Sparse Factor-Analytic Probit Models

BY JAMES G. SCOTT

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

PAUL R. HAHN

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

AND CARLOS M. CARVALHO

Graduate School of Business, University of Chicago,
Chicago, Illinois 60637, U.S.A.
carlos.carvalho@chicagogsb.edu

SUMMARY

We describe a class of sparse factor-analytic probit models for binomial and multinomial data. These models have two main advantages over traditional Bayesian probit models. First, they lead to more stable covariance estimation in high-dimensional settings, particularly where \( p \) is large compared to \( n \). Second, they can provide direct inference about lower-dimensional structure in data, which often has a useful subject-specific interpretation. We demonstrate these advantages both on synthetic data and then on two real data sets: one involving consumer preferences among different brands of Scotch whisky, and one that xyz. In both examples, we show how the inferred latent structure corresponds to plausible market forces. We also include a lengthy discussion of priors in sparse-factor models generally; the issues we consider here are crucial for probit models, but also have strong methodological implications for inference in factor models on continuous data.

Some key words: multivariate probit models; covariance estimation; factor models.
1 Introduction

2 Factor-analytic probit models

2.1 The multivariate probit model

Suppose we observe binary data $Y = (y_1 \ldots y_n)^t$, where each $y_i = (y_{i,1}, \ldots, y_{i,p})^t$ represents $p$ correlated Bernoulli trials for a single subject. The $n$ subjects might be, for example, participants in a marketing survey, with $y_{ij}$ indicating products the subjects have purchased over the last year; see Edwards and Allenby (2003) for just such an application.

The multivariate probit model (Chib and Greenberg, 1998) allows one to characterize the observed binary data $y$ in terms of an unobserved continuous quantity $z$, which is related to the Bernoulli success probability by the probit link function:

\[
y_{ij} \sim \text{BER}(r_{ij}) \\
r_{ij} = \Pr(z_{ij} > 0) \\
z_i \sim \text{N}(\alpha, \Sigma)
\]

where $\Sigma$, for identifiability reasons, must be in correlation form. To see this, note that if $z_i \sim \text{N}(\beta, \Sigma)$ is a parameterization in terms of covariances rather than correlations, then defining $D = \text{diag}(\tau_{11}^{-1/2}, \ldots, \tau_{pp}^{-1/2})$, $\Sigma = D\Sigma D$, and $\alpha = D\beta$ gives the same Bernoulli probabilities: $\Pr(y_{ij} = 1 \mid \alpha, \Sigma) = \Pr(y_{ij} = 1 \mid \beta, \Sigma)$. We later make use of this fact in post-processing our MCMC output (which is from the nonidentified space) to yield estimates of model parameters.

The latent vector $z$ can depend upon a linear function of covariates $X$; without loss of generality we denote this linear predictor by $\alpha$.

2.2 The multinomial probit model

If instead we observe subjects making one choice out of a finite set of $p$ possibilities, we have a multinomial model, where $y_i \in \{0, 1, \ldots, p-1\}$ indicates the choice made by subject $i$. Again, we use the probit link to relate the categorial variable $y$ to an underlying continuous quantity $z$ (which is now of dimension $p-1$). This model can be written as in McCulloch and Rossi (1994) and McCulloch et al. (2000):

\[
y_i = \begin{cases} 
0 & \text{if } \max z_i < 0 \\
j & \text{if } \max z_i = z_{ij} > 0
\end{cases} \\
z_i \sim \text{N}(\alpha, \Sigma)
\]
where, as above, $\Sigma$ must be a correlation matrix and $\alpha$ can be a linear predictor that depends upon covariates.

### 2.3 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 $z$ vector. In a typical covariance estimation problem with observed continuous data, information about $\Sigma$ would come from the cross-product matrix $Z'Z$. Even when the $z_i$ are observed directly, this can yield notoriously unstable estimates 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). Frustratingly, 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 $z$ is the largest (or that all elements are negative).

