

# Sparse Factor-Analytic Probit Models

P. RICHARD HAHN

*Department of Statistical Science, Duke University  
Durham, North Carolina 27708-0251, U.S.A.  
hahn@stat.duke.edu*

JAMES G. SCOTT

*McCombs School of Business, University of Texas at Austin  
Austin, Texas 78712, U.S.A.  
james.scott@mcombs.utexas.edu*

CARLOS M. CARVALHO

*Booth School of Business, The University of Chicago  
Chicago, Illinois 60637, U.S.A.  
carlos.carvalho@chicagobooth.edu*

November 2009

## ABSTRACT

We describe a class of sparse factor-analytic probit models for multivariate binomial and multinomial data. These models provide a parsimonious lower-dimensional representation of multivariate categorical data by imposing structure upon the covariance matrix of a latent normal parameter. This confers a number of advantages over traditional models. First, the factor-probit model can be used as a powerful exploratory tool for investigating underlying structure in categorical data. Second, it can be used to create well-behaved shrinkage estimators that make the multivariate probit model viable even when the number of variables is large relative to the number of observations. The use of sparsity priors contributes additional regularization, and also provides a natural probabilistic framework for investigating the number of factors driving the observed covariation. Finally, the factor model offers significant computational gains, as it circumvents the need to sample from a high-dimensional truncated multivariate normal distribution. After describing the model, we study its performance both on simulated data, and on a data set regarding consumer preferences in Scotch whisky that has been previously analyzed in the literature. We then turn to our motivating example: the analysis of partisanship patterns in sixty years of roll-call votes from the United States Senate. The factor-loadings matrix that emerges from this analysis corresponds to plausible political forces, and the manner in which this matrix changes over time

raises interesting questions regarding presidential election cycles. Moreover, the factor scores themselves provide a novel way of ranking senators in terms of the partisanship of their voting patterns.

*Some key words:* covariance estimation; factor models; multivariate probit models; political partisanship

## 1 INTRODUCTION

Correlated categorical data are ubiquitous both in the natural and social sciences. Yet even at their simplest, where outcomes are binary, such data sets pose significant modeling challenges. Estimators are ill-behaved; priors, hard to elicit. In recent years these challenges have been further complicated by the need to analyze models in which the number of variables ( $p$ ) can be as large as, or larger than, the number of available observations ( $n$ ).

In this paper, we extend the multivariate probit model (Chib and Greenberg, 1998) to encompass a sparse factor-analytic approach for inference about the underlying correlation structure of binary and ordinal data. Two main goals motivate our work:

- (i) **Interpretability in exploratory data analysis.** Sparse factor models provide a very natural and intuitive representation of latent structure in multivariate data. These models are especially useful when researchers are analyzing data without pre-set theories in hand, or with only loose ideas about relationships among the variables. This is because factors often have a useful subject-specific interpretation, and can be used to generate further hypotheses about the forces at play in the data. Moreover, our Bayesian framework allows for a more measured assessment of these forces, since uncertainty about all unknowns can be quantified using the full posterior distribution.
- (ii) **Regularization.** We improve estimator variance by drastically reducing the number of parameters that must be fit, and we do so with little compromise in flexibility. The key step is the imposition of structure on the covariance matrix, which creates estimators that are stable even in large problems—an advantage that can be decisive when the number of variables  $p$  is very large relative to the sample size  $n$ . It is well known that regularized estimators in general, and highly structured models in particular, can provide significant improvements over standard estimators in reconstructing large covariance matrices (Rajaratnam et al., 2008). This is highly relevant in cases where the data itself is only partially observed, as it is in a multivariate probit model. Sparse models provide

still further help here. Indeed, we will argue (via simulation) that sparse factor models can result in a highly favorable bias–variance trade-off, even when there is no particular reason to suspect an underlying factor structure.

Similar approaches have been proposed in a variety of different scientific contexts. Our work follows that of Chib and Greenberg (1998), Edwards and Allenby (2003), McCulloch et al. (2000), and to a lesser extent that of Elrod and Keane (1995). Our approach differs from the above work, however, in the imposition of highly structured, sparse Bayesian models for the latent covariance matrix. The generic pros and cons of being Bayesian are outside the scope of this paper; we will instead focus on the empirical advantages of conducting exploratory data analysis using our proposed methodology.

We also draw attention to three secondary, though still significant, advantages of our approach.

- (iii) **Computational efficiency:** Posterior sampling for a standard multivariate probit model require repeated draws from a multivariate truncated normal distribution whose parameters change at every step. This represents a significant bottleneck as  $p$  grows. Imposing a factor structure, however, reduces the multivariate truncation problem to a series of independent univariate truncations, which are significantly easier to handle, more scalable, and less prone to autocorrelation.
- (iv) **Missing data** can be imputed very straightforwardly.
- (v) **Modularity:** Sparse factor models can easily be embedded inside more complex hierarchical models—for example, those involving a spatial or temporal component. The motivating example in Section 6 is suggestive here.

The paper builds up to an analysis of our motivating example: a study of ideological and partisan patterns in sixty years of close roll-call votes from the United States Senate. Our results show an upward trend in partisan voting patterns over the last several decades, superimposed upon the usual ebb and flow of presidential election cycles. This is consistent with other analyses by political scientists using very different statistical methods. We also show how, as a byproduct of the analysis, individual senators can be ranked in terms of their partisan tendencies. This is both a novel feature of the model and a useful “sanity check” on our results: the analysis should, and does, tend to flag the majority whip as the among most partisan voters in any given Senate term.

Our goal in this analysis is not to construct a realistic model for how senators cast their votes. Such a model would likely go far beyond mere party membership to incorporate features such as geography, incumbency, committee membership, and much more besides. The sparse factor–probit model does not do this, and does not aspire to. If anything, we conceive of the method as a hypothesis-generating tool analogous to principal components, something to be applied before the hard work of formal model-building ever begins. It is in this exploratory capacity that we analyze the Senate roll-call data.

The paper begins, in Sections 2 and 3, by presenting the basic modeling ideas in sparse factor–probit analysis. Section 4 gives a detailed account of the sampling algorithm for posterior inference. Section 5 explores the practical importance of regularization in both a comprehensive simulation study as well as in a benchmark data set on Scotch preferences that has been previously analyzed in the literature. Finally, Section 6 presents the results of our analysis of Senate voting patterns.

