A Sparse Factor-Analytic Probit Model for Congressional Voting Patterns

P. Richard Hahn[1]

Booth School of Business
University of Chicago
Chicago, Illinois 60637 U.S.A.
richard.hahn@chicagobooth.edu

Carlos M. Carvalho

and

James G. Scott

Division of Statistics and Scientific Computing
and McCombs School of Business
University of Texas at Austin
Austin, Texas 78712, U.S.A.
{carlos.carvalho, james.scott}@mccombs.utexas.edu

Abstract

This paper adapts sparse factor models for exploring covariation in multivariate binary data, with an application to measuring latent factors in U.S. Congressional roll call voting patterns. This straightforward modification provides two advantages over traditional factor analysis of binary data. First, a sparsity prior can be used to assess the evidence that a given factor loading may be exactly zero, realizing a principled unification of exploratory and confirmatory factory analysis. Second, incorporating sparsity into existing factor-analytic probit models effects a favorable bias–variance tradeoff in estimating the covariance matrix of the multivariate Gaussian latent variables. Posterior summaries from this model applied to the roll-call data provide novel metrics of partisanship of a given Senate.

Key words: covariance estimation; factor models; multivariate probit models; voting patterns

---

[1]Corresponding author.

1   INTRODUCTION

In this paper, we extend the Bayesian multivariate probit model (Chib and Greenberg, 1998) to encompass a sparse factor-analytic approach for inference about the underlying correlation structure of binary data. We apply the proposed method to study partisan patterns in sixty years of non-lopsided roll call votes from the United States Senate, refining the similar analyses of Jackman (2001) and Clinton et al. (2004a). Our results show that the role of partisanship has risen distinctly in the last few decades, after hitting a low around 1970.

Our goal is not to propose a model that explains why Senators cast their votes; such a model would need to consider not only party membership but also geography, incumbency, committee membership, and much more besides. Rather, we undertake an exploratory analysis of the Senate roll call data. Our statistical approach differs from previous methods in that it allows practitioners to assesses whether individual elements of the factor loadings matrix—intuitively, the mapping between latent traits and observed votes—are identically equal to zero. In substantive terms, an inferred zero in the (j, g) entry of the factor loadings matrix corresponds to the judgment that the gth latent trait plays no discernible role in predicting whether Senators are likely to vote for or against bill j. In our approach, moreover, such judgments need not be encoded beforehand. Instead, they emerge naturally in the final analysis, assuming they are supported by the data.

The individual components of our model are the multivariate probit model, Gaussian factor models, and point-mass sparsity priors, each of which have been introduced in previous literature. In extending this literature by incorporating sparsity priors within a probit model, our paper has two motivating goals:

(i) **One-pass factor analysis.** Factor analysis has traditionally been used in two distinct modes: an exploratory phase, used to generate further hypotheses about the forces at play in the data, and a confirmatory phase where elements of the factor loadings are set to zero to reflect presumed conditional independencies. Our Bayesian framework with sparsity priors collapses these methods into a one-pass approach using ideas from Bayesian model averaging (Hoeting et al., 1999).

(ii) **Regularization.** Imposing a factor structure on a covariance matrix stabilizes estima-

tion, which is critical when the number of variables p is large relative to the sample size n (Rajaratnam et al., 2008). Such regularization is even more crucial when the estimated covariance matrix corresponds to an unobservable latent variable as it does in a multivariate probit model. Accordingly, our sparse factor model increases the degree of regularization by shrinking the elements of the factor loadings matrix towards zero. A simulation study demonstrates that sparse factor models effect a highly favorable bias–variance trade-off. This favorable trade-off persists even when there is no particular reason to suspect an underlying factor structure.

1.1 Background: Latent factor probit models

The latent factor probit model described in this section is well known in some research communities, going by various different names—ideal point estimation and item response modeling being two of the most common. Use of such models is currently widespread in many fields: political science (Jackman, 2001; Clinton et al., 2004a; Quinn, 2004; Jackman, 2009), statistics (Song and Lee, 2001, 2005), biostatistics (Bock and Gibbons, 1996), and marketing (Elrod and Keane, 1995). Two fine general references are Johnson and Albert (1999) and Bartholomew (1987).

Our presentation focuses on a normal latent–variable representation and the associated covariance-estimation problem because this perspective is well suited to the computational methods we employ. On this view, the model may be thought of as a Gaussian factor model embedded inside a multivariate probit model. But our approach can equally well be construed as an individual-level model; as Clinton et al. (2004a) observes, our probit model corresponds to the assumption that individual Senators have a quadratic utility function with normal errors, and occupy specific locations (or "ideal points") in a multi-dimensional Euclidean space.

1.1.1 The multivariate probit model

Let $\mathbf{Y} = (\mathbf{y}_1 \ldots \mathbf{y}_n)^t$, where each $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,p})^t$ represents p binary observations on a single subject. The multivariate probit model (Chib and Greenberg, 1998; Ashford and

3

Sowden, 1970) induces a probability distribution on y_i via an unobserved continuous quantity (utility) z_i, which is given a multivariate Gaussian distribution:

$$z_i \sim N(\alpha, \Sigma),$$

$$y_{ij} \mid z_{ij} \equiv \begin{cases} 0 & \text{if } z_{ij} \leq 0 \\ 1 & \text{if } z_{ij} > 0 \end{cases}. \tag{1}$$

In this way each of the 2^p possible binary vectors is associated with an orthant in $\mathbb{R}^p$ and assigned probability according to the corresponding multivariate Gaussian cumulative distribution function. Marginally,

$$y_{ij} \sim \text{Ber}(\rho_j)$$

$$\rho_j = \Pr(z_{ij} > 0). \tag{2}$$

Note that Σ is only identified up to its correlation structure, because scaling the latent utilities preserves the distribution of y_i as can be deduced from (1). Without loss of generality, we denote the mean of z_i by α, with the understanding that this may be a conditional predictor involving covariates (i.e., $\alpha_i \equiv \alpha(x_i)$).

The multivariate probit model reduces the problem of estimating 2^p probabilities to the problem of estimating the $p(p-1)/2$ pairwise correlations which compose Σ. The price of this reduction is the normality assumption on the latent utilities, which implies (among other things) a linear dependence structure. For many applications these assumptions are unobjectionable, and indeed the multivariate probit model is widely used (see, e.g., the examples discussed in Lesaffre and Molenberghs, 1991).

1.1.2 Gaussian factor models

While inference in a multivariate probit model is reduced to estimation of a correlation matrix, this task presents challenges of its own. Standard estimators are liable to be unstable when p is large compared to n, and can provide a distorted picture of the eigenstructure of Σ (Sun and Berger, 2006). These difficulties are magnified in the multivariate probit model, where the covariance matrix corresponds to an unobserved random variable.