With less information about each $z_i$, covariance estimation becomes more difficult. This motivates a simplifying factor structure, which improves estimability by constraining and regularizing $\Sigma$. Upon positing some fixed number $m$ of factors $f$ along with a $p \times m$ matrix of factor loadings $B$, we can write the latent variables for observation $i$ as:

$$
\begin{align*}
    z_i &= \alpha + Bf_i + \nu_i \\
    \nu_i &\sim N(0, \Psi) \\
    f_i &\sim N(0, I)
\end{align*}
$$

from which it follows that $\Sigma = BB' + \Psi$. Note that $\Psi$ is forced to be diagonal—either $\Psi = \sigma^2 I$ for a common scale among all variables, or $\Psi = \text{diag}(\psi_1, \ldots, \psi_p)$ for a more general model. We write $B'_j$ and $b_k$ 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 $z$ can be explained by mutual dependence upon a lower-dimensional vector $f$, whose components vary independently a priori. This dependence structure is encoded in $B$, 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

A factor model must be further constrained for the identity $\Sigma = BB' + \Psi$ to have a unique solution, i.e. for $B$ to be identifiable. These identification issues, discussed extensively in Aguilar (1998), 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 $B^* = \Gamma B$ and $f^*_i = \Gamma f_i$ give the same model for any orthogonal matrix $\Gamma$.

Two traditional solutions to this problem involve forcing $B$ to be orthogonal or forcing $B'\Psi B$ to be diagonal. We instead follow Geweke and Zhou (1996) and constrain $B$ to be zero for upper-triangular entries $\{b_{jk} : k > j\}$. This guarantees that $B$ is the unique matrix for which $BB'$ is positive-definite, and makes the choice of the $k$ leading variables (referred to as the founders of the factors) an important decision that directly impacts the fit and interpretation of the model.

3 Sparsity

We now describe a novel sparse factor-analytic probit model, where some of the unconstrained elements in the factor-loadings matrix $B$ can be identically 0. Introduced by West (2003) and further developed by Carvalho et al. (2008) in the context of gene-expression data, these models assume that each latent factor will be associated with only a small number of observed variables, yielding a more parsimonious covariance structure with even stronger regularization properties.

In sparse factor models, the pattern of non-zero elements in $B$ is unknown and must be estimated from the data. These models have typically taken the form:

\[
\begin{align*}
(b_{jk} | v_k, q_k) &\sim q_k \cdot N(0, v_k) + (1 - q_k)\delta_0 \quad (4) \\
v_k &\sim IG(a_v/2, b_v/2) \quad (5) \\
q_k &\sim Be(1, 1) \quad (6)
\end{align*}
\]

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 (Scott and Berger, 2008), thereby solving the implicit multiple-testing problem of simultaneously deciding whether to include each of hundreds, or even thousands, of possible nonzero entries in $B$.

But the prior on $v_k$, the variance of the nonzero factor loadings for factor $k$, can be tricky. For one thing, the prior must be proper, since this is a model-selection problem with $v_k$ not appearing in all models (specifically, it is missing from the model in which column $k$ of $B$ contains all zeros). West (2003), Carvalho et al. (2008), and others get around this problem by using the specification above, with an inverse-gamma
prior with fixed hyperparameters $a_v$ and $b_v$. Yet these priors suffer from the fact that $a_v$ and $b_v$ strongly influence the proportion of nonzero factor loadings, since these hyperparameters directly influence the Bayes factor for comparing $H_0 : b_{jk} = 0$ versus $H_A : b_{jk} \neq 0$. Hence $a_v$ and $b_v$ must either be chosen subjectively, which yields a subjective (and possibly questionable) fraction of nonzero factor loadings; or they must be tuned from the data, which some may view as philosophically dubious.

In our view, a better solution is to ensure that the prior for $v_k$ satisfies two criteria: (1) that it depends upon $\Psi$, which has an interpretation as the error variance for the latent linear predictor of $z$ and thus provides the only natural, objective scale for the problem; and (2) that it yields a marginal prior distribution for $b_{jk}$ with no moments (or at least has heavy tails if this is not possible). These two criteria have ample justification in the hypothesis-testing literature. They were proposed by Jeffreys (1961), who advocated Cauchy priors for testing normal means, and are also reflected in the recommended multiple-testing priors from Scott and Berger (2006). The scale-matching property is important for avoiding Lindley’s paradox (Lindley, 1957) and similar counter-intuitive behaviors, which could happen with careless choices of the inverse-gamma hyperparameters above; while the heavy-tailedness property is justified in detail by Liang et al. (2007).

