extensions, while theoretically possible, practically infeasible.

7 Acknowledgments

This work was supported by Award Number R01ES017436 from the National Institute of Environmental Health Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Environmental Health Sciences or the National Institutes of Health. This work has been partially supported by DTRA CNIMS Contract HDTRA1-11-D-0016-0001 and Virginia Tech internal funds.

References

J.H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679, 1993.

L.F. Burgette and P.R. Hahn. Symmetric bayesian multinomial probit models. *Duke University Statistical Science Technical Report*, pages 1–20, 2010.

L.F. Burgette and E.V. Nordheim. The trace restriction: An alternative identification strategy for the bayesian multinomial probit model. *Journal of Business & Economic Statistics*, 30(3):404–410, 2012.

C.M. Carvalho, N.G. Polson, and J.G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

J. Datta and J.K. Ghosh. Asymptotic properties of bayes risk for the horseshoe prior. *Bayesian Analysis*, 7(4):771–792, 2012.

A. Frank and A. Asuncion. UCI machine learning repository, 2010. URL http://archive.ics.uci.edu/ml.

M. Girolami and S. Rogers. Variational bayesian multinomial probit regression with gaussian process priors. *Neural Computation*, 18(8):1790–1817, 2006.

J.A. Hausman and D.A. Wise. A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica: Journal of the Econometric Society*, pages 403–426, 1978.

J.A. Hoeting, D. Madigan, A.E. Raftery, and C.T. Volinsky. Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401, 1999.

Chris C. Holmes and Leonhard Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesina Analysis*, 1(1):145–168, 2006.

K. Imai and D.A. Van Dyk. A bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of Econometrics*, 124(2):311–334, 2005.

R. McCulloch and P.E. Rossi. An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, 64(1):207–240, 1994.

R.E. McCulloch, N.G. Polson, and P.E. Rossi. A bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics*, 99(1):173–193, 2000.

S.M. O'Brien and D.B. Dunson. Bayesian multivariate logistic regression. *Biometrics*, 60(3):739–746, 2004.

Ari Pakman and Liam Paninski. Exact hamiltonian monte carlo for truncated multivariate gaussians. *arXiv preprint arXiv:1208.4118*, 2012.

Nick G. Polson, James G. Scott, and Jesse Windle. Bayesian inference for logistic models using Polya-Gamma latent variables. *ArXiv e-prints*, 2012.

J.G. Scott. Parameter expansion in local-shrinkage models. *arXiv preprint arXiv:1010.5265*, 2010.

K.E. Train. *Discrete choice methods with simulation.* Cambridge university press, 2003.

X. Zhang, W.J. Boscardin, and T.R. Belin. Sampling correlation matrices in bayesian models with correlated latent variables. *Journal of Computational and Graphical Statistics*, 15(4):880–896, 2006.

X. Zhang, W.J. Boscardin, and T.R. Belin. Bayesian analysis of multivariate nominal measures using multivariate multinomial probit models. *Computational statistics & data analysis*, 52(7):3697–3708, 2008.