4

A popular approach to address this instability is by imposing a factor structure, letting

$$\text{cov}(z_i) = \mathbf{B}\mathbf{B}^t + \mathbf{\Psi}, \tag{3}$$

where $\mathbf{\Psi}$ is p-by-p diagonal with non-negative elements and $\text{rank}(\mathbf{B}) = k < p$. We may rewrite this model by augmenting the representation to include latent factor scores f_i. Conditional on $\mathbf{B}$ and f_i, the elements of z_i are independent:

$$
\begin{aligned}
z_i &= \alpha + \mathbf{B}f_i + \epsilon_i, \quad \epsilon_i \sim \text{N}(0, \mathbf{\Psi}) \\
f_i &\sim \text{N}(0, \mathbf{I}_k).
\end{aligned}
\tag{4}
$$

For $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}^t + \mathbf{\Psi}$ to have a unique solution in $\mathbf{B}$, $\mathbf{B}$ must be constrained in some manner. In particular, two sorts of unidentifiability must be addressed: sign indeterminacy and rotation indeterminacy. Traditional solutions to this problem include forcing $\mathbf{B}$ to be have mutually orthogonal columns or $\mathbf{B}^t \mathbf{\Psi} \mathbf{B}$ to be diagonal. Another approach, popularized in the Bayesian community by Geweke and Zhou (1996) and adopted here, is to constrain $\mathbf{B}$ to be zero for upper-triangular entries $\{b_{js} = 0 : s > j\}$ for $s \leq k$ (addressing rotational indeterminacy) and positive along the diagonal $\{b_{jj} > 0\}$ for all j (addressing sign indeterminacy).

It is additionally possible to permit correlated factors so that $f_i \sim \text{N}(0, \mathbf{V})$. In this case the covariance is expressible as $\mathbf{\Sigma} = \mathbf{B}\mathbf{V}\mathbf{B}^t + \mathbf{\Psi}$ and identification of $\mathbf{B}$ and $\mathbf{V}$ can be achieved by setting certain $k(k-1)/2$ additional elements of $\mathbf{B}$ to zero (Bartholomew, 1987). Correlated factors are often preferable from the standpoint of interpretability. We revisit this topic in Section 2, where correlated factors combine with sparsity priors to provide valuable a priori biases towards easily interpretable factor rotations.

Factor models have been a topic of research for more than 100 years. A seminal reference is Spearman (1904), while Press (1982) and Bartholomew (1987) are key contemporary references. Bayesian factor models for continuous data have been developed by many authors, including Geweke and Zhou (1996), Aguilar and West (2000), and West (2003). A comprehensive bibliography may be found in Lopes (2003).

1.1.3 Latent factor probit models

Putting these pieces together—by assuming that the covariance parameter in a multivariate probit model admits a factor decomposition as in (3)—yields a latent factor probit model. This is equivalent to the individual level model

$$\Pr(y_{ij} = 1 \mid \alpha_j, \mathrm{b}_j, \mathrm{f}_i) = \Phi((\alpha_j + \mathrm{b}_j \mathrm{f}_i)/\sqrt{\psi_j}), \tag{5}$$

where $\Phi(\cdot)$ denotes the standard Normal cumulative distribution function and b_j denotes the jth row of $\mathbf{B}$.

In this context, the factor decomposition allows a simple identification strategy for the correlation structure, which is to let $\mathbf{\Psi} = \mathbf{I}$. Priors on element b_{js} of $\mathbf{B}$ then induce a prior distribution on the correlation coefficient ρ_{js} for $j \neq s$ (Rossi et al., 2006); the magnitude of these elements describe the amount of the variation attributable to the factor structure as opposed to idiosyncratic noise. Thus, the scale of the prior distribution governs prior expectations about the strength of the factor structure in terms of describing observed patterns of covariation.

1.2 A model for roll call votes

Political scientists have long sought to understand the historical forces leading 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, one that is too vast to recapitulate here. A recent book-length treatment and a long list of references may be found in McCarty et al. (2006). An important first step, which is the present focus, is simply to measure (rather than explain) ideological polarization.

For example, if we know that the majority whip votes against a particular bill, then we may believe that most other members of the majority will vote against it, too. The goal then becomes to quantify the strength of this association by estimating the amount of variation in observed voting records that can be accounted for by a so-called "partisanship factor." To this end, we analyze publicly available United States Congressional roll call data, restricting

our attention to votes in the U.S. Senate between 1949 and 2009.

In our analysis we let $y_i = (y_{i,1}, \ldots, y_{i,p})^t$ denote the vector of yea (1) or nay (0) votes of senator i. Our main data set contains the 30 closest votes in each two-year Senate term ($p = 30$). 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 decisions (Jackman, 2001). Missing data in the form of no-votes are easily handled in our framework, as will be described later.

Often party affiliation is not included specifically into a factor analysis of roll call votes; rather it emerges implicitly in that senators from different parties tend to cluster in factor space (Jackman, 2001; Clinton et al., 2004a). Our approach is to include this information explicitly by forcing the first factor f_{1i} to be left or right truncated depending on senator i's party affiliation. This ensures that the first factor is "pure" partisanship. Note that this approach is equivalent to a factor model with $p + 1$ observed dimensions, the first of which is a pseudo-vote recording each senator's party affiliation, where the residual variance on that element is set to zero ($\psi_1 = 0$); then $z_{1i} = f_{1i}$ up to a scale factor. To distinguish this privileged role, we denote the partisanship factor for Senator i by γ_i and the remaining latent factors by f_i. Similarly, we relabel the factor loadings column associated with γ as λ, so that

$$\mathrm{E}(z_i \mid f_i, \gamma_i, \alpha, \lambda, \mathbf{B}) = \alpha + \lambda\gamma_i + \mathbf{B}f_i,$$

and we define the dummy vote R_i recording whether Senator i is a Republican as $R_i = \mathbb{1}(\gamma_i > 0)$.

Under this parametrization, large positive entries in λ can be interpreted as a constellation of Republican-supported issues while large negative loadings on λ correspond to Democrat-supported issues. Likewise, the inferred γ's for each Senator can be interpreted as individual measures of partisanship: a large positive γ_i indicates a tendency for Senator i to vote for Democrat-supported issues with high probability. Meanwhile $\mathbf{B}$ encodes commonalities in voting behavior that are independent of party membership. These patterns can be used to generate hypotheses about why Senators vote the way they do, irrespective of party affiliation.

2   A Sparse factor–probit model

A sparse model is one in which certain of the parameters are permitted to be exactly zero. The canonical example is a linear model in which subsets of the regression coefficients may be estimated to be zero. The sparsity framework spans the areas of regularized prediction, hypothesis testing, and model selection, depending on whether it is viewed as a means, an end, or both.