## 2 FACTOR-ANALYTIC PROBIT MODELS

### 2.1 Multivariate and multinomial probit models

Suppose we observe  $Y = (\mathbf{y}_1 \dots \mathbf{y}_n)^t$ , where each  $\mathbf{y}_i = (y_{i,1}, \dots, y_{i,p})^t$  represents  $p$  correlated binary observations on a single subject. The multivariate probit model (Chib and Greenberg, 1998) characterizes  $\mathbf{y}_i$  in terms of an unobserved continuous quantity  $\mathbf{z}_i$ , related to the data by the probit link function:

$$\begin{aligned} y_{ij} &\sim \text{BER}(r_{ij}) \\ r_{ij} &= \Pr(z_{ij} > 0) \\ \mathbf{z}_i &\sim N(\boldsymbol{\alpha}, \boldsymbol{\Sigma}). \end{aligned} \tag{1}$$

The trick of augmenting the parameter space with  $\mathbf{z}_i$  makes computation simple, since these variables can be treated as “observed” Gaussian data in a way that preserves the correct marginal distribution for  $\mathbf{y}_i$ .

A related model arises when we instead observe  $y_i \in \{0, 1, \dots, p - 1\}$  indicating the choice made by subject  $i$  among  $p$  mutually exclusive possibilities. The same link function can be used in a multinomial probit model to relate the categorical variable  $y$  to an underlying continuous quantity  $z$  (which is now of dimension  $p - 1$ ). Following McCulloch and Rossi (1994) and McCulloch

et al. (2000),

$$y_i = \begin{cases} 0 & \text{if } \max \mathbf{z}_i < 0 \\ j & \text{if } \max \mathbf{z}_i = z_{ij} > 0 \end{cases} \quad (2)$$

$$\mathbf{z}_i \sim N(\boldsymbol{\alpha}, \Sigma).$$

In principle these two models could be nested to arrive at a multivariate multinomial model similar to Zhang et al. (2008), but such an approach is not further explored here.

Observe that, for identifiability reasons,  $\Sigma$  must be a correlation matrix. If  $\mathbf{z}_i \sim N(\beta, \mathbf{T})$  is a parameterization in terms of covariances rather than correlations, then defining  $D = \text{diag}(\tau_{11}^{-1/2}, \dots, \tau_{pp}^{-1/2})$ ,  $\Sigma = D\mathbf{T}D$ , and  $\alpha = D\beta$  gives the same Bernoulli probabilities:  $\Pr(y_{ij} = 1 \mid \alpha, \Sigma) = \Pr(y_{ij} = 1 \mid \beta, \mathbf{T})$ . We later make use of this fact in post-processing our MCMC output to yield estimates of model parameters.

Without loss of generality, we denote the mean of  $\mathbf{z}_i$  by  $\boldsymbol{\alpha}$ , with the understanding that this may be a linear predictor involving covariates  $X$ .

## 2.2 Factor models for latent variables

The difficulty with both of these models lies in estimating the set of  $p(p-1)/2$  pairwise correlations for the latent  $\mathbf{z}$  vector. In a typical covariance-estimation problem, one would observe the actual values of  $\mathbf{z}_i$ , with information about  $\Sigma$  coming from the cross-product matrix  $Z'Z$ . Even here, standard estimators can be notoriously unstable when  $p$  is fairly large compared to  $n$ , and in particular can yield a distorted picture of the eigenstructure of  $\Sigma$ ; see the discussion in, for example, Sun and Berger (2006).

These troubles are compounded in each of the two probit models. For binary data, only the sign of each  $z_{ij}$  is observed, and not its actual value; for multinomial data, one knows only which element of  $\mathbf{z}_i$  is the largest, or that all elements are negative. With less information about each  $\mathbf{z}_i$  than in the standard case, estimating  $\Sigma$  becomes even more difficult.

This motivates a simplifying factor structure, which improves estimability by constraining and regularizing  $\Sigma$ . Upon positing some fixed number  $k$  of factors  $\mathbf{f}$  along with a  $p \times k$  matrix of factor loadings  $\mathbf{B}$ , we can write the latent variables for observation  $i$  as:

$$\begin{aligned} \mathbf{z}_i &= \boldsymbol{\alpha} + \mathbf{B}\mathbf{f}_i + \boldsymbol{\nu}_i \\ \boldsymbol{\nu}_i &\sim N(0, \Psi) \\ \mathbf{f}_i &\sim N(0, \mathbf{I}) \end{aligned} \quad (3)$$

from which it follows that  $\Sigma = \mathbf{B}\mathbf{B}' + \Psi$ . Note that  $\Psi$  is forced to be diagonal—either  $\Psi = \sigma^2 I$  for a common scale among all variables, or  $\Psi = \text{diag}(\psi_1, \dots, \psi_p)$  more generally. Write  $B'_j$  and  $\mathbf{b}_k$  for the vectors representing row  $j$  and column  $k$  of the factor-loadings matrix, respectively.

A factor model says that all correlation among the elements of the high-dimensional latent vector  $\mathbf{z}$  can be explained by mutual dependence upon a lower-dimensional vector  $\mathbf{f}$ , whose components vary independently *a priori*. This dependence structure is encoded in  $\mathbf{B}$ , with idiosyncratic variation captured by  $\nu$ , a vector of random noise whose elements are called uniquenesses.

Even if we do not wish to interpret the factors as having any substantive meaning, we have reduced the number of parameters in  $\Sigma$  that must be estimated to  $k(p+1)$ , rather than  $p(p-1)/2$  (which gives a natural upper bound for  $k$ ). Since  $k$  is usually much less than  $p$ , this structural regularization can be quite helpful in high-dimensional problems.

Factor models originally date to Spearman (1904); a more modern discussion can be found in Press (1982). Bayesian factor models for continuous data owe their development to many authors, including Geweke and Zhou (1996) and Aguilar and West (2000).

A factor model must be further constrained for the identity  $\Sigma = \mathbf{B}\mathbf{B}' + \Psi$  to have a unique solution, i.e. for  $\mathbf{B}$  to be identifiable. These identification issues, discussed extensively in Aguilar (1998) and Frühwirth-Schnatter and Lopes (2009), are logically distinct from the fact that  $\Sigma$  must be a correlation matrix in the probit setting. Our chief concern is that the model in (3) is invariant under an orthogonal rotation of the factors, since  $\mathbf{B}^* = \mathbf{B}\Gamma$  and  $\mathbf{f}_i^* = \Gamma'\mathbf{f}_i$  give the same model for any orthogonal matrix  $\Gamma$ .

Two traditional solutions to this problem involve forcing  $\mathbf{B}$  to be orthogonal or forcing  $\mathbf{B}^t\Psi\mathbf{B}$  to be diagonal. We instead follow Geweke and Zhou (1996) and constrain  $\mathbf{B}$  to be zero for upper-triangular entries  $\{b_{jk} : k > j\}$  and positive along the diagonal  $\{b_{jj} > 0\}$ . This guarantees that  $\mathbf{B}$  is the unique matrix for which  $\mathbf{B}\mathbf{B}^t$  is positive-definite, and makes the choice of the  $k$  leading variables – referred to as the founders of the factors – a key modeling decision. This approach permits an analyst to easily associate factors with known qualitative attributes of observables.

