Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Partial Factor Modeling: Predictor-Dependent Shrinkage for Linear Regression

P. Richard Hahn [a], Carlos M. Carvalho [b] & Sayan Mukherjee [c]

[a] Booth School of Business, University of Chicago, Chicago, IL, 60637

[b] McCombs School of Business, The University of Texas, Austin, TX, 78712

[c] Departments of Statistical Science, Computer Science, Mathematics, and Institute for Genome Sciences Policy, Duke University, Durham, NC, 27708

Accepted author version posted online: 29 Mar 2013.Published online: 27 Sep 2013.

To cite this article: P. Richard Hahn, Carlos M. Carvalho & Sayan Mukherjee (2013) Partial Factor Modeling: Predictor-Dependent Shrinkage for Linear Regression, Journal of the American Statistical Association, 108:503, 999-1008, DOI: 10.1080/01621459.2013.779843

To link to this article: http://dx.doi.org/10.1080/01621459.2013.779843

Partial Factor Modeling: Predictor-Dependent Shrinkage for Linear Regression

P. Richard HAHN, Carlos M. CARVALHO, and Sayan MUKHERJEE

We develop a modified Gaussian factor model for the purpose of inducing predictor-dependent shrinkage for linear regression. The new model predicts well across a wide range of covariance structures, on real and simulated data. Furthermore, the new model facilitates variable selection in the case of correlated predictor variables, which often stymies other methods.

KEY WORDS: g Prior; Prediction; Shrinkage estimators; Variable selection.

1. INTRODUCTION

This article considers modifications to a Gaussian factor model, which make it better suited for regression and variable selection. These purpose-specific modifications make the new model of interest not only as a new Bayesian factor model (Aguilar and West 2000; Lopes 2003; West 2003), but also as a novel regularized regression technique and as a new model for Bayesian variable selection. Our model differs from previous work on Bayesian variable selection (Mitchell and Beauchamp 1988; George and McCulloch 1997; Clyde and George 2004) in that it explicitly accounts for predictor correlation structure.

The factor regression framework may be written in two parts, as a linear regression model for a scalar response variable Y_i and as a marginal model for a p-dimensional column vector of predictor variables $\mathbf{X}_i$. Specifically, for Gaussian factor models

$$(Y_i \mid \mathbf{X}_i, \boldsymbol{\beta}, \sigma^2) \sim \mathrm{N}\big(\mathbf{X}_i^t \boldsymbol{\beta}, \sigma^2\big), \tag{1}$$

with marginal predictor model

$$\begin{aligned} \mathbf{X}_i &= \mathbf{B}\mathbf{f}_i + \boldsymbol{\nu}_i, \quad \boldsymbol{\nu}_i \sim \mathrm{N}(0, \boldsymbol{\Psi}) \\ \mathbf{f}_i &\sim \mathrm{N}(0, \mathbf{I}_k), \end{aligned} \tag{2}$$

where $\mathbf{B}$ is a $p \times k$ real-valued matrix. If $\boldsymbol{\Psi}$ is assumed diagonal, the elements of $\mathbf{X}_i$ are conditionally independent given the k-vector of common unobserved (latent) factors $\mathbf{f}_i$, so that $\boldsymbol{\nu}_i$ represents idiosyncratic errors. Without loss of generality we assume throughout that our response and predictor variables are centered at zero.

This article asks the question: how should the prior on $\boldsymbol{\beta}$ depend on $\mathbf{B}$ and $\boldsymbol{\Psi}$? Two prevailing approaches represent extreme answers to this question. At one extreme, a pure linear regression model ignores the marginal distribution of the predictors, $\pi(\mathbf{X})$, entirely, which is equivalent to setting $\pi(\boldsymbol{\beta} \mid \mathbf{B}, \boldsymbol{\Psi}) = \pi(\boldsymbol{\beta})$.

At the other extreme, a pure factor model approach assumes that each Y_i depends linearly on the same k latent factors that capture the covariation in $\mathbf{X}_i$ and is conditionally independent of $\mathbf{X}_i$ given these factors: $\pi(Y_i \mid \mathbf{X}_i, \mathbf{f}_i, \boldsymbol{\theta}) = \pi(Y_i \mid \boldsymbol{\theta}, \mathbf{f}_i)$ and more specifically $\mathrm{E}(Y_i \mid \boldsymbol{\theta}, \mathbf{f}_i) = \boldsymbol{\theta}\mathbf{f}_i$, where $\boldsymbol{\theta}$ denotes a $1 \times k$ vector of regression coefficients. This entails that $\boldsymbol{\beta}$ can be written as a deterministic function of $\mathbf{B}$ and $\boldsymbol{\Psi}$:

$$\boldsymbol{\beta}^t = \boldsymbol{\theta}\mathbf{B}^t(\mathbf{B}\mathbf{B}^t + \boldsymbol{\Psi})^{-1}. \tag{3}$$

This very strong prior can lead to poor estimates of $\boldsymbol{\beta}$ if the chosen k is too small; specifically it may be mistakenly inferred that the response is entirely uncorrelated with *any* of the predictors if Y_i is in fact independent of $\mathbf{B}^t(\mathbf{B}\mathbf{B}^t + \boldsymbol{\Psi})^{-1}\mathbf{X}_i$, the projection of the predictors onto the k-dimensional subspace defined by $\mathbf{B}$ and $\boldsymbol{\Psi}$. An analogous problem in principal component regression is well known; the *least-eigenvalue scenario* is when the response is associated strongly with only the least important principal component (Hotelling 1957; Cox 1968; Jolliffe 1982).

Placing a prior over k, allowing the number of factors to be learned from the data, is an obvious way around this difficulty. However, specifying a prior over k that respects the goal of prediction within the framework of the joint distribution is nontrivial (see the example in Section 1.2). The joint likelihood is dominated by predictor matrix $\mathbf{X}$, even if our practical goal is to use the $\mathbf{X}_i$ to predict the corresponding Y_i. Our solution is instead to construct a hierarchical model, which is centered at the Bayesian factor regression model (conditional on some fixed number of factors):

$$\mathrm{E}(\boldsymbol{\beta}^t \mid \mathbf{B}, \boldsymbol{\Psi}, \boldsymbol{\theta}) = \boldsymbol{\theta}\mathbf{B}^t(\mathbf{B}\mathbf{B}^t + \boldsymbol{\Psi})^{-1}. \tag{4}$$

Permitting deviations from the pure factor model safeguards inference against misspecification in the sense that for any fixed values of $\mathbf{B}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\theta}$, our prior on $\boldsymbol{\beta}$ has full support in $\mathbb{R}^p$, unlike the point-mass prior of the usual factor model (Equation (3)).