Sparsity can be achieved in a variety of ways, such as direct testing or thresholding. Here we take an implicit testing approach via sparsity priors which affix a point-mass probability at zero (George and McCulloch, 1993; Mitchell and Beauchamp, 1988). A detailed discussion of this and similar Bayesian approaches to model selection in linear regression can be found in Ishwaran and Rao (2005). Our development closely follows West (2003) and Carvalho et al. (2008) who 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 observed variables, yielding a more parsimonious covariance structure. Specifically, the prior on the loadings matrix $\mathbf{B}$ takes the form:

$$(b_{js} \mid \nu_s, q_s) \sim q_s \cdot \mathrm{N}(0, \nu_s) + (1 - q_s)\delta_0(b_{js}) \tag{6}$$

$$\nu_s \sim \mathrm{IG}(c_s/2, c_s d_s/2) \tag{7}$$

$$q_s \sim \mathrm{Beta}(1, 1) \tag{8}$$

where there is a different variance component v_s and prior inclusion probability q_s associated with each column ($s = 1, \ldots, k$) of the loadings matrix. Here IG refers to an inverse Gamma distribution. The second term in the mixture $\delta_0(b_{js})$ denotes a point-mass measure at $b_{js} = 0$. The hyperparameters c_s and d_s modulate prior expectations about the scale of each column.

Note that although q_s is a prior probability shared by elements in the sth column of $\mathbf{B}$, this common prior parameter does not result in shared posterior inclusion probabilities across elements in a given column (excepting the uninteresting extremal cases of $q_s = 0$ or $q_s = 1$, which preclude posterior learning of sparsity patterns altogether). Even with a shared prior inclusion probability, different elements of the sth column of $\mathbf{B}$ will emerge with distinct

posterior probabilities. By treating the prior inclusion probabilities as model parameters to be estimated from the data, this model induces a strong multiplicity correction, automatically adjusting for the multiple-testing problem implicit in trying to learn the nonzero entries in each column of $\mathbf{B}$ (Scott and Berger, 2006, 2010). Analogous priors are placed on the elements of $\boldsymbol{\lambda}$ (which is simply a renamed column of $\mathbf{B}$).

Note also that in the context of sparsity, allowing correlated factors confers an additional advantage beyond possible ease of interpretation, which is that a greater range of covariance structures admit sparse representations in $\mathbf{B}$ if $\mathrm{cov}(\mathrm{f}_i) = \mathbf{V}$ is free to differ from the identity. To see this, observe that under the lower triangular identification scheme one may move freely between independent factors and dependent factors via the transformation

$$\tilde{\mathbf{B}} = \mathbf{BL} \tag{9}$$

where $\mathbf{V} = \mathbf{LL}^t$ denotes the (lower) Cholesky decomposition of the factor covariance. Both $\tilde{\mathbf{B}}$ and $\mathbf{B}$ are lower triangular, because the group of lower triangular matrices is closed under multiplication, and the two models are equivalent because

$$\mathbf{BVB}^t = \tilde{\mathbf{B}}\tilde{\mathbf{B}}^t. \tag{10}$$

But, in general, $\tilde{\mathbf{B}}$ will be much less sparse. That is, if $\tilde{\mathbf{B}}$ is sparse, setting $\mathbf{V} = \mathbf{I}$ gives a sparse $\mathbf{B}$, whereas a sparse $\mathbf{B}$ does not similarly yield a sparse $\tilde{\mathbf{B}}$. The columns of $\tilde{\mathbf{B}}$ are linear combinations of those of $\mathbf{B}$ and sums of sparse vectors need not be sparse.

Thus, sparsity priors on $\mathbf{B}$, combined with a prior on $\mathbf{V}$ centered at the identity matrix, serve to capture the inherent trade-off between parsimonious factor loadings and independent factors, automatically shrinking estimates towards more interpretable configurations.

As suggested in the previous section, we chose to let the partisanship factor be indepen-

dent of all other factors, while allowing those additional factors be correlated:

$$\begin{pmatrix} \gamma_i \\ f_i \end{pmatrix} \sim N(0, \tilde{\mathbf{V}}),$$

$$\tilde{\mathbf{V}} = \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{V} \end{pmatrix},$$

(11)

where $\mathbf{V}$ is an arbitrary correlation matrix. Though this approach may result in a potentially non-sparse partisanship loading λ, it has the interpretive advantage that a zero element of λ means that partisanship is wholly non-predictive of the corresponding roll call vote. Meanwhile, allowing the remaining partisanship-independent factors to be correlated fosters sparser loadings, as described above.

One may think of the resulting inferences as coming from a mixture over different candidate "confirmatory" models, where each model is weighted by its ability to describe the observed data (Hoeting et al., 1999). Crucially, by fixing certain elements of $\mathbf{B}$ to zero for identification purposes, we guarantee that all the models being mixed over observe these sameconstraints. This means that while different patterns of sparsity may emerge during posterior sampling, the interpretation of the factors within each of these models (as defined by the identification constraints) remains fixed.

2.1   Synopsis: the sparse factor model

For ease of reference we write down the full model here, including prior distributions and hyperparameter specifications. The model can be thought of as having three parts: the data $\mathbf{Y} = (y_1, y_2, \ldots, y_n)$ given the latent utilities $\mathbf{Z} = (z_1, z_2, \ldots, z_n)$, the latent factors $\mathbf{F} = (f_1, f_2, \ldots, f_n)$ and $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_n)$ given the model parameters and hyperparameters $\{\alpha = (\alpha_1, \ldots, \alpha_p)^t, \mathbf{B}, \mathbf{V}, \lambda = (\lambda_1, \ldots, \lambda_p)^t, q = (q_1, \ldots, q_k), \nu = (\nu_1, \ldots, \nu_k)\}$, and finally the prior over the model parameters. Altogether, for $j = 1, \ldots, p$, $i = 1, \ldots, n$ and $s = 1, \ldots, k$,

we have

$$
\begin{aligned}
R_i &= \mathbb{1}(\gamma_i > 0) \\
y_i \mid z_i &= \mathbb{1}(z_i > 0), \\
z_i &\sim N(\alpha + \lambda\gamma_i + \mathbf{B}f_i, \mathbf{I}_p), \\
f_i &\sim N(0, \mathbf{V}), \quad \gamma_i \sim N(a, 1), \\
\alpha_j &\sim N(0, \nu_\alpha), \quad a \sim N(0, \nu_a) \\
\mathbf{V} &\sim IW(\mathbf{V}_0, h), \\
(b_{js} \mid \nu_s, q_s) &\sim q_s \cdot N(0, \nu_s) + (1 - q_s)\delta_0(b_{js}), \\
\nu_s &\sim IG(c_s, c_s d_s/2), \quad q_s \sim Beta(1, 1), \\
(\lambda_j \mid \nu_\lambda, q_\lambda) &\sim q_\lambda \cdot N(0, \nu_\lambda) + (1 - q_\lambda)\delta_0(\lambda_j), \\
\nu_\lambda &\sim IG(c_s, c_s d_s/2), \quad q_\lambda \sim Beta(1, 1).
\end{aligned}
$$

Here $IW(\mathbf{V}_0, h)$ denotes an inverse-Wishart distribution with mode matrix $\mathbf{V}_0$ and degrees of freedom h. Note that $\Phi(a)$ can be interpreted as the latent fraction of Republicans among hypothetical Senators.