Priors over  $\alpha$ ,  $\mathbf{B}$ , and  $\Psi$  complete a Bayesian specification, which we detail in subsequent sections.

### 3 SPARSITY PRIORS

West (2003) and Carvalho et al. (2008) develop sparse factor models for continuous data in the context of gene-expression studies. These models assume that each latent factor will be associated with only a small number of ob-

served variables, yielding a more parsimonious covariance structure with even stronger regularization properties. (As we shall see, the development of a similar methodology for the probit case must confront some unique challenges due to the constraint that  $\Sigma$  must be a correlation matrix.)

We now describe a novel sparse Bayesian factor-analytic probit model, where some of the unconstrained elements in the factor-loadings matrix  $\mathbf{B}$  can be identically 0.

In a sparse factor model, the pattern of non-zero elements in  $\mathbf{B}$  is unknown and must be estimated from the data. Previous authors have assumed that the prior for the loadings matrix  $\mathbf{B}$  takes the following form:

$$(b_{jk} \mid v_k, q_k) \sim q_k \cdot \mathcal{N}(0, v_k) + (1 - q_k)\delta_0 \quad (4)$$

$$v_k \sim \text{IG}(a_v/2, b_v/2) \quad (5)$$

$$q_k \sim \text{BE}(1, 1) \quad (6)$$

where there is a different variance component  $v_k$  and prior inclusion probability  $q_k$  associated with each column of the loadings matrix. By treating the prior inclusion probabilities as model parameters to be estimated from the data, this model induces a strong multiplicity-correction effect, thereby solving the implicit multiple-testing problem of simultaneously deciding whether to include each of hundreds, or even thousands, of possible nonzero entries in  $\mathbf{B}$  (Scott and Berger, 2006).

At one extreme, entire columns of the loadings matrix can be set to exactly zero with probability near one, effectively selecting the number of necessary factors automatically. In this respect, the sparse factor model is an alternative to approaches such as that presented in Lopes and West (2004).

We modify this now-standard model by grouping the variance components by *row* rather than by column:

$$(b_{jk} \mid v_j, q_k) \sim q_k \cdot \mathcal{N}(0, v_j) + (1 - q_k)\delta_0 \quad (7)$$

$$v_j \sim \text{IG}(c/2, d/2) \quad (8)$$

$$q_k \sim \text{BE}(1, 1). \quad (9)$$

This change reflects the fact that while the sparsity (that is, the fraction of exactly-zero factor loadings) is naturally a factor-specific property, the variability of the factor loadings should instead be a row-specific property. Latent variables, after all, have no intrinsic scale, meaning that scale differences across the dimensions of our observation vector can only arise due to scale differences in the rows of our loadings matrix.

Consider, as a simple example, a tuple recording daily high and low tem-

peratures for a sample of cities, where the daily high is recorded in Celsius and the daily low is recorded in Fahrenheit. Here we might interpret a latent factor as reflecting some geographical attributes of the cities' locations. Yet because  $\mathbf{f}_i$  has an arbitrary, fixed scale, the extra variability in Fahrenheit temperatures must be reflected in our loadings matrix in a row-wise manner.

Formally, note that a column-wise change in scale can be interpreted simply as a change in the variability of factor scores: if  $\mathbf{b}_k \sim N(0, s_k v_k \mathbf{I})$  for each  $k$ , then an equivalent model is given by  $\mathbf{b}_k \sim N(0, v_k \mathbf{I})$  for each  $k$  with  $\mathbf{f}_i \sim N(0, S)$  where  $S$  is the diagonal  $K$ -by- $K$  matrix of scale components  $s_k$ . Since the scale of latent factor scores is arbitrary, the data cannot inform us about a scale parameter shared by column.

Though we do not further address this point here, this modeling decision becomes crucial for appropriately applying sparsity priors in the continuous setting, where the probability of an element of  $\mathbf{B}$  being exactly zero should depend on the ratio  $\psi_j/v_j$ .

For estimation of a probit model, where the  $\mathbf{z}_i$  themselves are latent we can safely fix  $\Psi \equiv \mathbf{I}$ . Hyperparameters  $c$  and  $d$  can then be chosen to match this scale appropriately; we use  $c = d = 1$ . Inferences concerning sparsity flow from the relative size of the elements  $b_{jk}$  on this unit scale.

#### 4 MODEL FITTING

We employ a Gibbs sampler to draw correlated samples from the joint posterior distribution of all parameters (Gelfand and Smith, 1990; Geman and Geman, 1984). In what follows we describe how to sample from each of the full conditional distributions.

We sample the nonidentified parameters and post-process the output to yield estimates of quantities that are identified. This post processing simply amounts to rescaling  $\mathbf{B}$  so that  $\Sigma$  is a correlation matrix, and then similarly scaling  $\alpha$ . This is explained further in McCulloch et al. (2000).

Sampling proceeds as follows.

1. Draw the latent observation matrix  $Z = (z_{ij})$  by drawing each  $z_{ij} \sim N(\alpha_j + B'_j f_i, 1)$  truncated above at 0 if  $y_{ij} = 0$  and below at 0 if  $y_{ij} = 1$ . Here  $B'_j$  is the row of the factor loadings matrix corresponding to component  $j$  of the random vector  $\mathbf{z}$ .
2. Sample the mean vector  $\alpha$ ; this standard step will be context dependent, and is not considered here.
3. Sample the vectors of factor scores independently as

$$(\mathbf{f}_i \mid \mathbf{z}_i) \sim N(\mathbf{B}'[\mathbf{B}\mathbf{B}' + \mathbf{I}]^{-1}(\mathbf{z}_i - \alpha), \mathbf{I} - \mathbf{B}'[\mathbf{B}\mathbf{B}' + \mathbf{I}]^{-1}\mathbf{B}).$$

4. To sample the unconstrained elements of  $\mathbf{B}$ , define  $z_{ij}^* = z_{ij} - \alpha - \sum_{l=1, l \neq k}^m B_{j,l} f_{k,i}$ . Then sample

$$b_{jk} \sim (1 - \hat{q}_{jk})\delta_0 + \hat{q}_{jk} \cdot \text{N}(\hat{b}_{jk}, \hat{v}_{jk}),$$

where

$$\begin{aligned} \hat{v}_{jk} &= \left( \sum_{i=1}^n f_{k,i}^2 + v_j^{-1} \right)^{-1} \\ \hat{b}_{jk} &= \hat{v}_{jk} \left( \sum_{i=1}^n f_{k,i} z_{ij}^* \right) \\ \frac{\hat{q}_{jk}}{1 - \hat{q}_{jk}} &= \frac{\text{N}(0 \mid 0, v_j)}{\text{N}(0 \mid \hat{b}_{jk}, \hat{v}_{jk})} \frac{q_k}{1 - q_k}. \end{aligned}$$

5. Let  $s_j$  be the number of the elements in  $B'_j$  currently not set to zero. Using this, draw

$$v_j \sim \text{IG}\{(1 + s_j)/2, (1 + B_j B'_j)/2\}.$$

6. Finally, draw  $q_k \sim \text{BE}(1 + s_k, 1 + \tilde{s}_k - s_k)$ , where  $s_k$  is as in the previous step and  $\tilde{s}_k$  is the maximum possible number of non-zero elements for column  $k$ .