The rest of the article is organized as follows. The remainder of this section briefly reviews previous work and demonstrates the challenges of factor model selection in terms of obtaining good regression parameter estimates. Section 2 develops the new partial factor model in detail and compares its out-of-sample prediction performance to common alternatives, such

P. Richard Hahn is Assistant Professor of Econometrics and Statistics, Booth School of Business, University of Chicago, Chicago, IL 60637 (E-mail: *richard.hahn@chicagobooth.edu*). Carlos M. Carvalho is Associate Professor of Statistics, McCombs School of Business, The University of Texas, Austin, TX 78712 (E-mail: *carlos.carvalho@mccombs.utexas.edu*). Sayan Mukherjee is Associate Professor of Statistics, Departments of Statistical Science, Computer Science, Mathematics, and Institute for Genome Sciences Policy, Duke University, Durham, NC 27708 (E-mail: *sayan@stat.duke.edu*). P. Richard Hahn thanks Dan Merl for helpful discussions. Carlos M. Carvalho thanks the McCombs School of Business Research Excellence Grant. Sayan Mukherjee acknowledges AFOSR FA9550-10-1-0436, NSF DMS-1045153, and NSF CCF-1049290 for partial support. Sayan Mukherjee acknowledges Mike West, Anirban Bhattacharya, and David Dunson for useful comments. The authors thank the referees for helpful suggestions.

Journal of the American Statistical Association
September 2013, Vol. 108, No. 503, Theory and Methods
DOI: 10.1080/01621459.2013.779843

as ridge regression, partial least squares, principal component regression (Hastie, Tibshirani, and Friedman 2001), and the lasso (Tibshirani 1996). Section 3 adapts the new model for the purpose of variable selection in the presence of correlated predictors. Section 4 concludes with a brief discussion about further connections and generalizations.

1.1 Bayesian Linear Factor Models

We briefly provide details of a typical Bayesian linear factor model. Any multivariate normal distribution may be written in *factor form* as in (2). As above, $\mathbf{B}$ is a $p \times k$ matrix and $\mathbf{\Psi}$ is assumed diagonal. The matrix $\mathbf{B}$ is referred to as a loadings matrix, the elements of $\mathbf{\Psi}$ are referred to as idiosyncratic variances, and the $\mathbf{f}_i$ are called as factor scores. Conditional on $\mathbf{B}$ and $\mathbf{f}_i$, the elements of each observation are independent. Integrating over $\mathbf{f}_i$, we see

$$\operatorname{cov}(\mathbf{X}_i) \equiv \mathbf{\Sigma}_X = \mathbf{B}\mathbf{B}^t + \mathbf{\Psi}. \tag{5}$$

When $k \geq p - 1$, this form is unrestricted in that any positive definite matrix can be written as (5). We say that a positive definite matrix admits a k-factor form if it can be written in factor form $\mathbf{B}\mathbf{B}^t + \mathbf{\Psi}$, where $\operatorname{rank}(\mathbf{B}) \leq k$. Note that $\mathbf{B}\mathbf{B}^t + \mathbf{\Psi}$ has full rank whenever the idiosyncratic variances are strictly positive, while $\mathbf{B}$, which encodes the covariance structure, may have much lower rank.

If we further assume that the p predictors influence each response Y_i only through the k-dimensional latent variable $\mathbf{f}_i$, we arrive at the Bayesian factor regression model:

$$\begin{aligned} Y_i &= \boldsymbol{\theta}\mathbf{f}_i + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2) \\ \mathbf{\Sigma} = \operatorname{cov}\begin{pmatrix} \mathbf{X}_i \\ Y_i \end{pmatrix} &= \begin{bmatrix} \mathbf{B}\mathbf{B}^t + \mathbf{\Psi} & \mathbf{V}^t \\ \mathbf{V} & \xi \end{bmatrix}, \\ \mathbf{V} &= \boldsymbol{\theta}\mathbf{B}^t, \\ \xi &= \sigma^2 + \boldsymbol{\theta}\boldsymbol{\theta}^t. \end{aligned} \tag{6}$$

As the norm of $\mathbf{\Psi}$ goes to zero, this model recovers singular value regression (West 2003). Again, $\boldsymbol{\theta}$ is a $1 \times k$ row vector; effectively it is an additional row of the loadings matrix ($\boldsymbol{\theta} \equiv \mathbf{b}_{p+1}$ and $Y_i = \mathbf{X}_{p+1,i}$).

Factor models have been a topic of research for over a century, with increased recent interest spurred by the ready availability of computational implementations. A seminal reference is Spearman (1904); Press (1982) and Bartholomew and Moustaki (2011) are key modern references. Bayesian factor models for continuous data have been developed by many authors, including Geweke and Zhou (1996) and Aguilar and West (2000). A thorough bibliography can be found in Lopes (2003). Notable applications include finance (Chamberlain 1983; Chamberlain and Rothschild 1983; Fama and French 1992, 1993; Aguilar and West 2000; Bai 2003; Lopes and Carvalho 2007; Fan, Fan, and Lv 2008) and gene expression studies (Carvalho et al. 2008; Merl et al. 2009; Lucas et al. 2012). The area continues to see new methodological developments focusing on a variety of issues: prior specification (Ghosh and Dunson 2009), model selection (Lopes and West 2004; Bhattacharya and Dunson 2011), and identification (Fruhwirth-Schnatter and Lopes 2012). In this work, we highlight the use of factor models for prediction and variable selection.

1.2 The Effects of Misspecifying k

If k is chosen too small, inferences can be unreliable as a trivial consequence of misspecification. Less appreciated, however, is that minute misspecifications in terms of overall (joint) model fit can drastically impair the suitability of the regression induced by the joint model. The following example demonstrates that the evidence provided by the data may be indifferent between two-factor models that differ only by the presence of one factor, even though the larger model is clearly superior in terms of prediction.