Recall that this model has structural zeros in $\mathbf{B}$ for identification purposes. For example, in a three factor model ($k = 3$), the identification structure could could be achieved by setting elements b_{41}, b_{52} and b_{61} to zero, allowing $\mathbf{\Sigma} = \mathbf{B}\mathbf{V}\mathbf{B}^t + \mathbf{I}$ to be solved uniquely for $\mathbf{B}$ and $\mathbf{V}$, given $\mathbf{\Sigma}$. Having ones on the diagonal permits $\mathbf{V}$ to be a covariance matrix, rather than being restricted to a correlation matrix. One should note that fixing this structure beforehand is convenient notationally, but renders the ordering of the votes a nontrivial modeling decision.

Our analysis of the roll-call data (summarized in Section 3) was performed with $\nu_\alpha = \nu_a = 10$, $h = 6$, $\mathbf{V}_0 = \mathbf{I}_k$ and $c_s = 2$ and $d_s = 1$ for all s. We used a three factor model (the partisanship factor γ, plus $k = 2$ additional factors). Larger models are possible, but in practice we did not encounter a case where a fourth factor exhibited significant loadings.

2.2 Posterior sampling

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.

1. Draw the latent observation matrix $\mathbf{Z} = (z_{ij})$ by drawing each $(z_{ij} \mid -) \sim \mathrm{N}(\alpha_j + \lambda_j \gamma_i + \mathrm{b}_j \mathrm{f}_i, 1)$ truncated above at 0 if $y_{ij} = 0$ and below at 0 if $y_{ij} = 1$.

2. Sample γ_i as $(\gamma_i \mid -) \sim \mathrm{N}((1 + \lambda^t \lambda)^{-1}(\nu_a^{-1} a + \lambda^t(\mathrm{z}_i - \boldsymbol{\alpha} - \mathbf{B}\mathrm{f}_i)), (1 + \lambda^t \lambda)^{-1})$, truncated above at 0 if senator i is Republican and below at 0 if senator i is Democrat. Handling Independents is as simple as treating the sign of their latent partisanship as missing data and simply bypassing the truncation.

3. Sample a as $(a \mid -) \sim \mathrm{N}\left((\nu_a^{-1} + n)^{-1} \sum_i \gamma_i, (\nu_a^{-1} + n)^{-1}\right)$.

4. Sample the elements of $\boldsymbol{\alpha}$ as $(\alpha_j \mid -) \sim \mathrm{N}(m_j, M)$ with

$$
\begin{aligned}
\tilde{\mathrm{z}}_i &= \boldsymbol{\Psi}^{\frac{1}{2}}(\mathrm{z}_i - \lambda \gamma_i - \mathbf{B}\mathrm{f}_i), \\
M &= (\nu_\alpha^{-1} + n)^{-1}, \\
m_j &= M \sum_i \tilde{\mathrm{z}}_{ji}.
\end{aligned}
\tag{12}
$$

5. Sample the vectors of factor scores independently as $(\mathrm{f}_i \mid -) \sim \mathrm{N}(\mathrm{m}, \mathbf{M})$, with

$$
\begin{aligned}
\mathbf{M} &= \mathbf{V} - \mathbf{V}^t \mathbf{B}^t (\mathbf{B}\mathbf{V}\mathbf{B}^t + \mathbf{I}_p)^{-1} \mathbf{B}\mathbf{V}, \\
&= (\mathbf{V}^{-1} + \mathbf{B}^t \mathbf{B})^{-1}, \\
\mathrm{m} &= \mathbf{M}\mathbf{B}^t(\mathrm{z}_i - \boldsymbol{\alpha} - \lambda \gamma_i).
\end{aligned}
$$

6. Sample $(\mathbf{V} \mid -) \sim \mathrm{IW}(\mathbf{V}_0 + \mathbf{F}\mathbf{F}^t, n + h)$.

7. To sample the unconstrained loadings of column s, define $\tilde{\mathrm{z}}_i = \mathrm{z}_i - \boldsymbol{\alpha} - \lambda \gamma_i - \mathbf{B}_{-s}\mathrm{f}_{-s,i}$ where $\mathbf{B}_{-s}$ is $\mathbf{B}$ with column s removed and $\mathrm{f}_{-s,i}$ is the vector of factor scores for

individual i with factor score s removed. Sample

$$(b_{js} \mid -) \sim (1 - \hat{q}_{js})\delta_0 + \hat{q}_{js} \cdot \mathrm{N}(\hat{b}_{js}, \hat{\nu}_{js}),$$

where

$$\begin{aligned}
\hat{\nu}_{js} &= \left(\sum_{i=1}^{n} f_{s,i}^2 + \nu_s^{-1} \right)^{-1}, \\
\hat{b}_{js} &= \hat{\nu}_{js} \left(\sum_{i=1}^{n} f_{s,i} \tilde{z}_{ij} \right), \\
\hat{r}_{js} &= \frac{\mathrm{N}(0 \mid 0, \nu_j)}{\mathrm{N}(0 \mid \hat{b}_{js}, \hat{\nu}_{js})}, \\
\hat{q}_{js} &= \frac{\hat{r}_{js}}{\frac{1 - q_s}{q_s} + \hat{r}_{js}}.
\end{aligned}$$

8. Let m_s be the number of (unconstrained) elements in B_s currently set to non-zero. Draw
$$\nu_s \sim \mathrm{IG}\{(c_s + m_s)/2, (c_s d_s + \mathrm{B}_s^t \mathrm{B}_s)/2\}.$$

9. Draw $(q_s \mid -) \sim \mathrm{Be}(1 + m_s, 1 + \tilde{m}_s - m_s)$, where m_s is as in the previous step and $\tilde{m}_s$ is the maximum possible number of non-zero elements for column s (given the fixed identification constraints).

10. The elements of λ are drawn analogously to those of the B_s: define $\tilde{z}_i = z_i - \alpha - \mathbf{B}\mathrm{f}_i$ and sample
$$(\lambda_j \mid -) \sim (1 - \hat{q}_{\lambda,j})\delta_0 + \hat{q}_{\lambda,j} \cdot \mathrm{N}(\hat{\lambda}_j, \hat{\nu}_\lambda),$$

where

$$\hat{\nu}_{\lambda,j} = \left(\sum_{i=1}^{n}\gamma_i^2 + \nu_\lambda^{-1}\right)^{-1},$$

$$\hat{\lambda}_j = \hat{\nu}_{\lambda,j}\left(\sum_{i=1}^{n}\gamma_i\tilde{z}_{ij}\right),$$

$$\hat{r}_{\lambda,j} = \frac{\mathrm{N}(0 \mid 0, \nu_\lambda)}{\mathrm{N}(0 \mid \hat{\lambda}_j, \hat{\nu}_\lambda)},$$

$$\hat{q}_{\lambda,j} = \frac{\hat{r}_{\lambda,j}}{\frac{1-q_\lambda}{q_\lambda} + \hat{r}_{\lambda,j}}.$$

11. Likewise, let m_λ be the number of elements in λ currently set to non-zero, and draw

$$(\nu_\lambda \mid -) \sim \mathrm{IG}\{(c_\lambda + m_\lambda)/2, (c_\lambda d_\lambda + \lambda^t\lambda)/2\}.$$

12. Again similarly, draw $(q_\lambda \mid -) \sim \mathrm{BE}(1 + m_\lambda, 1 + p - m_\lambda)$.

In this sampler, as highlighted in Song and Lee (2005) in the context of an analogous E-M algorithm, it is not necessary to draw from a high-dimensional truncated multivariate normal distribution; all the dependence among the elements of z_i is encoded in $\mathbf{B}$ so that the truncations arising from the observed data y_{ij} can be handled independently, leading to a series of easier univariate truncations.