Notice the substantial computational savings implicit in Step 1. In our method, it is not necessary to draw from a high-dimensional truncated multivariate normal distribution, as in the Wishart model of Chib and Greenberg (1998). All the dependence among the elements of  $\mathbf{z}_i$  is encoded in  $B$ . Hence the truncations arising from the observed data  $y_{ij}$  can be handled independently, given  $B$ .

## 5 PERFORMANCE ON BENCHMARK EXAMPLES

### 5.1 Simulated data

The following simulation study illustrates the manner in which modeling the latent covariance as  $\Sigma = \mathbf{B}\mathbf{B}' + \Psi$  for  $\mathbf{B}$  can produce a regularized estimator with a favorable bias–variance tradeoff. We compare three models of the covariance structure: the diffuse inverse-Wishart model, centered at the identity matrix; a  $k = 6$  factor model; and a  $k = 6$  sparse factor model. We examined the performance of each of these models on data corresponding to four distinct regimes:

- The true  $\Sigma$  has a factor structure with three factors.
- The true  $\Sigma$  has a factor structure with ten factors.
- The true  $\Sigma$  has no factor structure (equivalently, the number of factors is equal to number of dimensions).
- The true  $\Sigma$  is the identity matrix.

For a given  $\Sigma$  and mean vector  $\alpha$ , the data was constructed as:

$$Z \sim \mathbf{N}(\alpha, \mathbf{R}) \quad (10)$$

$$X = [Z > 0], \quad (11)$$

where  $\mathbf{R} = D^{-\frac{1}{2}}\Sigma D^{-\frac{1}{2}}$  for  $D = \text{diag}(\Sigma)$ . In all regimes  $\alpha$  was drawn as  $\mathbf{N}(0, 0.2\mathbf{I})$ .

For all simulations, we used  $n = 50$  observations. An estimated correlation matrix  $\tilde{\mathbf{R}}$  was obtained for  $p = 20$  and  $p = 100$ , and the mean Frobenius and Stein losses were computed over 100 replications. Recall the Frobenius and Stein losses are given, respectively, as:

$$L_F(\tilde{\mathbf{R}}, \mathbf{R}) = \text{tr}\{(\tilde{\mathbf{R}} - \mathbf{R})^2\} \quad (12)$$

$$L_S(\tilde{\mathbf{R}}, \mathbf{R}) = \text{tr}(\tilde{\mathbf{R}}\mathbf{R}^{-1}) - \log \det(\tilde{\mathbf{R}}\mathbf{R}^{-1}) - p, . \quad (13)$$

The study includes cases where  $p > n$  and  $n > p$ ; cases where the factor model has both too few and too many factors compared to the truth; and cases where there is no factor structure at all. This goal is to benchmark the performance of the three models across a range plausible real-data scenarios. Results are reported in Table 1; the top half of the table shows results when  $n > p$ , and the bottom half shows results when  $p > n$ .

The differences between the various models when  $n > p$  are modest, but the factor models still dominate the inverse-Wishart model. The sparse model performs better in the highly sparse settings (identity and  $k = 3$ ), while the traditional factor model performs better in the less-sparse settings ( $k = 10$  and  $k = 20$ ).

When  $p = 100$  and  $n = 50$ , however, the benefit of the factor model over the inverse-Wishart model becomes more stark. For instance, the sparse model on the identity matrix gives outstanding performance. The results are especially interesting in the  $(p = 100, k = 10)$  and  $(p = 100, k = 100)$  cases. Here, six factors are clearly insufficient to reconstruct  $\Sigma$ ; the true covariance matrix is not even in the support of the prior, which means the resulting estimator must inevitably be highly biased. Yet the factor model still outperforms the

Table 1: Mean Stein and Frobenius losses suffered in reconstructing the true correlation matrix  $\mathbf{R}$  in various configurations.

Loss function	True model	Fitted Model		
		Wishart	6-Factor	Sparse 6-Factor
Stein	$p = 20, k = 3$	74.7	13.9	9.9
	$p = 20, k = 10$	91.0	24.0	29.7
	$p = 20, k = 20$	53.8	12.1	18.0
	$p = 20, \text{identity}$	3.6	2.9	0.4
Frobenius	$p = 20, k = 3$	89.6	14.6	12.9
	$p = 20, k = 10$	40.3	12.3	14.0
	$p = 20, k = 20$	37.6	14.6	13.0
	$p = 20, \text{identity}$	8.1	6.7	0.89
Stein	$p = 100, k = 3$	503.1	136.7	43.4
	$p = 100, k = 10$	1323.2	357.4	394.2
	$p = 100, k = 100$	827.8	454.2	667.3
	$p = 100, \text{identity}$	28.3	26.2	1.0
Frobenius	$p = 100, k = 3$	2573.8	430.5	234.0
	$p = 100, k = 10$	1143.1	403.8	410.0
	$p = 100, k = 100$	305.7	275.7	160.9
	$p = 100, \text{identity}$	94.6	136.3	2.1

Wishart—in some cases drastically—due to the dramatically lower sampling variance of the posterior mean.

In short, it would appear that there are strong reasons to use the factor model in large  $p$ , small  $n$  settings, even if we do not believe that such a factor structure accurately describes  $\Sigma$ .

## 5.2 Data on preferences in Scotch whisky

### 5.2.1 Exploratory Analysis

In the following example, we use the Scotch-preference data previously analyzed by McCulloch and Rossi (1994) and Edwards and Allenby (2003) to benchmark the factor-probit model, draw attention to its data-exploration properties, and highlight the practical relevance of regularization.

This data set comes from the Simmons Study of Media and Markets (1997). It consists of  $n = 2,218$  binary vectors indicating which of 21 Scotch whisky brands individual  $i$  had purchased in the preceding year. In fitting a factor model, we hope to recover patterns consistent with the notion that preferences are shaped by a relatively small number of market forces.