Example 1. Consider the 10-dimensional two-factor Gaussian model with loadings matrix

$$\mathbf{B}^t = \begin{bmatrix} 0 & -4 & 0 & -8 & -4 & -6 & 1 & -1 & 4 & 0 \\ 1 & 0 & 0 & -1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$

and idiosyncratic variances $\psi_{jj} = 0.2$ for all $j \in \{1, \ldots, p\}$. Now consider the one-factor model that is closest in Kullback–Leibler divergence to this model, with loadings matrix

$$\begin{aligned} \mathbf{A}^t = [&0.0004 \quad -3.9967 \quad 0 \quad -7.9713 \quad -3.9967 \\ &-5.9778 \quad 0.9990 \quad -0.9960 \quad 3.9967 \quad -0.0004] \end{aligned}$$

and idiosyncratic variances given by the vector

$$\begin{aligned} \mathbf{D} = [&1.2000 \quad 0.1871 \quad 0.2000 \quad 1.5032 \quad 0.1871 \quad 1.3762 \\ &0.1996 \quad 1.2054 \quad 0.1872 \quad 1.2000]. \end{aligned}$$

Observe that the one-factor loadings matrix $\mathbf{A}$ is very nearly equal to the first factor of $\mathbf{B}$, but that the idiosyncratic variances are notably different. In particular, consider the problem of using the one-factor approximation to predict future observations of the 10th dimension of X, which does not load on the first factor. The true idiosyncratic variance is $\psi_{10} = 0.2$, but the approximate model has $D_{10} = 1.2$, suggesting that prediction on this dimension will be inaccurate. However, as measured by the joint likelihood, the one-factor model is an excellent approximation. These mismatched conclusions are reflected in Figure 1, which plots the difference in mean squared prediction error (MSPE) between the two models against the difference in log-likelihood; each point represents a realization of 10 observations. Above zero on the vertical axis favors the true model, while below zero favors the one-factor approximation. The horizontal axis represents approximation loss due to the missing factor. The average likelihood ratio is approximately one, while prediction performance is always worse with the smaller model.

More importantly, this discrepancy does not fade as more data are used. With only 10 observations, the likelihood ratio favors the true model only 47% of the time; with 100 observations this number creeps up to 51% and at 1000 observations it stays at 51%. By the likelihood criterion the two models are nearly identical. However, in terms of predicting X_{10}, the one-factor approximation is literally useless: the inferred conditional and marginal variances are virtually identical.

Thus, we see that relying on a prior distribution to correctly choose between a one- versus two-factor model is a difficult task: the prior would have to be strong enough to overwhelm more than 1000 observations worth of evidence, which favors

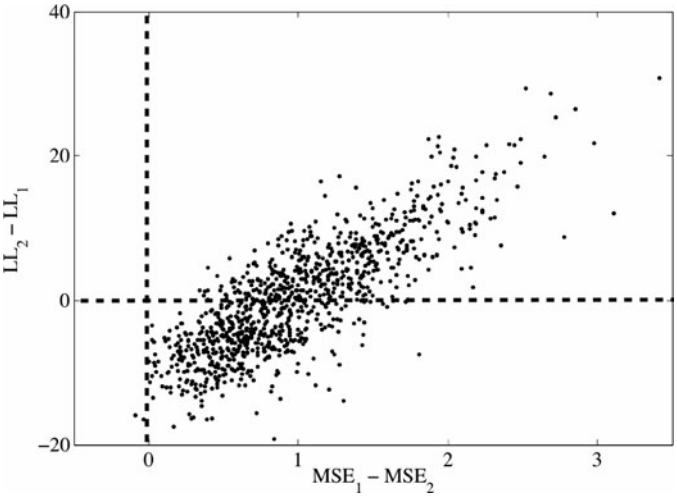

Figure 1. Points denote realizations from the true two-factor model. For points above the dashed horizontal line, the likelihood ratio favors the true model. The distance to the right of the dashed vertical line measures how much worse than the true model the one-factor approximation did in terms of predicting X_{10}. Model selection based on the full likelihood favors the larger model half the time, while model selection based on predictive fit favors the larger model nearly always.

the wrong model about half the time. Moreover, to work appropriately it would need to be a prior, which favored *more* factors a priori. Observe also that model averaging does not improve the situation, as inclusion probabilities would hover around 1/2 for each model, resulting in a prediction halfway between the prediction of the correct model and zero, the prediction of the incorrect model; the problem grows worse as the number of possible values of k increases.

By contrast, a cross-validation approach would uncover the predictive superiority of the two-factor model directly. While a joint distribution allows one to borrow information from the marginal predictor distribution, which may be useful for prediction, using an unmodified high-dimensional joint distribution subordinates the prediction task to the more difficult task of high-dimensional model selection.

These difficulties persist even with the use of sophisticated nonparametric model selection priors for factor models (Bhattacharya and Dunson 2011), because they are logically distinct from any particular prior, the problem lying rather with the assumption that the latent factors $\mathbf{f}_i$ explain *all* of the variability in *both* the predictors and the responses. That is, the problem lies in the particular latent variable representation rather than the prior over k—it is the implied prior over the regression coefficients *as a function of the observed predictors* that is ultimately at issue. This issue is clearly illustrated in the simulation study of Section 2.1.

In the next section, we surmount this obstacle directly, by relaxing the assumption that the latent factors capturing the predictor covariance are sufficient for predicting the response.

2. PARTIAL FACTOR REGRESSION

Our new model—referred to here as the *partial factor* model—circumvents the prior specification difficulties described in the previous section by positing a lower-dimensional covariance structure for the predictors, but permitting the relationship between the predictors and the response to be linear in up to the full p dimensions. This is achieved by using the following covariance structure for the joint normal distribution:

$$\begin{pmatrix} \mathbf{X}_i \\ Y_i \end{pmatrix} \sim \mathrm{N}(0, \boldsymbol{\Sigma})$$
$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{B}\mathbf{B}^t + \boldsymbol{\Psi} & \mathbf{V}^t \\ \mathbf{V} & \xi \end{bmatrix}. \tag{7}$$

The difference between (6) and (7) is simply that in (7) $\mathbf{V}$ is not required to exactly equal $\boldsymbol{\theta}\mathbf{B}^t$. The matrix $\mathbf{B}$ is still a $p \times k$ matrix with $k < p - 1$ so that the $p \times p$ predictor covariance matrix is constrained to the $\mathbf{B}\mathbf{B}^t + \boldsymbol{\Psi}$ form, but the full covariance matrix $\boldsymbol{\Sigma}$ is not simultaneously restricted. This way, the response can depend on directions in predictor space, which are not dominant directions of variability, but inference and prediction still benefit from structural regularization of $\boldsymbol{\Sigma}_X$.