As with independent senators, missing data (uncast votes, in this case) may be accomodated by simply drawing the corresponding latent utilities without truncation in the first step of the sampler, under the assumption of noninformative missingness. Informative missingness may also be incorporated by truncating with some pre-determined probability, possibly as a function of additional covariates.

2.3 Simulation study

This section shows via simulation that sparse factor models result in a highly favorable bias-variance trade-off. We compare three models of the covariance structure, the Wishart model, a $k = 6$ factor model, and a $k = 6$ sparse factor model. We examine the performance of

each of these models under four distinct regimes: (1) data drawn with underlying covariance matrix which possesses a factor structure with three factors; (2) data drawn with underlying covariance matrix which possesses a factor structure with ten factors; (3) data drawn with underlying covariance matrix with no factor structure (equivalently, with number of factors equal to number of dimensions); and (4) data drawn with underlying covariance matrix given by the identity matrix.

Specifically, for a given covariance matrix $\boldsymbol{\Sigma}$ and mean vector $\boldsymbol{\alpha}$, we let

$$\mathbf{C} = \mathbf{D}^{-\frac{1}{2}}\boldsymbol{\Sigma}\mathbf{D}^{-\frac{1}{2}} \tag{13}$$

$$\mathbf{D} = \operatorname{diag}(\boldsymbol{\Sigma}) \tag{14}$$

$$z_i \sim \mathrm{N}(\alpha, \mathbf{C}) \tag{15}$$

$$y_i = \mathbb{1}(z_i > 0). \tag{16}$$

In all regimes $\boldsymbol{\alpha}$ was drawn as $\mathrm{N}(0, 0.2\mathbf{I})$.

For all simulations the number of observations was fixed at $n = 50$. 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. Frobenius and Stein losses are given, respectively, as:

$$L_F(\tilde{\mathbf{C}}, \mathbf{C}) = \operatorname{tr}((\tilde{\mathbf{C}} - \mathbf{C})^2) \tag{17}$$

$$L_S(\tilde{\mathbf{C}}, \mathbf{C}) = \operatorname{tr}(\tilde{\mathbf{C}}\mathbf{C}^{-1}) - \ln(\det(\tilde{\mathbf{C}}\mathbf{C}^{-1})) - p. \tag{18}$$

Note that the regimes examined here include cases where $p > n$ and also $n > p$, cases where the assumed factor structure has both too few and too many included factors, and that both the factor models (sparse and non-sparse) and the Wishart model are centered at the identity matrix (since $\mathrm{E}(\mathbf{B}) = 0$). As such, this battery provides a good snap-shot of the performance of the three models across a range of plausible real-data scenarios. Results are reported in Table 1.

The differences between the various models when $n > p$ are modest, but the factor models is seen to dominate the Wishart model. Meanwhile, the difference between sparse

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

| | | | Fitted Model | |
Loss function	True 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
	$p = 20$, 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
	$p = 20$, 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
	$p = 100$, 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
	$p = 100$, identity	94.6	136.3	2.1

versus nonsparse factor models can be attributed to which of these models is closest a priori to the generating model—so that the sparse model performs better for the identity and for the $k = 3$ true models, while the non-sparse does better for the $k = 10$ and $k = 20$ generated data. However, with $p = 100$ and $n = 50$ the slight penalty the sparse model incurs for underestimating the number of factors is shown to be relatively minor. In this setting, the benefit over the Wishart distribution becomes more stark. Naturally, whichever model favors the truth (a priori) still comes out on top. For instance, the sparse model on the identity matrix gives outstanding performance.

That said, the $p = 100$, $k = 10$ (row two) results are most interesting: since six factors are insufficient to perfectly reconstruct $\mathbf{\Sigma}$ in this case, it is striking that the factor model still outperforms the Wishart. Furthermore, incorporating the sparsity component does not suffer much at all in this case, while we can see that when the true number of factors is less than six, adding the sparsity offers a substantial benefit (row one). In short, it would appear that the bias induced by the factor structure assumption is more than compensated by the reduced variance when $p > n$.

3 Analysis of U.S. Senate roll call votes, 1949–2009

3.1 Overview of approach

Our analysis of U.S. Senate roll call votes focuses on three different, complementary methods for characterizing partisan voting behavior. First, we can rank individual Senators by partisanship, along the lines of Clinton et al. (2004b). This is done by examining the posterior distribution of each Senator's partisanship factor score, $p(\gamma_i|\mathbf{Y})$. We observe that there are typically large differences among Senators in both the mean of this distribution as well as the associated uncertainty (see, for example, Figure 3).

Second, we can ascertain which roll calls—if any—were plausibly free of party influence. We do this by examining the posterior probabilities that each roll call loads on each of the $k + 1$ factors.

Third, we can study historical changes in the importance of partisanship for predicting roll–call votes for the Senate as a whole (rather than for individual Senators, as above). This can be operationalized in two ways for each Senate: (1) as the percentage of variation in roll-call votes attributable to the partisanship factor; and (2) as the fraction of non–zero elements in the partisanship factor loadings vector $\boldsymbol{\lambda}$. In principle, this permits two different notions of increasingly partisan behavior to be detected: individual roll calls becoming more severely partisan, versus more roll calls overall becoming at least somewhat partisan.