We use the study presented in Edwards and Allenby (2003) as a benchmark for our analysis. In that paper, an unconstrained multivariate probit model was used under the assumption of an inverse-Wishart prior for  $\Sigma$ ; all explo-

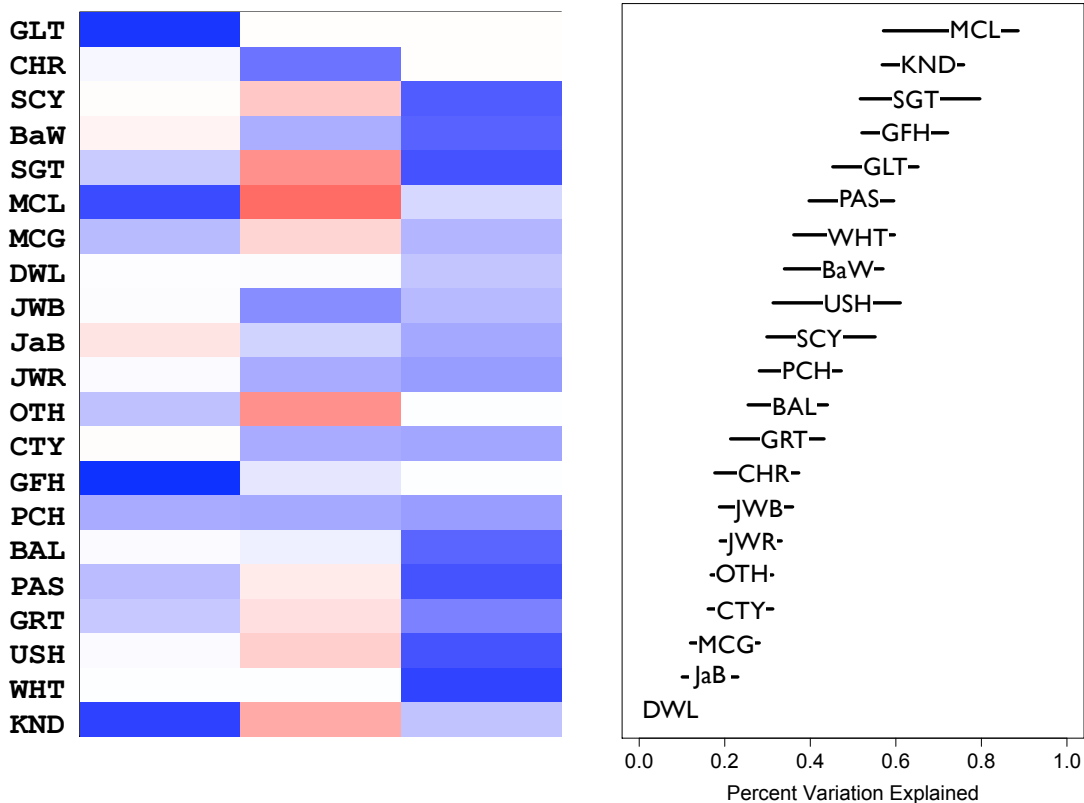


Figure 1: Left: The loadings matrix of each scotch upon the three latent factors. Note how the sparsity prior yields factor loadings on the first factor that easily identify it as the “single malt” factor. Right: 90% posterior credible intervals for the percent variation in each scotch explained by common factors, with the remainder explained idiosyncratically.

ration of lower-dimensional features was done after the fact. Given the large number of observations, working with the unconstrained model is reasonable, and a good basis for comparison.

In our analysis, we fit a 3-factor model to the data using Glenlivet, Chivas Regal, and Scoresby as the founding factors. This choice reflects the prior belief that two factors may be important in Scotch sales: how expensive the scotch is, and whether it is a single malt or a blend. (Fitting a four-factor model resulted in a largely zero-loaded fourth factor, suggesting that three is enough to capture most common variation.)

Figure 2 shows the posterior mean of the correlation matrix for the latent  $\mathbf{z}$  vector, and has been reordered to expose the groupings that emerged. The first five Scotches (reading top to bottom on the left hand side) represent, perhaps, the connoisseur’s Scotches; all of the single malts appear here, and the

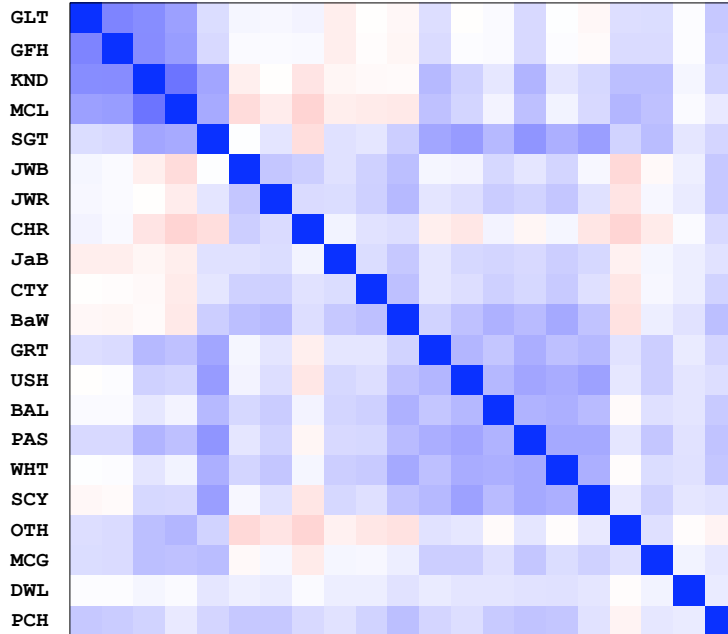


Figure 2: The posterior mean of the implied correlation matrix.

correlation within this group is strongly positive. The next group of six represents the popular mid-level Scotches. Finally, the third group of 11 includes mostly budget Scotches.

The story that emerges from the three-factor analysis is consistent with prior judgments about the importance of price and prestige. There are, however, some interesting twists. For example, while the first two factors are clearly dominant, the third factor still has non-trivial loadings (Figure 1). Clearly there is additional common variation in purchasing decisions, beyond that explained merely by prestige and price. Uncovering a plausible interpretation for this factor may suggest interesting possibilities for market researchers.

Figure 1 is intended to assess the overall variation in each Scotch’s sales that can be explained by commonalities among all the Scotches. This measure, which is implicit in the decomposition  $\Sigma = \mathbf{B}\mathbf{B}' + \Psi$ , is obtained by computing the ratio  $B_j' B_j / \Sigma_{jj}$  for the  $j$ th scotch at each step in the MCMC. This computation also provides a natural gauge of the posterior uncertainty in the “percent variation explained” metric (as shown by the error bars in the plot). Additional insight can be generated by computing the percent variation explained by the  $k$ th factor via the ratio  $b_{jk}^2 / \Sigma_{jj}$ .