In sparse factor models for continuous data such as in West (2003), these issues are not too distressing. Heavy-tailed priors certainly yield nice theoretical properties, but lighter-tailed versions often perform just fine in practice. Hence one can simply choose $a_v$ and $b_v$ to yield appropriate ad hoc scale-matching to the observed data without worrying too much about theoretical niceties.

For our factor model, however, the scale-matching problem is far more serious, because there is no intrinsic scale to the latent $z_{ij}$’s. When we sample from the posterior distribution of $\mathbf{B}$ and $\Psi$, we operate in a space that is not likelihood-identified and rely upon post-processed MCMC output to yield estimates of the mean vector $\alpha$ and the correlation matrix of latent variables (both of which are identified). This is a standard computational trick in multivariate probit models to avoid the need for restricting $\Sigma$ to be a correlation matrix “on the fly” (Chib and Greenberg, 1998; McCulloch et al., 2000). Yet as a result, $\Psi$ is not constrained by the data and instead can wander rather freely around the parameter space—even while maintaining the identifiability of the correlation matrix corresponding to $\mathbf{BB}^t + \Psi$.

This will stymie any attempt to choose fixed values of $a_v$ and $b_v$ that roughly match the scale of the latent “error variance.” Instead, the prior on $v_k$ must depend directly on $\Psi$, so that when $\Psi$ wanders around during posterior sampling, the scale of the prior distribution for nonzero factor loadings will match it step-for-step, always yielding a reasonable test of the null hypothesis that $b_{jk} = 0$.

To see this problem more clearly, it helps to look at the implied marginal likelihoods for testing $H_0 : b_{jk} = 0$ versus $H_A : b_{jk} \neq 0$. Consider only a single component $z_j$ of the latent vector $\mathbf{z}$, which we can assume for the moment is actually observed. The factor model says that $z_j = \alpha_j + B_j^t f + \nu_j$, with $\nu_j \sim N(0, \Psi_j)$. 

5
Let \( \tilde{z}_j^{(k)} \) stand for the residual \( z_j - \alpha_j - \sum_{l \neq k} b_{jl} f_l \). We know that \( (\tilde{z}_j^{(k)} \mid \Psi_j) \sim N(0, \Psi_j) \) under \( H_0 \), or alternatively \( (\tilde{z}_j^{(k)} \mid v_k, \Psi_j) \sim N(0, f_k^2 v_k + \Psi_j) \) under \( H_A \). If \( v_k \sim IG(a/2, b/2) \), then the marginal likelihood under \( H_A \) is not available in closed form. Yet to get intuition about the effect of the prior, one can imagine simulating draws from the predictive distributions of these two hypotheses:

- To make a draw under \( H_0 \), set \( \tilde{z}_j^{(k)} = \epsilon_1 \), with \( \epsilon_1 \sim N(0, \Psi_j) \).
- To make a draw under \( H_A \), set \( \tilde{z}_j^{(k)} = \epsilon_1 + \epsilon_2 \), with \( \epsilon_1 \sim N(0, \Psi_j) \) and \( \epsilon_2 \sim T_a(0, bf_k^2/a) \). Here \( T_a(m, s) \) represents a \( t \) distribution (with obvious notation).

The above expressions make it clear how \( a \) and \( b \) affect the relative scales of the two predictive distributions, and thus the marginal likelihoods of the two hypotheses for particular observed \( \tilde{z}_j^{(k)} \). Fixed choices of \( a \) and \( b \) make this ratio of scales—and hence the Bayes factor between the two models—completely arbitrary in a probit model, since \( \Psi_j \) itself is arbitrary.

We now describe three possible solutions to this scale-matching problem.

1. Choose \( \Psi = \sigma^2 I, \) reflecting a common “residual variance” (after regressing upon the factors) along every dimension of the latent \( z \) vector. Setting \( \pi(\sigma^2) \propto 1/\sigma^2 \) and \( (v_k \mid \sigma^2) \sim IG(1/2, \sigma^2/2) \) then yields an appropriately scaled prior on the nonzero factor loadings.