3.2 Partisanship over time

The three characterizations of partisanship sketched above may be estimated for any Congress in aggregate by considering the posterior distribution of quantities involving all votes. Specifically, for each of the last 30 Congresses, spanning the 60 year period from 1949 to 2009, we consider the posterior distributions of the following quantities:

1. the overall proportion of variation attributable to partisanship, defined as

$$p^{-1} \sum_j \frac{\lambda_j^2}{\lambda_j^2 + \mathbf{b}_j \mathbf{V} \mathbf{b}_j^t + 1};$$
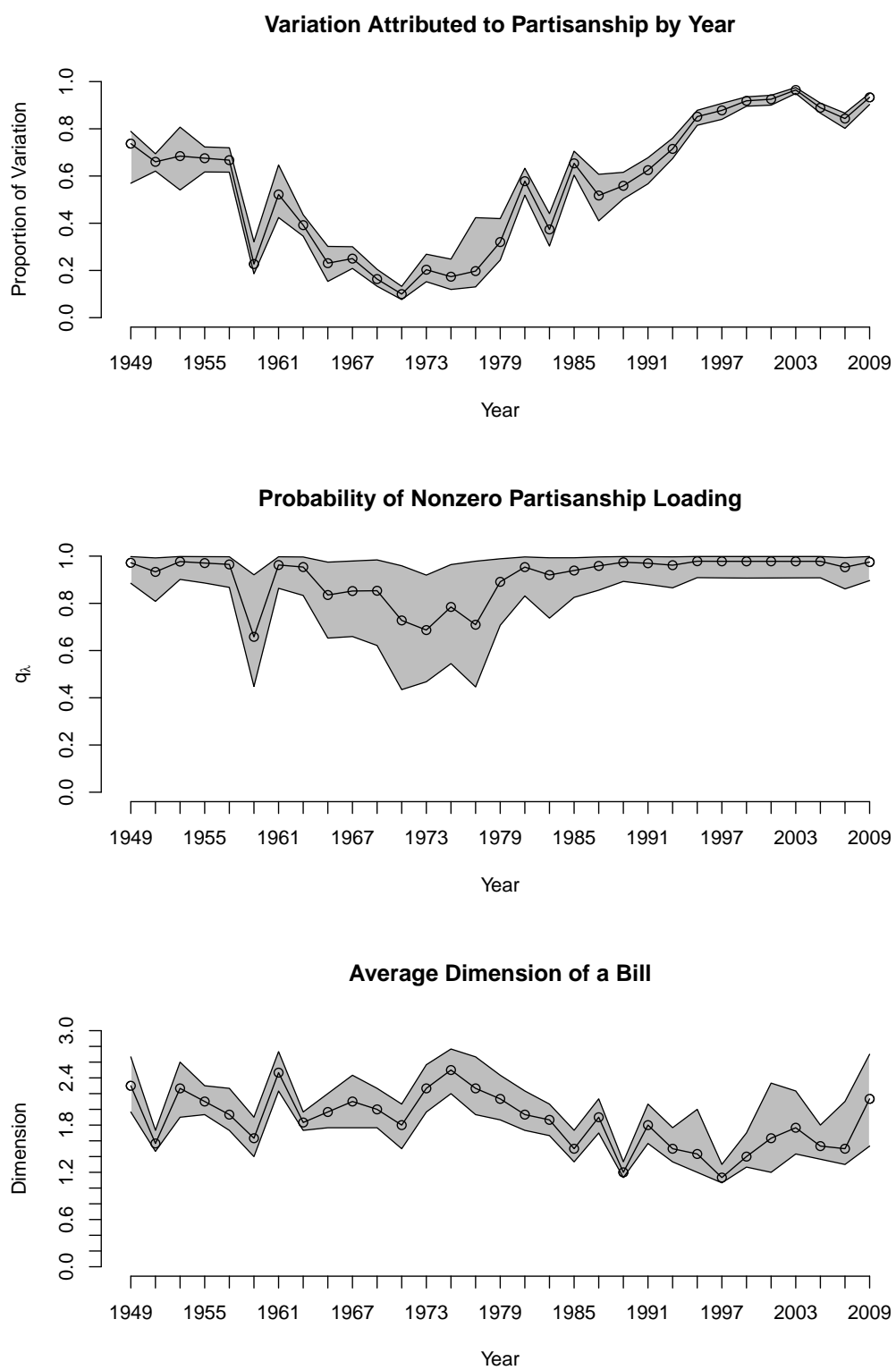
(19)

Figure 1: Three measures of partisan behavior in Senate roll-call votes, 1949–2009.

2. the overall probability of a nonzero partisanship loading, given by the parameter q_λ; and

3. the average dimensionality across the bills, defined as

$$p^{-1} \sum_{j,g} \mathbb{1}(b_{jg} \neq 0). \tag{20}$$

These summaries provide a snapshot of time-varying partisanship. A key feature of the inferred partisanship trajectory shown in the top panel of Figure 1 is that Senate voting patterns of the 60s and 70s suggest forces at play beyond those captured by party affiliation; this much has long been recognized. Our analysis shows, however, that as partisanship has fluctuated noticeably, the dimensionality of the voting landscape has remained relatively stable. Justifying this finding from a theoretical political science perspective would be an interesting line for future inquiry.

An important issue in high-dimensional Bayesian analysis is prior sensitivity. Throughout, our focus has been on specifying a default, or conventional, set of hyperparameters in an attempt to minimize the influence of prior inputs. These choices follow closely along the lines of Carvalho et al. (2008). In many cases such default choices are natural—for example, a uniform distribution for the prior inclusion probabilities q_s. It is important to observe that improper priors cannot be used for the variances ν_s for each column of the factor loadings matrix, since these variances enter into an implicit Bayes-factor computation in the sparse model. In a very real sense, no "objective" choice for this prior is possible. In this case our use of an inverse-gamma(1,1) prior recognizes the need for propriety, while simultaneously attempting to exhibit as little influence as possible over the final answers.

Space—and the sheer combinatorial explosion of hyperparameters—both preclude a full discussion of prior robustness, but we observe two important generalities. First, the prior for q_s (the prior inclusion probabilities) makes the greatest amount of difference in posterior inference, which is consistent with the findings of Scott and Berger (2006) in the parallel context of multiple hypothesis testing. In our experiments, we observed that changing the hyperparameters that govern q_s could predictably change the corresponding posterior inclu-

19

Magnitude of Posterior Mean Factor Loadings

Bill No.	λ	B_1	B_2
166			
184			
117			
123			
165			
399			
36			
37			
133			
191			
195			
28			
153			
96			
317			
70			
85			
332			
342			
122			
376			
190			
299			
316			
55			
325			
69			
225			
226			
167			