Also, the scotches in the second “mid-level” category are all negatively correlated with “Other,” the catch-all category for scotches not explicitly appearing on the list. This may reflect brand loyalty specific to the category;

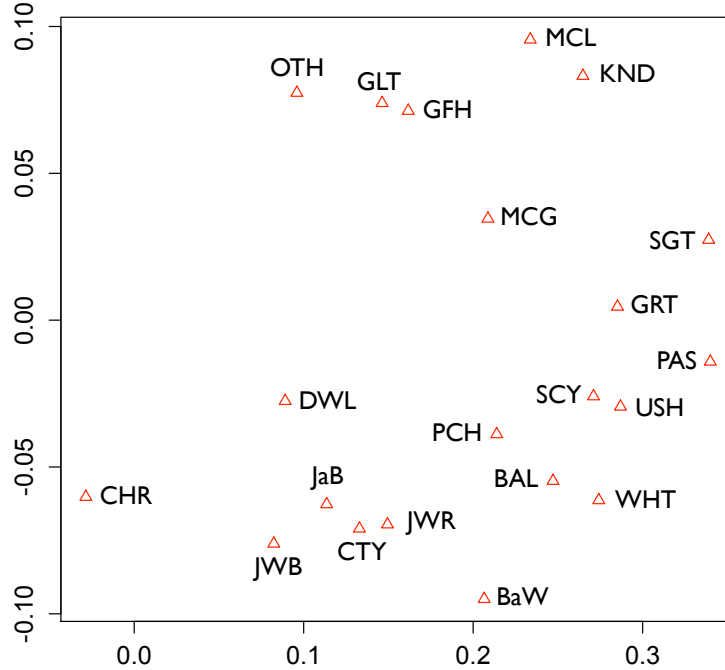


Figure 3: The first two mean eigenvectors of the implied correlation matrix. Compare to Figure 3 of Edwards and Allenby (2003).

many of these Scotches (such as Johnny Walker and Chivas Regal) are backed with significant advertising budgets.

Figure 3 is intended to show the similarity between our results to the study in Edwards and Allenby (2003) (Figure 3 in their work). This plot tries to spatially characterize different types of Scotches by looking at their relative position in the two-dimensional latent space defined by the first and second factors. In Edwards and Allenby (2003) this is done by looking at the loadings of each variable on the first and second principal components of the estimate for the latent covariance matrix. Here we present a ergodic average (based on the sequence of MCMC draws) of loadings on factors one and two from a orthogonal rotation of  $\mathbf{B}$ .

The arbitrariness of the scales notwithstanding, the substantive similarity between the two plots is striking. Two points are worth noting. First this should not come as a surprise as the *post-hoc* empirical strategy of Edwards and Allenby (2003) should recover the latent structure given the relative large number of observations. Second, it is very reassuring to see that a our model identifies this latent structure by working directly with a parsimonious representation, rather than by trying to recover such parsimony after the fact.

## 6.1 Goals

Political scientists have long sought to understand the historical forces that have led to the entrenched partisan rancor of modern American politics. Untangling the relative contributions of various polarizing factors is the subject of a vigorous scholarly debate. Putative explanations abound; these include the British colonial origins of the American political system, the effect of television news networks, the rush to gerrymander Congressional districts, the rise of immigration and income inequality in the 20th century, and the basic role that geography plays in representative democracy. As one might imagine, there is an enormous body of scholarly work on the subject, one that is far too large to cover here. A recent book-length treatment and a long list of references can be found in McCarty et al. (2006).

A more narrowly drawn ambition is simply to measure, rather than explain, ideological polarization. Indeed, if the folk wisdom is true and partisan behavior is really on the rise, then we would expect to see some signature of this behavior in Congressional voting records.

Many interest groups, such as the American Conservative Union or the National Rifle Association, attempt to detect this signature when they publish annual ratings of elected representatives in terms of how strongly they toe a “party line” or support a particular stance on an issue. The ratings themselves, however, are usually little more than a measure of how often the voter agreed with the interest group on a particular set of important votes.

Among social scientists, there are two common approaches for measuring a partisanship signature. In the political science literature this is often referred to as “ideal-point estimation” or “spatial voting” (Jessee, 2009). The first approach is to represent Congressional roll-call votes using some kind of discrete-choice regression model, such as a probit or logit regression. In this framework, party membership explicitly enters the model as a regressor. Partisanship for individual legislators, or groups of legislators, can then be measured by estimating, testing, or clustering regression coefficients. See the discussion in, for example, Bafumi et al. (2005).

The second commonly used approach is to represent votes in an underlying latent Euclidean space, and then to draw a cutting plane through this space that maximizes the correct party labeling of legislators. Each legislator can then be characterized by projecting her votes onto the cutting plane, and computing some summary measure of the votes’ location in that plane (e.g. the NOMINATE procedure of Poole and Rosenthal, 1997). The method is essentially a measure of who votes together, and how often they do it. Further

## Voting the Party Line

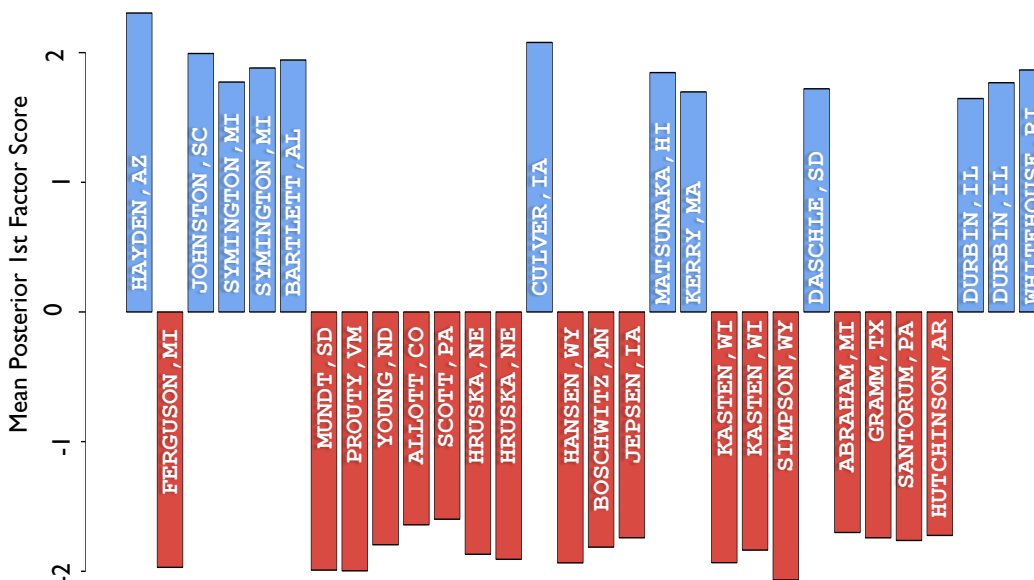


Figure 4: The most partisan voter of each of the past 30 congresses, ordered consecutively. The height of each bar represents the posterior mean of the respective senators’ first factor score. Familiar names on this list help to confidence in the model.

information on these methods can be found in McCarty et al. (1997).

## 6.2 Data, method, and results

The key intuition of the factor model is that observed variation can be decomposed into two pieces: a piece that depends upon common factors, and a piece that is idiosyncratic. This provides a rich alternative framework for measuring ideological polarization in voting bodies. Partisan behavior is, after all, predictable behavior; if we know that Diane Feinstein will vote against a particular bill, for example, then we can be fairly certain that most other Democrats will vote against it, too.