Just as crucially, the prior on $\mathbf{V}$ may be conditioned on $\boldsymbol{\Sigma}_X$. Specifically, we may suggest via the prior that higher variance directions in predictor space are more apt to be predictive of the response. Unlike principal component regression or factor models, the prior furnishes this bias as a hint rather than a rigid assumption.

The hierarchical specification arises from the jointly normal distribution between $\mathbf{X}_i$, Y_i, and the k latent factors, which have covariance:

$$\operatorname{cov}\begin{pmatrix} \mathbf{X}_i \\ \mathbf{f}_i \\ Y_i \end{pmatrix} = \begin{bmatrix} \mathbf{B}\mathbf{B}^t + \boldsymbol{\Psi} & \mathbf{B}^t & \mathbf{V}^t \\ \mathbf{B} & \mathbf{I}_k & \boldsymbol{\theta}^t \\ \mathbf{V} & \boldsymbol{\theta} & \xi \end{bmatrix}. \tag{8}$$

Again, recall that $\mathbf{V}$ is not constrained as in (6). From this covariance, the conditional moments of the response can be expressed as

$$\mathrm{E}(Y_i \mid \mathbf{f}_i, \mathbf{X}_i) = \boldsymbol{\theta}\mathbf{f}_i + \left\{(\mathbf{V} - \boldsymbol{\theta}\mathbf{B}^t)\boldsymbol{\Psi}^{-\frac{1}{2}}\right\}\left\{\boldsymbol{\Psi}^{-\frac{1}{2}}(\mathbf{X}_i - \mathbf{B}\mathbf{f}_i)\right\} \tag{9}$$
$$\operatorname{var}(Y_i \mid \mathbf{f}_i, \mathbf{X}_i) = \xi - [\mathbf{V} \quad \boldsymbol{\theta}]\boldsymbol{\Sigma}_{X,f}^{-1}[\mathbf{V} \quad \boldsymbol{\theta}]^t \quad \equiv \sigma^2. \tag{10}$$

To center at a standard factor model, our prior for $\mathbf{V}$, conditional on $\boldsymbol{\theta}$, $\mathbf{B}$ and $\boldsymbol{\Psi}$ is

$$v_j \sim \mathrm{N}\big(\boldsymbol{\theta}\mathbf{B}^t, \omega^2 w_j^2 \psi_j^2\big), \tag{11}$$

for element j, implying that a priori the error piece plays no role in the regression. Here, ω^2 is a global prior variance and w_j^2 is a predictor-specific prior variance, following work on robust shrinkage priors (Carvalho, Polson, and Scott 2010); ψ_j^2 is the jth diagonal element of $\boldsymbol{\Psi}$. This effect is perhaps easiest to see via the following reparameterization: define

$$\boldsymbol{\Lambda} = (\mathbf{V} - \boldsymbol{\theta}\mathbf{B}^t)\boldsymbol{\Psi}^{-\frac{1}{2}}. \tag{12}$$

Then

$$\lambda_j \sim \mathrm{N}\big(0, w^2\omega_j^2\big) \tag{13}$$

is equivalent to (11).

Using this formulation (along with some additional hyperparameters), the model may be expressed as

$$\begin{aligned}
\mathbf{X}_i \mid \mathbf{B}, \mathbf{f}_i, \boldsymbol{\Psi} &\sim \mathrm{N}(\mathbf{B}\mathbf{f}_i, \boldsymbol{\Psi}) \\
Y_i \mid \mathbf{X}_i, \mathbf{B}, \boldsymbol{\theta}, \boldsymbol{\Lambda}, \mathbf{f}_i, \boldsymbol{\Psi}, \sigma^2 &\sim \mathrm{N}\left(\boldsymbol{\theta}\mathbf{f}_i + \boldsymbol{\Lambda}\left\{\boldsymbol{\Psi}^{-\frac{1}{2}}(\mathbf{X}_i - \mathbf{B}\mathbf{f}_i)\right\}, \sigma^2\right) \\
\lambda_j &\sim \mathrm{N}\left(0, \omega^2 w_j^2\right), \\
\mathbf{f}_i &\sim \mathrm{N}(0, \mathbf{I}_k) \\
\theta_h &\sim \mathrm{N}\left(0, \tau^2 q_h^2\right) \\
b_{jh} &\sim \mathrm{N}\left(0, \tau^2 t_{jh}^2\right), \quad h = 1, \ldots, k, \\
& \qquad\qquad\qquad\quad j = 1, \ldots, p.
\end{aligned} \tag{14}$$

Independent half-Cauchy priors are placed over τ and ω and the individual elements of the vectors $\mathbf{t}$, $\mathbf{w}$, and $\mathbf{q}$. This corresponds to the so-called horseshoe priors (Carvalho, Polson, and Scott 2010) over the elements of $\mathbf{B}$, $\boldsymbol{\theta}$, and $\boldsymbol{\Lambda}$, respectively. The residual standard deviations (σ and each element of $\boldsymbol{\Psi}^{1/2}$) are given Strawderman–Berger priors with density $p(s) \propto s(1+s^2)^{-3/2}$. The model can be fit using a Gibbs sampling approach; computational details are deferred to Appendix A.

Appendix B describes a brief simulation study focusing on how priors over the elements of $\boldsymbol{\Lambda}$ affect posterior estimates of $\boldsymbol{\beta}$.

2.1 Out-of-Sample Prediction Simulation Study

This section considers the predictive performance of the partial factor model relative to the two models between which it strikes a compromise: a pure linear regression model and a full factor model. To anticipate the results below, the partial factor model predicts as well as a factor model and outperforms a pure regression model when the factor structure is informative of the response, and it predicts as well as a pure regression model and outperforms the factor model when the factor structure is only weakly predictive. This profile is consistent with that of the multiple-shrinkage principal component regression model of George and Oman (1996), which has a similar motivation—seeking to mimic principal component regression but to protect against the least-eigenvalue scenario—but is not derived from a joint sampling model.