86th Senate

Magnitude of Posterior Mean Factor Loadings

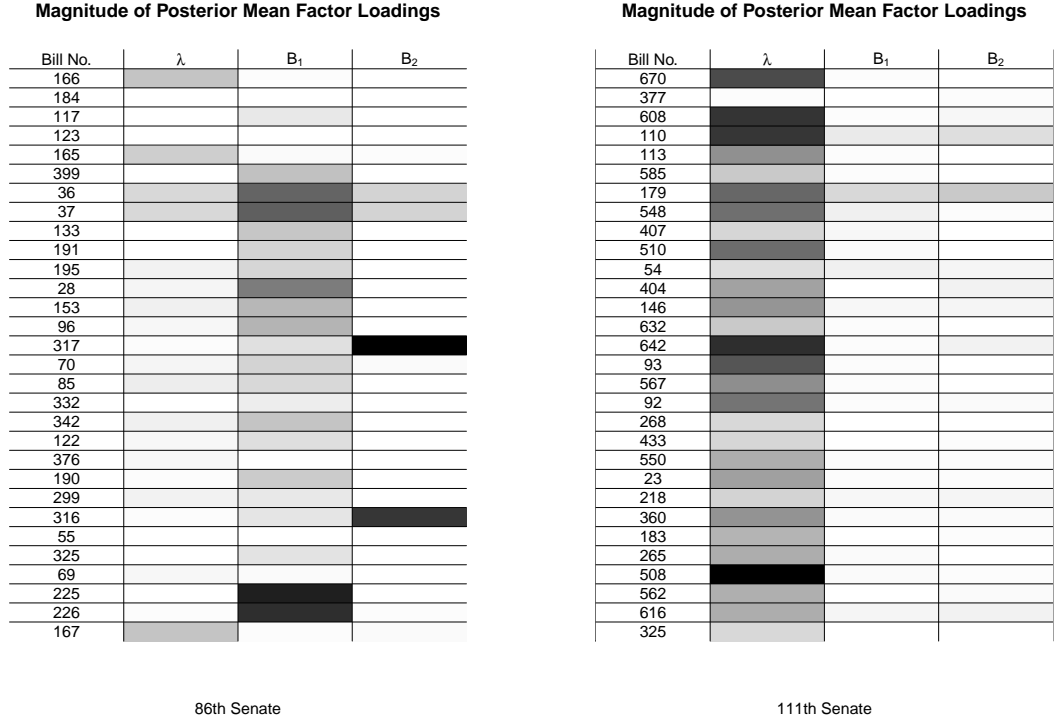| Bill No.	λ	B_1	B_2
670			
377			
608			
110			
113			
585			
179			
548			
407			
510			
54			
404			
146			
632			
642			
93			
567			
92			
268			
433			
550			
23			
218			
360			
183			
265			
508			
562			
616			
325			

111th Senate

Figure 2: Grayscale shades depict relative magnitudes of posterior mean factor loadings. The 86th Senate has two prominent factors unrelated to partisanship, while the 111th Senate shows that party affiliation accounts for the majority of variation in the observed votes.

sion probabilities, but very rarely change other measures of partisanship such as percent-variation-explained. Given the existence of an obvious default (uniform) prior for q_s, this would seem untroubling. Second, results for later Congresses are less sensitive to the prior than results of earlier Congresses. This is fully consistent with the notion that the statistical signature of partisanship has become more pronounced.

3.3 Comparing the 86th and 111th Congresses

Within a given congress, individual senators and individual bills can be evaluated and ranked in terms of their inferred partisanship. In terms of ranking senators, we recapitulate the basic findings of Clinton et al. (2004a) for the 108th Congress. Our findings differ from theirs mainly in the larger posterior intervals, likely owing simply to the fewer number of votes we considered, but also perhaps to the increased flexibility of our sparse model. We

Table 2: Posterior summaries for the 30 closest Senate votes in the 86th and 111th Congresses. PV = Proportion of Variation attributed to partisanship. PIP = Posterior Inclusion Probability. MPD = Mean Posterior Dimension.

	86th Senate				111th Senate		
Bill No.	PV	PIP	MPD	Bill No.	PV	PIP	MPD
166	0.95	1.00	1.37	670	1.00	1.00	1.58
184	0.02	0.32	0.39	377	0.12	0.86	1.42
117	0.01	0.23	1.33	608	1.00	1.00	2.13
123	0.02	0.32	0.60	110	0.98	1.00	2.67
165	0.92	1.00	1.46	113	0.98	1.00	1.94
399	0.01	0.26	1.32	585	0.95	1.00	1.93
36	0.34	1.00	2.92	179	0.95	1.00	2.73
37	0.34	1.00	2.93	548	0.98	1.00	2.43
133	0.01	0.25	1.31	407	0.91	1.00	2.16
191	0.01	0.24	1.34	510	0.98	1.00	2.35
195	0.36	1.00	2.13	54	0.82	1.00	2.30
28	0.07	0.87	1.96	404	0.96	1.00	2.41
153	0.27	1.00	2.08	146	0.97	1.00	2.40
96	0.10	0.83	1.93	632	0.95	1.00	1.99
317	0.01	0.52	1.93	642	0.99	1.00	2.45
70	0.26	0.86	2.11	93	0.99	1.00	2.11
85	0.49	1.00	2.07	567	0.99	1.00	2.05
332	0.05	0.44	1.42	92	0.99	1.00	2.11
342	0.31	1.00	1.26	268	0.92	1.00	1.95
122	0.23	0.97	2.03	433	0.93	1.00	2.00
376	0.38	1.00	1.26	550	0.97	1.00	2.00
190	0.06	0.58	1.67	23	0.98	1.00	2.13
299	0.46	1.00	2.03	218	0.84	1.00	2.40
316	0.01	0.49	1.87	360	0.98	1.00	2.07
55	0.00	0.15	0.25	183	0.97	1.00	1.91
325	0.01	0.21	1.30	265	0.97	1.00	2.14
69	0.38	0.98	1.51	508	0.99	1.00	2.16
225	0.01	0.39	1.51	562	0.97	1.00	2.02
226	0.02	0.39	1.57	616	0.96	1.00	2.27
167	0.94	1.00	1.44	325	0.93	1.00	1.77

also include the most recent 111th Congress for comparison as a supplementary file.

In terms of analyzing the bills themselves, we compare inferences from the most recent 111th Senate and the 86th Senate from 1960–62. As witnessed in the time series above, we expect to see a less fervent partisanship signature in the 1960s votes, and indeed we do. By looking more closely at the inferred factor loadings we are able to venture interpretations of the non-partisan columns on the basis of the emergent sparsity patterns. In particular we find that the first non-partisan factor loads heavily on two labor union bills and two bills concerning appropriations for teachers' salaries. These findings are reported in the table below.

4 DISCUSSION

The sparse factor–probit model represents a principled approach to bridging the gap between exploratory and confirmatory factor analyses using ideas from Bayesian model averaging.

The result is a one-pass approach to uncovering covariance patterns with ready substantive interpretations, as our application to U.S. Senate roll call votes demonstrates.

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 ℓ^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.

As noted in Jackman (2001), extending Bayesian models of voting is relatively straightforward and our sparse model naturally inherits this property. The use of covariates in the linear predictor $\alpha(x_i)$ could easily be incorporated to sharpen the investigation of hypotheses suggested by an initial analysis. A most interesting extension of the method—in light of our approach to incorporating partisanship—would be to add an autocorrelation component, be it spatial or temporal, to the factor scores. This would account for Senators serving in consecutive congresses, or Senators in nearby states, and would intrinsically handle the interesting (but relatively infrequent) case of Senators switching party affiliation (Clinton et al., 2004a). (In our framework the switched truncation in consecutive congresses for such a senator, in combination with temporal autocorrelation, would pull those factor score estimates towards neutral zero.) 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.

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

References

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. Ashford and R. Sowden. Multivariate probit analysis. *Biometrics*, 26:535–546, 1970.