It is therefore natural to quantify the strength of this association by estimating the amount of variation in observed voting records that can be explained by a so-called “partisanship factor.” To demonstrate this, we analyze publicly available United States Congressional roll call data, restricting our attention to votes in the U.S. Senate between 1949 and 2009. Our main data set contains the 20 closest votes in each two-year Senate term. The close votes are typically the most interesting ones, and also allow us to sidestep the many near-unanimous votes which tend to be wholly unrelated to major policy de-

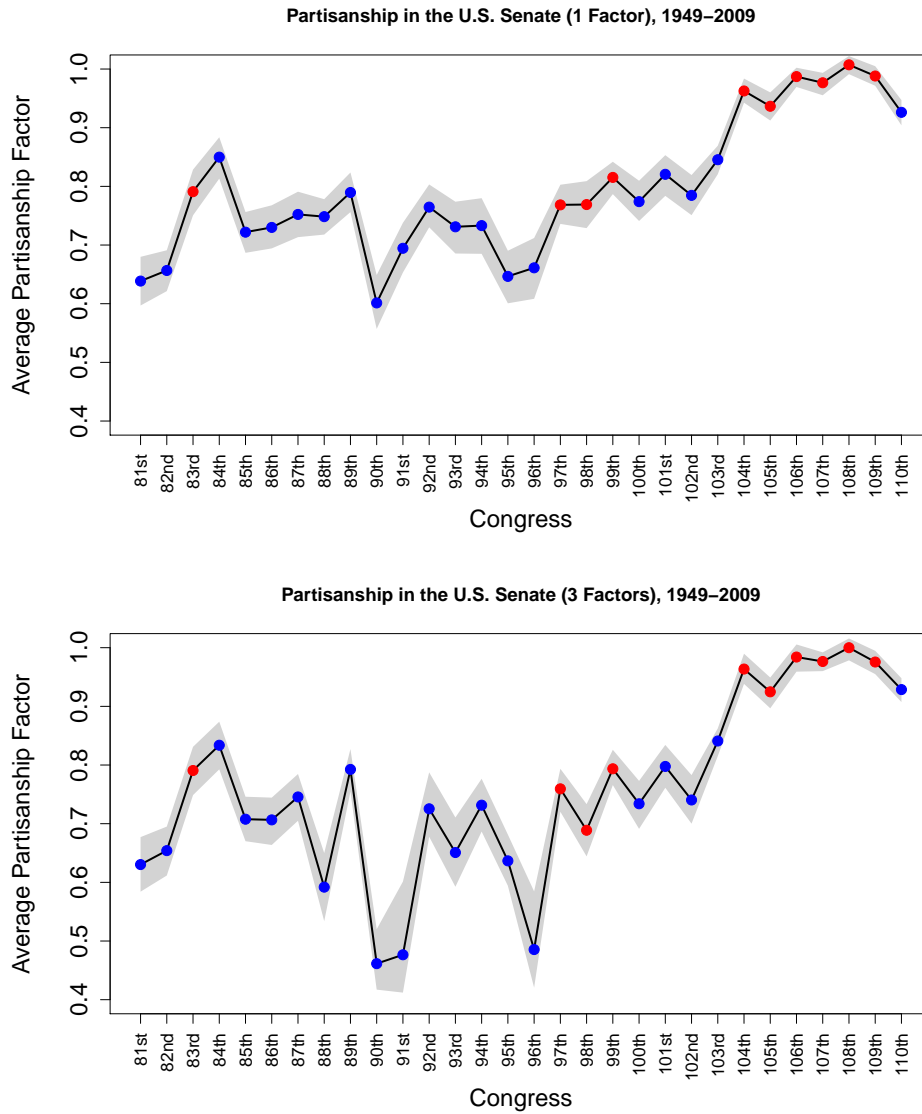


Figure 5: Normalized posterior magnitude of the “partisanship” factor. When two additional factors are added, the pattern in the series’ middle portion changes while the rest remains largely unchanged.

cisions. As was mentioned in the introduction, missing data in the form of no-votes are easily handled in our framework by simply drawing the latent Gaussian  $\mathbf{z}$  variables without truncation.

We associate the first factor in our analysis with the party membership of the senators by recording pseudo-votes for whether each is a Democrat. This vote is then used to “found” the first factor—which must be founded by some vote, of course, in light of the upper triangular structure imposed to ensure statistical identification of  $B$ . This is very different from the common approach of explicitly regressing votes upon party membership, since here, both the “design matrix”  $\mathbf{F}$  and the matrix of regression coefficients  $\mathbf{B}$  are estimated from the data (subject to appropriate identifying restrictions). This is far more flexible, and allows more interesting patterns to emerge:

- The large positive entries in the first column of  $\mathbf{B}$  can be interpreted as a constellation of Democrat-supported issues. Large negative loadings, meanwhile, correspond to Republican-supported issues. The patterns in the loadings matrix allow one to immediately spot “hot” issues in any given year. Large positive loadings were associated with, for example, the Equal Rights Amendment in the 97th Congress and the Brady Bill in the 103rd Congress.
- Changes in the first-factor loadings over time suggest structural changes in the way that policy issues map onto Republican and Democratic preferences. The nature of these changes may be particularly interesting during, for example, the era of the civil-rights movement.
- Other columns of  $\mathbf{B}$  suggest commonalities in voting behavior that is independent of party membership. These patterns can be used to generate hypotheses about why senators vote the way they do.
- The first factor score for each senator can be interpreted as an individual measure of partisanship. If  $f_{i,1}$  is large and positive, that indicates a tendency for senator  $i$  to vote for Democrat-supported issues with high probability. The scores also provide an interesting way to categorize and visualize senatorial voting patterns. Graphics such as Figure 4, which shows the most partisan senator of each of the past 30 congresses as measured by posterior mean factor score, may be of independent interest. They also provide us with a novel model validation tool by confirming that our latent factors square appropriately with expert qualitative judgements. Notice, for example, the group of highly partisan “Dixiecrats” from southern states in the 83rd through 86th Congresses.

- By analogy with Figure 1, the posterior distribution for  $\mathbf{B}$  and  $\Sigma$  allows us to see, for a given Congress, the extent to which the factor founded by party membership explains the observed variability in voting patterns.

To provide a summary measure of the amount of variation explained by the partisanship factor in each Congress, we examined the posterior distribution of the magnitude of the first column of the loadings matrix (appropriately normalized by the largest observed magnitude). In a sense, this allows us to examine “pure partisanship,” since the factor scores are independent a priori. Figure 5 plots the posterior mean magnitude of the first factor as it changes over time, with the shaded error representing a 95% posterior credible interval. We show this measure both for a one-factor model and a two-factor model. Three interesting facts emerge here.