For this simulation study, we draw $\mathbf{X}_i$ from a $k = 10$ factor model with $p = 80$ and $n = 50$. We simulated 100 datasets according to the following specifications. For $j = 1, \ldots, p$ and $g = 1, \ldots, k$

$$\begin{aligned}
\mathbf{B} &\equiv \mathbf{A}\mathbf{D} \\
a_{j,g} &\sim \mathrm{N}(0, 1) \\
d_g &\equiv 1 + |\epsilon_g|, \text{ s.t. } |d_g| \geq |d_{g'}| \text{ if } g < g', \\
\epsilon_g &\sim \mathrm{t}(0, df = 5) \\
\psi_j &= \sqrt{\mathbf{b}_j \mathbf{b}_j^t / u_j}, \quad u_j \sim \mathrm{Unif}(1/4, 15),
\end{aligned} \tag{15}$$

where $\mathbf{D}$ is a k-by-k diagonal matrix with diagonal elements d_g. This procedure allows direct control over the signal-to-noise ratio for each dimension of $\mathbf{X}_i$ in terms of u_j. We similarly draw Y_i from the factor model (6), considering two cases. In the first case, the first and most dominant factor (in the sense of $|d_{g,g}|$ being largest) is solely predictive of Y_i:

$$\boldsymbol{\theta} = (1, 0, \ldots, 0).$$

In the second case, the least dominant factor is the one that is solely predictive of Y_i:

$$\boldsymbol{\theta} = (0, 0, \ldots, 1).$$

In each case, we set $\sigma = 1/5$ (for a 5-to-1 signal-to-noise ratio), so that if relevant the factor can be accurately inferred, it is highly predictive of the response.

In addition to the partial factor model, we fit a factor model employing the model selection prior of Bhattacharya and Dunson (2011), and a pure regression model using the horseshoe prior of Carvalho, Polson, and Scott (2010). We compare the out-of-sample prediction in terms of two metrics, first the relative (to optimal) MSPE

$$\mathrm{MSPE}(\beta_{\mathrm{est}}) = \frac{\mathrm{E}\left((Y_{n+1} - \mathbf{X}_{n+1}\beta_{\mathrm{est}})^2\right)}{\mathrm{E}\left((Y_{n+1} - \mathbf{X}_{n+1}\beta_{\mathrm{true}})^2\right)}, \tag{16}$$

where the expectation is taken over $(Y_{n+1}, \mathbf{X}_{n+1})$. The denominator can be computed analytically using the known parameter values; the numerator is computed via Monte Carlo simulations. Second, we record how frequently a given method performed better than the other two, which we denote by Pr(optimal).

For this simulation, the number of factors in the partial factor model was chosen using the following heuristic: either chose the value between 1 and n that gives the largest difference in consecutive singular values of $\mathbf{X}$, or 3, whichever is *smaller*. We do not advocate this heuristic in general; rather this simulation study highlights the strength of the partial factor model in mitigating the price one pays in terms of predictive degradation when the number of factors is underestimated. This robustness allows one to safely choose the lowest plausible number of factors based on subject matter knowledge or to use a crude heuristic without risking leaving a lot of predictive power on the table. Within this context, choosing a particular value of k induces a certain prior on the implied regression coefficients, $\boldsymbol{\beta}$. Incorporating model selection priors such as Bhattacharya and Dunson (2011) within the partial factor framework is a potentially fruitful line of future research.

Different simulation schemes will highlight the strengths of different methods; the scheme described here serves to communicate what can go wrong when one banks too heavily on factor structure being predictive of a response variable. Numerical results are displayed in Tables 1 and 2. Because the factor structure is always present in this simulation ($k = 10$ being much less than $p = 80$), we observe that the pure regression model does not adequately capitalize on this structure. However, when it happens that the factor structure is less strongly predictive, the strong bias of the factor model can degrade predictions more than necessary relative to the pure regression model. So, even in the favorable case where the factor model is best 50% of

Table 1. Case one: when the factor structure is highly predictive of the response, the partial factor model performs on par with the learned factor model

Method	MSPE	Pr(optimal)
Partial factor regression	1.31	0.33
Bhattacharya et al.	1.33	0.58
Carvalho et al.	1.49	0.09

Table 2. Case two: when the factor structure is less predictive of the response, partial factor regression performs on par with the pure regression model, while the full factor model suffers dramatic overshrinkage

Method	MSPE	Pr(optimal)
Partial factor regression	1.59	0.54
Bhattacharya et al.	5.86	0.41
Carvalho et al.	1.84	0.05

the time, when it is not best it can be far from optimal. This pattern becomes more pronounced in the less favorable case where the prediction error can be dramatically suboptimal. The simulation results demonstrate that the partial factor model successfully avoids this pitfall, while still capitalizing on strong factor structure when it is evident.

2.2 Out-of-Sample Prediction Applied Example

In this section, we extend our comparisons to additional methods and to the case of real data. We compare partial factor regression to five other methods: principal component regression, partial least squares, lasso regression (Tibshirani 1996), ridge regression, and unadjusted Bayesian factor modeling using the model selection prior of Bhattacharya and Dunson (2011). We observe the same pattern of robust prediction performance as in the simulation study. Partial factor regression shows itself to be the best or nearly the best among the methods considered in terms of out-of-sample MSPE.

Five real datasets in the $p > n$ regime are analyzed; the data are available from the R packages `pls` (Mevik and Wehrens 2007), `chemometrics` (Varmuza and Filzmoser 2009), and `mixOmics` (Cao, Gonzalez, and Dejean 2009). These data were selected because they are publicly available and fall within the $p > n$ regime that is most germane to our comparisons.

To test the methods, each of the datasets is split into training and test samples, with 75% of the observations used for training. Each model is then fit using the training data, with tuning parameters for the four non-Bayesian methods chosen by ten-fold cross-validation on only the training data. Out-of-sample predictive performance on the holdout data is measured by sum of squared prediction error.

As shown in Table 3, the partial factor model outperforms the other models on three of the five datasets and is never much worse than the best in the remaining cases. By comparison the final column shows that the unadjusted Bayesian factor regression using a modern factor selection prior does very poorly at recovering a satisfactory regression model for prediction; its tendency to radically overshrinkage is illustrated by the yarn data shown in row three.