2. Allow \( \Psi = \text{diag}(\sigma^2 \psi_1, \ldots, \sigma^2 \psi_p) \) and, as above, choose \( (v_k \mid \sigma^2) \sim IG(1/2, \sigma^2/2) \). This model is not identified, but one can choose a noninformative prior on \( \sigma^2 \) in conjunction with sharp priors on each \( \psi_j \) to yield a model that is weakly identified in the Bayesian sense. For example, setting \( \psi_j \sim IG(\kappa + 2, \kappa + 1) \) yields a prior mean of 1 and variance of \( 1/\kappa \). This gives \( \sigma^2 \) an interpretation as a baseline variance, with individual \( \psi_j \)'s inflating or shrinking component-wise variances with respect to this baseline. The hyperparameter \( \kappa \) can thus be chosen to reflect the expected scale of this heterogeneity.

3. Allow \( \Psi = \text{diag}(\psi_1, \ldots, \psi_p) \), except instead of choosing a common variance parameter for each column of the factor loadings matrix as above, now choose a common variance for each row of this matrix:

\[
(b_{jk} \mid q_k, v_j) \sim q_k \cdot N(0, v_j) + (1 - q_k)\delta_0 \quad (7)
\]
\[
(v_j \mid \Psi_j) \sim IG(1/2, \Psi_j/2) \quad (8)
\]
\[
\pi(\Psi_j) \propto 1/\Psi_j \quad (9)
\]
\[
q_k \sim \text{Be}(1, 1) \quad (10)
\]

When it is appropriate to assume a common scale of residual variance across all dimensions, the first option should be perfectly adequate. If this is not realistic,
then either the second or the third should yield better answers. We prefer (and use in this paper) the third option because of its simplicity, and because it does not require specifying an expected scale for deviations from the baseline residual variance. The second option may be attractive in some cases, however, since it preserves the interpretation from Carvalho et al. (2008) of hyper-variances as parameters adhering to latent factors rather than to observed variables.

4 Model Fitting

We employ a Gibbs sampler to draw MCMC samples from the joint posterior distribution (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. See the discussion in McCulloch et al. (2000).

1. Draw the latent observation matrix $Z = (z_{ij})$ by drawing each $z_{ij} \sim N(\mu_j + B_j'f_i, 1)$ 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 $z$.

2. Sample $\mu$.

3. Sample the vector-valued elements of $F$ independently as $(f_i | z_i) \sim N(B'[BB' + I]^{-1}(z_i - \mu), I - B'[BB' + I]^{-1}B)$.

4. Sample the unconstrained elements of $b_k$, column $k$ of the factor loadings matrix, by defining $z_{ij}^* = z_{ij} - \sum_{l=1,l\neq k}^{m} B_{l,j} f_{k,i}$. Then sample $b_{jk} \sim (1 - \hat{q}_{jk})\delta_0 + \hat{q}_{jk} \cdot N(\hat{b}_{jk}, \hat{v}_{jk})$ where

$$\hat{v}_{jk} = \left( \sum_{i=1}^{n} f_{k,i}^2 + v_k^{-1} \right)^{-1}$$

$$\hat{b}_{jk} = \hat{v}_{jk} \left( \sum_{i=1}^{n} f_{k,i} z_{ij}^* \right)$$

$$\hat{q}_{ik} = \frac{N(0 \mid 0, v)}{N(0 \mid \tilde{b}_{ik}, \hat{q}_{ik})} N(0 \mid \hat{b}_{ik}, \hat{q}_{ik}) \frac{1}{1 - \hat{q}_{ik}}$$

5. Let $\tilde{s}_k$ be the maximum number of non-zero elements in column $b_k$ (those not explicitly set to zero by design), and let $s_k$ be the number of these elements
currently set to zero. Then we have
\[ v_k \sim IG\left((1 + s_k)/2, (1 + b'_k b_k)/2\right) \]
by the standard conjugate updates.

6. And finally, also by conjugacy, \( q_k \sim Be(1 + s_k, 1 + \tilde{s}_k - s_k) \).

5 Performance on Simulated Data

6 Examples

6.1 Multivariate binary data

Scotch.

6.2 Multinomial data

7 Discussion

New priors for sparsity component: we emphasize that this is not simply a “fix” for a probit-specific issue, but rather a general methodological recommendation for sparse factor models on all kinds of data.

References