First, both factor models show an upward trend in the overall amount of variation that can be explained by partisanship. This is consistent with the findings in the political science literature referenced above, and indeed buttresses those findings, given the very different methods that we have used to quantify partisanship.

Second, there is an obvious cyclical component in the partisanship measure over time. There are various theories for explaining this cyclicity in terms of the difference between presidential and midterm elections (Campbell, 1993; Gershtenson, 2006). By and large, this notion seems to be supported by the data; the cyclical component has a period of four years for most of the observation window. (Note that even-numbered Congresses convene after midterm elections, and odd-numbered Congresses convene after presidential elections.)

There are twists, however. Between the 91st and 95th Congresses, and again between the 104th and 109th Congresses, partisanship was locally higher after midterms. But between the 96th and 103rd Congresses, partisanship was locally higher after presidential elections. (These relationships are true both the one- and three-factor models.) This strange inversion pattern suggests that partisanship cycles may be more complicated than the regular quadrennial march of presidential elections would imply. It also raises the possibility that the apparent cyclicity may be a mirage, and that the observed changes are caused by other, non-cyclical factors.

Third, the one-factor and three-factor models are remarkably similar, except for the period between the 87th and 96th congresses. There are surely plausible stories to be told about the unique political forces at play, and about the extra dimension that is required to explain variability in voting patterns, during this period beginning in 1963.

We also note that there are many fruitful possibilities for extending the model. The use of covariates in the linear predictor  $\alpha(X)$  could easily be

incorporated to sharpen the investigation of hypotheses suggested by an initial analysis. Additionally, covariates may be incorporated at the level of the factor scores, fostering even greater ease of interpretation. Another interesting extension of the method would be to add an autocorrelation component, be it spatial or temporal, on the factor scores. This could account for senators serving in consecutive congresses, or senators in nearby states. This is just one example of how larger models could be constructed that would allow flexible borrowing of information across spatial and temporal dimensions, all within a factor-analytic framework.

## 7 DISCUSSION

We propose that the sparse factor-analytic probit model can serve the same role that principal-components analysis has long played in the exploration of continuous observations. The model may be especially helpful in social science and marketing applications, where categorical data can be the norm rather than the exception, and where latent factors confer an interpretational advantage—especially when they are carefully tied to germane observables. Our real examples demonstrate this approach.

Our simulations also demonstrate the beneficial regularizing properties of both the factor structure and the sparsity prior. Together, these allow the multivariate probit model to be effective even when the dimension  $p$  is quite large. Many other approaches to covariance estimation in this setting, such as banding or  $\ell^1$  regularization, do not offer the interpretational benefits of our method, nor do they easily accommodate additional modeling structure—for example, time series or spatial models.

Finally, there is the compelling computational advantage of the model: it avoids the need to sample from a truncated multivariate normal distribution, replacing it with a much easier problem involving a series of independent univariate truncations.

Taken together, these reasons suggest that the factor–probit model can be a useful default exploratory tool in the increasingly common situation of high-dimensional, correlated categorical data.

## REFERENCES

- O. Aguilar. *Latent Structure in Bayesian Multivariate Time Series Models*. PhD. Thesis, Duke University, 1998.
- O. Aguilar and M. West. Bayesian dynamic factor models and variance matrix discounting for portfolio allocation. *Journal of Business and Economic Statistics*, 18:338–357, 2000.
- J. Bafumi, A. Gelman, D. K. Park, and N. Kaplan. Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis*, 13:171–87, 2005.
- J. E. Campbell. *The Presidential Pulse of Congressional Elections*. The University of Kentucky Press, 1993.
- C. M. Carvalho, J. Lucas, Q. Wang, J. Nevins, and M. West. High-dimensional sparse factor modelling: Applications in gene expression genomics. *Journal of the American Statistical Association*, 103(4):1438–56, 2008.
- S. Chib and E. Greenberg. Analysis of multivariate probit models. *Biometrika*, 85(2):347–361, 1998.
- Y. D. Edwards and G. M. Allenby. Multivariate analysis of multiple response data. *Journal of Marketing Research*, 40:321–34, 2003.
- T. Elrod and M. P. Keane. A factor-analytic probit model for representing the market structure in panel data. *Journal of Marketing Research*, 32:1–16, 1995.
- S. Frühwirth-Schnatter and H. F. Lopes. Parsimonious Bayesian factor analysis when the number of factors is unknown. Technical report, University of Chicago Booth School of Business, 2009.
- A. E. Gelfand and A. F. M. Smith. Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409, 1990.
- S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.
- J. Gershtenson. Election cycles and partisanship in the u.s. house of representatives, 1857–2000. *Politics and Policy*, 34(4):690–705, 2006.
- J. Geweke and G. Zhou. Measuring the pricing error of the arbitrage pricing theory. *The Review of Financial Studies*, 9:557–587, 1996.
- S. Jessee. Spatial voting in the 2004 presidential election. *American Political Science Review*, 103(1):59–81, 2009.

- H. Lopes and M. West. Bayesian model assessment in factor analysis. *Statistica Sinica*, 14:41–67, 2004.
- N. McCarty, K. T. Poole, and H. Rosenthal. *Income Redistribution and the Realignment of American Politics*. American Enterprise Institute, 1997.
- N. McCarty, K. T. Poole, and H. Rosenthal. *Polarized America: The Dance of Ideology and Unequal Riches*. MIT Press, 2006.
- R. McCulloch and P. E. Rossi. An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, 64:207–240, 1994.
- R. McCulloch, P. Rossi, and N. Polson. Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics*, 99 (173–193), 2000.
- K. T. Poole and H. Rosenthal. *Congress: A Political-Economic History of Roll-Call Voting*. Oxford University Press, 1997.
- S. Press. *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference (2nd edition)*. New York: Krieger, 1982.
- B. Rajaratnam, H. Massam, and C. M. Carvalho. Flexible covariance estimation in graphical Gaussian models. *The Annals of Statistics*, 36(6): 2818–2849, 2008.
- J. G. Scott and J. O. Berger. An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136(7):2144–2162, 2006.
- C. Spearman. General intelligence, objectively determined and measured. *American Journal of Psychology*, 15:201–293, 1904.
- D. Sun and J. O. Berger. Objective priors for the multivariate normal model. In *Proceedings of the 8th Valencia World Meeting on Bayesian Statistics*. ISBA, June 2006.
- M. West. Bayesian factor regression models in the “large p, small n” paradigm. In J. M. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, editors, *Bayesian Statistics 7*, pages 723–732. Oxford University Press, 2003.
- X. Zhang, W. J. Boscardin, and T. R. Belin. Bayesian analysis of multivariate nominal measures using multivariate multinomial probit models. *Comput. Stat. Data Anal.*, 52(7):3697–3708, 2008. ISSN 0167-9473. doi: <http://dx.doi.org/10.1016/j.csda.2007.12.012>.