3. SPARSITY PRIORS FOR VARIABLE SELECTION

In this section, we consider adapting the partial factor model for the purpose of variable selection. Variable selection is a pervasive problem in applied statistics, see, for example, George and Foster (2000), Liang et al. (2008), and references therein. Notable Bayesian approaches to this problem in the canonical normal linear regression setup include Zellner (1971), George and McCulloch (1997), Liang et al. (2008), Clyde and George (2004), among many others. With the additional assumption that the predictors and the data come from a joint normal distribution, the variable selection problem becomes a question of inferring exactly zero entries in (a particular row of) the associated precision matrix $\mathbf{\Sigma}_{X,Y}^{-1}$, bringing the problem into the territory of Gaussian graphical models (Dempster 1972; Speed and Kiiveri 1986; Dawid and Lauritzen 1993). Previous work has also advocated covariance regularization for variable selection problems (Jeng and Daye 2011).

Here, intuitive representations of the regression vector $\boldsymbol{\beta}$ will follow directly from the generative structure of the partial factor model and prove helpful in a variable selection context. The guiding intuition is simply this: if a subset of predictors is mutually dependent, then they should either all be in or out of the model as a group.

Consider the parameterization of the partial factor model given in Equation (12), whence the latent regression (conditional on $\mathbf{f}_i$) follows as

$$Y_i = \boldsymbol{\theta}\mathbf{f}_i + \boldsymbol{\Lambda}\boldsymbol{\Psi}^{-\frac{1}{2}}(\mathbf{X}_i - \mathbf{B}\mathbf{f}_i) + \epsilon_i, \qquad \epsilon_i \sim \mathrm{N}(0, \sigma^2). \quad (17)$$

Fixing $\boldsymbol{\Lambda}$ to a vector of zeros represents a pure factor model. If $\lambda_j = 0$, predictor X_j appears in the regression of Y_i only via its dependence on the latent factors.

Accordingly, positive a priori probability of zero elements in $\boldsymbol{\Lambda}$, $\boldsymbol{\theta}$, and $\mathbf{B}$ implies positive probability of zero elements in $\boldsymbol{\beta}$. For instance, a point-mass prior at zero (sometimes called a two-groups model or a spike-and-slab model) on the jth element of $\boldsymbol{\Lambda}$ can be written as

$$\pi(\lambda_j \mid \alpha_\lambda) = \alpha_\lambda \mathrm{N}\left(\lambda_j; 0, \omega^2 w_j^2\right) + (1 - \alpha_\lambda)\delta_0(\lambda_j), \quad (18)$$

Table 3. PFR: partial factor regression. RIDGE: ridge regression. PLS: partial least squares. LASSO: lasso regression. PCR: principal component regression. BFR: Bayesian factor regression

			Average out-of-sample error					
Dataset	n	p	PFR	RIDGE	PLS	LASSO	PCR	BFR
Nutrimouse	40	120	377.3 (27%)	**296.2**	418.5 (41%)	492.3 (66%)	391.2 (32%)	517.73 (75%)
Cereal	15	145	**31.8**	41.86 (32%)	51.30 (61%)	42.97 (35%)	45.01 (42%)	62.34 (96%)
Yarn	28	268	**0.29**	0.50 (72%)	0.37 (28%)	0.30 (3%)	0.42 (45%)	7.80 (260%)
Gasoline	60	401	**0.54**	0.68 (26%)	0.74 (37%)	0.81 (50%)	0.70 (30%)	0.71 (31%)
Multidrug	60	853	183.0 (18%)	**154.0**	164.0 (6%)	220.4 (43%)	170.6 (11%)	212.13 (38%)

NOTE: Percentages shown are amount worse than the best method, reported in bold type.

where δ_0 denotes a point mass at zero. Analogous priors are placed on the elements of $\mathbf{B}$ and $\boldsymbol{\theta}$, with corresponding hyperparameters α_b and α_θ:

$$\begin{aligned}\pi(\theta_h \mid \alpha_\theta) &= \alpha_\theta \mathrm{N}\left(\theta_h; 0, \tau^2 q_h^2\right) + (1-\alpha_\theta)\delta_0(\theta_h),\\ \pi(\beta_{jh} \mid \alpha_\beta) &= \alpha_\beta \mathrm{N}\left(\beta_{jh}; 0, \tau^2 t_{jh}^2\right) + (1-\alpha_\beta)\delta_0(\beta_{jh}).\end{aligned} \tag{19}$$

Sparsity of $\boldsymbol{\beta}$ is then induced via the identity

$$\boldsymbol{\beta}^t = \left(\boldsymbol{\theta} - \boldsymbol{\Lambda}\boldsymbol{\Psi}^{-\frac{1}{2}}\mathbf{B}\right)\mathbf{B}^t(\mathbf{B}\mathbf{B}^t + \boldsymbol{\Psi})^{-1} + \boldsymbol{\Lambda}\boldsymbol{\Psi}^{-\frac{1}{2}}. \tag{20}$$

Unlike the prediction context, which we used to motivate the partial factor model, in the variable selection setting, identifiability of $\mathbf{B}$ becomes a relevant concern. A complete investigation of the identification issues associated with linear factor models is Fruhwirth-Schnatter and Lopes (2012). Presently, we develop the sparse partial factor model on the working assumption that inferences concerning $\mathbf{B}$ are being handled in an appropriate fashion; for our simulation study we adopt the convenient lower-triangular restriction of Geweke and Zhou (1996).