D. Bartholomew. *Latent variable models and factor analysis.* Charles Griffin, 1987.

R. D. Bock and R. D. Gibbons. High-dimensional multivariate probit analysis. *Biometrics*, 52(4): 1183–1194, 1996.

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(484):1438–1456, December 2008.

S. Chib and E. Greenberg. Analysis of multivariate probit models. *Biometrika*, 85(2):347–361, 1998.

J. Clinton, S. Jackman, and D. Rivers. The statistical analysis of roll call data. *American Political Science Review*, 90(2), 2004a.

J. Clinton, S. Jackman, and D. Rivers. "The most liberal senator"? Analyzing and interpreting congressional roll calls. *Political Science & Politics*, October 2004b.

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.

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.

E. I. George and R. E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88:881–889, 1993.

J. Geweke and G. Zhou. Measuring the pricing error of the arbitrage pricing theory. *The Review of Financial Studies*, 9:557–587, 1996.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinksy. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4), 1999.

H. Ishwaran and J. S. Rao. Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730–773, 2005.

S. Jackman. Multidimensional analysis of roll call data via Bayesian simulation: Identification, estimation, inference, and model checking. *Political Analysis*, 9(3), 2001.

S. Jackman. *Bayesian Analysis for the Social Sciences*. John Wiley and Sons, 2009.

V. Johnson and J. Albert. *Ordinal data modeling*. Springer, New York, 1999.

E. Lesaffre and G. Molenberghs. Multivariate probit analysis: a neglected procedure in medical statistics. *Statistics in Medicine*, 10:1391–1403, 1991.

H. Lopes. Factor models: An annotated bibliography. *Bulletin of the International Society for Bayesian Analysis*, 2003.

N. McCarty, K. T. Poole, and H. Rosenthal. *Polarized America: The Dance of Ideology and Unequal Riches*. MIT Press, 2006.

T. Mitchell and J. Beauchamp. Bayesian variable selection in linear regression (with discussion). *Journal of the American Statistical Association*, 83:1023–36, 1988.

S. Press. *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference (2nd edition)*. New York: Krieger, 1982.

K. M. Quinn. Bayesian factor analysis for mixed ordinal and continuous responses. *Political Analysis*, 12:338–353, 2004.

B. Rajaratnam, H. Massam, and C. M. Carvalho. Flexible covariance estimation in graphical Gaussian models. *The Annals of Statistics*, 36(6):2818–2849, 2008.

P. E. Rossi, G. M. Allenby, and R. McCulloch. *Bayesian statistics and marketing*. Series in Probability and Statistics. Wiley, 2006.

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.

J. G. Scott and J. O. Berger. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 2010. to appear.

X. Song and S. Lee. Bayesian estimation and test for factor analysis model with continuous and polytomous data in several populations. *British Journal of Mathematical and Statistical Psychology*, 54:237–63, 2001.

X.-Y. Song and S.-Y. Lee. A multivariate probit latent variable model for analyzing dichotomous responses. *Statistica Sinica*, 15:645–664, 2005.

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.
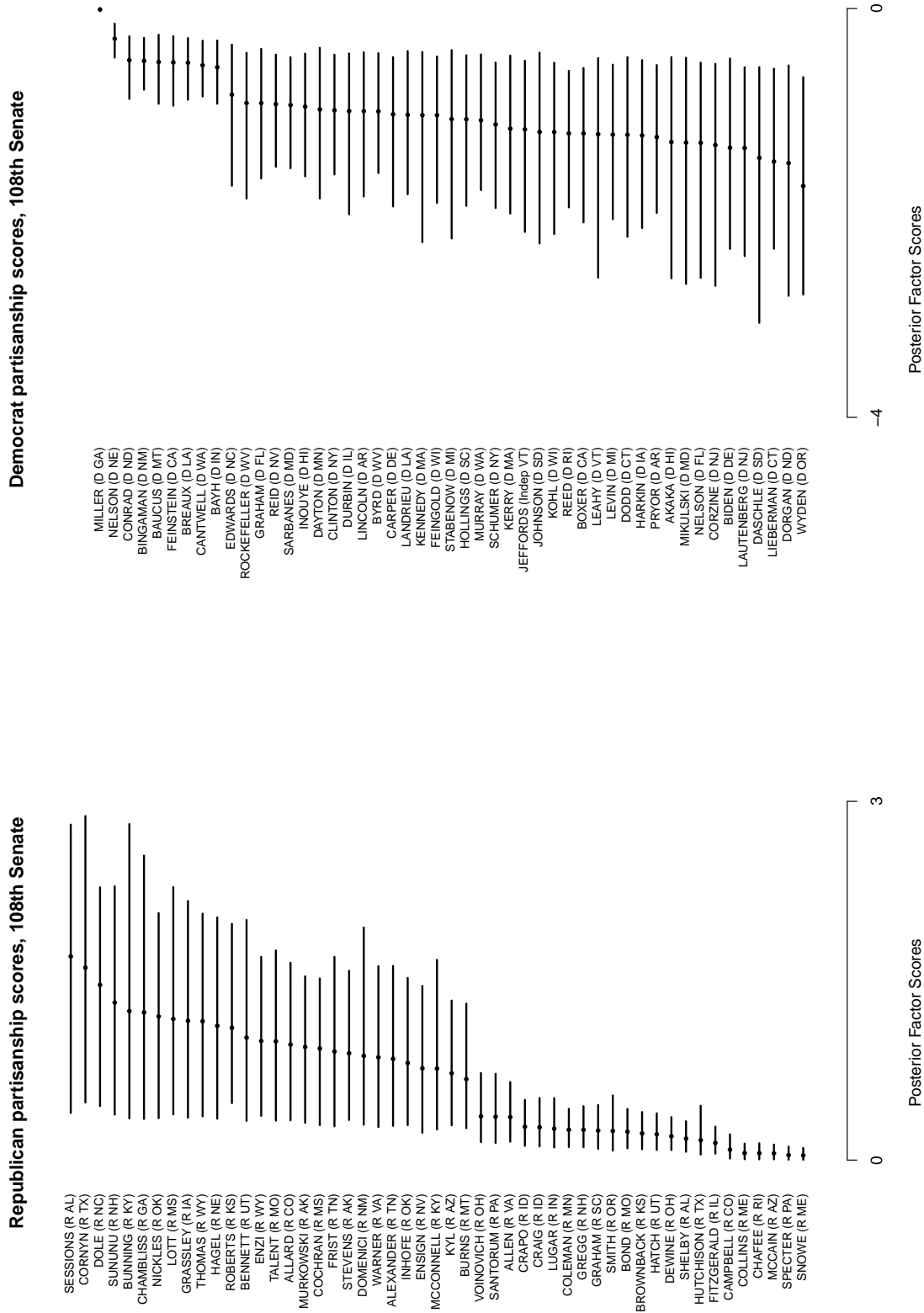
Figure 3: Posterior mean partisanship scores are shown here for each Senator of the 108th Congress, along with associated 95% credible intervals.