Zeros in $\boldsymbol{\beta}$ arise only when $\boldsymbol{\Sigma}_X$ has a block diagonal structure (or can be permuted to have such a structure) via elements of $\mathbf{B}$ being exactly zero. Each block then represents a subset of the predictors, which is jointly independent of the remaining elements, and each block can be associated with its own set of latent factors. For d such groups, we have

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_1 & 0 & 0 & \cdots \\ 0 & \mathbf{B}_2 & 0 & \cdots \\ 0 & 0 & \ddots & \\ \vdots & \vdots & & \mathbf{B}_d \end{bmatrix}, \quad \boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{\Psi}_1 & 0 & 0 & \cdots \\ 0 & \boldsymbol{\Psi}_2 & 0 & \cdots \\ 0 & 0 & \ddots & \\ \vdots & \vdots & & \boldsymbol{\Psi}_d \end{bmatrix}, \tag{21}$$

and $\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}_1 \ \boldsymbol{\Lambda}_2 \ \ldots \ \boldsymbol{\Lambda}_d)$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1 \ \boldsymbol{\theta}_2 \ \ldots \ \boldsymbol{\theta}_d)$ represent conformable partitions. From these definitions it follows straightforwardly that $\boldsymbol{\beta}$ also partitions as $\boldsymbol{\beta}^t = (\boldsymbol{\beta}_1^t \ \boldsymbol{\beta}_2^t \cdots \boldsymbol{\beta}_d^t)$, with each group being defined as in (20):

$$\boldsymbol{\beta}_l^t = \left(\boldsymbol{\theta}_l - \boldsymbol{\Lambda}_l\boldsymbol{\Psi}_l^{-\frac{1}{2}}\mathbf{B}_l\right)\mathbf{B}_l^t\left(\mathbf{B}_l\mathbf{B}_l^t + \boldsymbol{\Psi}_l\right)^{-1} + \boldsymbol{\Lambda}_l\boldsymbol{\Psi}_l^{-\frac{1}{2}}. \tag{22}$$

This equivalence follows from the fact that the inverse of a block diagonal matrix is the block diagonal matrix composed of the inverses of the original block components. From this expression, one observes that the regression coefficient group $\boldsymbol{\beta}_l$ is a zero vector precisely when $\boldsymbol{\Lambda}_l$ and $\boldsymbol{\theta}_l$ are both zero vectors. If any one element of these subvectors is nonzero, then the whole block becomes nonzero via their interdependence. The partial factor model gives nonzero prior probability to an exactly zero regression coefficient for a given predictor only if the response variable is independent of any latent factors governing that predictor ($\boldsymbol{\theta}_l = 0$) and is also unrelated to that predictor residually ($\boldsymbol{\Lambda}_l = 0$).

Arguably this approach to sparsity is more intuitive than a pure regression model in cases of correlated predictors in the following sense. In a linear regression, a zero coefficient in $\boldsymbol{\beta}$ may arise (in principle) if a given variable has (say) a negative effect, but is correlated with another variable having an *exactly* countervailing effect. The partial factor model of sparsity assigns zero prior probability to such implausible balancing acts.

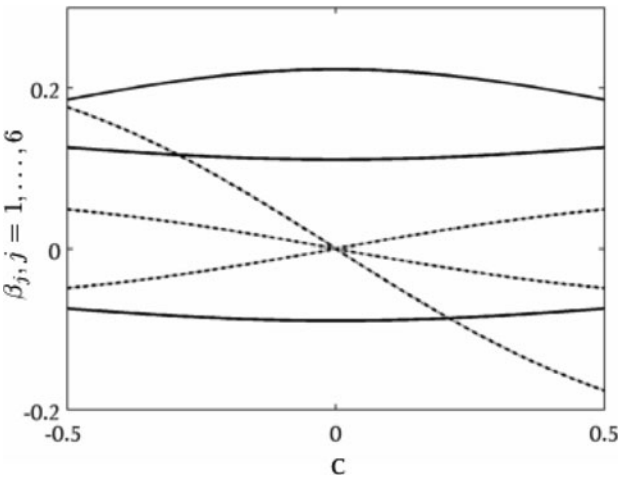

Figure 2. Elements of $\boldsymbol{\beta}_1$ are solid. Elements of $\boldsymbol{\beta}_2$ are dashed. $\boldsymbol{\Lambda} = (0 \ 0.1 \ 0 \ 0 \ 0 \ 0)$. Note that the elements of $\boldsymbol{\beta}_2$ all converge at the origin.

Instead, the induced sparsity in $\boldsymbol{\beta}$ has to happen "directly" via sparsity of $\boldsymbol{\theta}$, $\mathbf{B}$, and $\boldsymbol{\Lambda}$.

Example 2. Consider a partial factor regression (7) with the following parameters:

$$\begin{aligned}\mathbf{B}^t &= \begin{bmatrix} -0.25 & 0.1 & 0.1 & c & 0 & 0 \\ 0 & 0 & 0 & -0.8 & 0.2 & -0.2 \end{bmatrix}, \quad \boldsymbol{\Psi} = 0.5\mathbf{I}\\ \Lambda &= (0 \quad 0.1 \quad 0 \quad 0 \quad 0 \quad 0), \quad \theta = -0.5 \quad 0).\end{aligned} \tag{23}$$

Now consider how $\boldsymbol{\beta}$ changes as a function of c, noting that when $c = 0$ we have the desired block independence of $\mathbf{X}_i$. Consider the following partitioning: $\boldsymbol{\theta} = (\theta_1 \, \theta_2)$, $\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}_1 \quad \boldsymbol{\Lambda}_2)$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1 \quad \boldsymbol{\beta}_2)$. Specifically $\boldsymbol{\Lambda}_1 = (\Lambda_1 \quad \Lambda_2 \quad \Lambda_3)$, $\boldsymbol{\Lambda}_2 = (\Lambda_4 \quad \Lambda_5 \quad \Lambda_6)$ and $\boldsymbol{\beta}_1 = (\beta_1 \quad \beta_2 \quad \beta_3)^t$, $\boldsymbol{\beta}_2 = (\beta_4 \quad \beta_5 \quad \beta_6)^t$. We see that because $\theta_2 = 0$ and $\boldsymbol{\Lambda}_2 = 0$, the only way elements of $\boldsymbol{\beta}_2$ become nonzero is via nonzero c. When $c = 0$, all the elements of $\boldsymbol{\beta}_2$ equal zero, as seen in Figure 2. Information from predictor dimensions four through six becomes informative about the response via correlation with predictor dimensions one through three; in this example, when $c = 0$ there is no such informative correlation.

Observe, if we keep the same setup as above, but set $\boldsymbol{\Lambda} = (0 \quad 0.1 \quad 0 \quad 0 \quad 0 \quad 0.25)$, the picture changes dramatically, as shown in Figure 3. This perhaps is counterintuitive because λ_6 represents the "residual" dependence of the sixth predictor variable, but we must remember that this interpretation is conditional on the latent factors. Because in practice the latent factors are unobserved and must be inferred from the data, nonzero elements of $\boldsymbol{\Lambda}_2$ dictate that the response depends on the corresponding predictors in complicated ways defined by expression (20). The partial factor model provides the following new functionality: if the data provide evidence that λ_j is nonzero, the posteriori probability that β_j is exactly zero becomes correspondingly less likely.

3.1 Variable Selection Simulation Study

To demonstrate the effectiveness of the sparse factor model for variable selection, we compare it to three alternatives. First

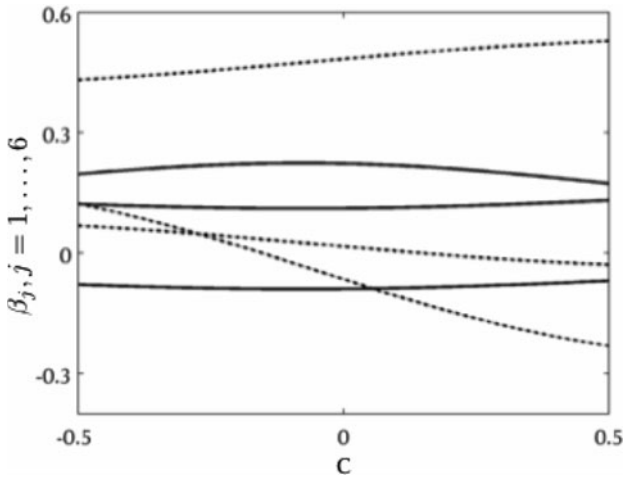

Figure 3. Elements of $\boldsymbol{\beta}_1$ are solid. Elements of $\boldsymbol{\beta}_2$ are dashed. The convergence at the origin noted in the previous plot is no longer evident now that $\Lambda = (0 \;\; 0.1 \;\; 0 \;\; 0 \;\; 0 \;\; 0.25)$.

is a two-groups regression model with an independent prior on $\boldsymbol{\beta}$:

$$\begin{aligned} \pi(\beta_j \mid v_j, v) &= \frac{1}{2}\delta_0(\beta_j) + \frac{1}{2}\mathrm{N}\left(0, \omega^2 w_j^2\right), \\ \omega &\sim \mathrm{C}^+(0, 1), \\ w_j &\sim \mathrm{C}^+(0, 1). \end{aligned} \tag{24}$$

This prior is an amalgam of the usual two-groups model (with normal density for the continuous portion) and a shrinkage model called the horseshoe (Carvalho, Polson, and Scott 2010).

The second model we consider is a two-groups model with a g prior on the nonzero regression coefficients:

$$\pi(\beta_{\mathcal{M}} \mid \mathbf{X}, \mathcal{M}) = \mathrm{N}\big(0, g^{-1}\big(\mathbf{X}_{\mathcal{M}}\mathbf{X}_{\mathcal{M}}^t\big)^{\dagger}\big), \tag{25}$$

where $\mathcal{M}$ picks out which elements of the regression are nonzero and $\dagger$ denotes the Moore–Penrose pseudoinverse. This model was fit using the R package BAS (Liang et al. 2008; Clyde, Ghosh, and Littman 2011). Under this model the prior inclusion probability α is given a uniform hyperprior; the hyper g prior setting is used with the recommended default hyperparameters.

Third, we use the elastic net model described in Zou and Hastie (2005), which compromises between a lasso l_1 penalty and a ridge l_2 penalty. Like the partial factor model, the elastic net has the property that related predictors will group in and out of the model together (see Zou and Hastie 2005, sec. 2.3).

For the partial factor model, the individual elements of $\boldsymbol{\theta}$, $\mathbf{B}$, and $\boldsymbol{\Lambda}$ are given horseshoe point-mass priors as in (24). However, different prior probabilities are chosen (fixed): $\alpha_b = 0.5$, $\alpha_\lambda = 0.1$, and $\alpha_\theta = 0.9$ (for all elements of the indicated parameter). These choices encode the prior beliefs that any given predictor is as likely to depend on any given latent factor as not, but that it is much more likely that Y depends on the common latent factors than it is to depend on any of the individual predictors residually.

We compare performance according to three metrics:

- the usual mean squared error (SE) of regression coefficient estimates: $p^{-1}\sum_j(\beta_j - \hat{\beta}_j)^2$,
- the average misclassified error (ME): $p^{-1}\sum_j(\mathbb{1}\{\beta_j \neq 0\} - \mathbb{1}\{\hat{\alpha}_{\beta_j} > \frac{1}{2}\})^2$,
- and the squared difference between the posterior inclusion probability and the true model indicator vector, or the "probability estimate" (PE) error: $p^{-1}\sum_j(\mathbb{1}\{\beta_j \neq 0\} - \hat{\alpha}_{\beta_j})^2$,

where $\hat{\alpha}_{\beta_j}$ denotes the posterior probability of inclusion. The mean SE measures overall goodness of the estimate, the average misclassification error measures performance explicitly in terms of variable selection and the mean squared probability error provides a notion of sharpness to accompany the misclassification error.

Table 4. SE: squared error. ME: misclassification error. PE: probability estimation error. The mean value across simulations is reported for each type of error. The blank element in the table is due to the elastic net not providing inclusion probability estimates

Method	SE	ME	PE
Sparse partial factor	0.10	0.10	0.08
Variable selection g-prior	0.11	0.51	0.32
Two-groups horseshoe	0.10	0.44	0.24
Elastic net	0.15	0.41	–

Data were generated according to a sparse factor model with k drawn uniformly at random on $\{0, \ldots, 5\}$. For this study $p = 20$ and $n = 50$. Elements of $\boldsymbol{\Lambda}$ are drawn from independent normal distributions with mean zero and variance 0.25, and the elements of $\boldsymbol{\theta}$ are drawn from independent normal distributions with mean zero and variance 4. This situation corresponds to there being strong factor structure, which is predictive of the response. Results are depicted in Table 4.

The partial factor model soundly outperforms these three widely used methods on our simulated data in terms of misclassification error. The g prior model and the two-groups model give similar SE estimation performance to the partial factor model, but the variable selection performance of the partial factor model is markedly better. For these data, patterns of covariation in the predictors are informative about the sparsity of $\boldsymbol{\beta}$. The independent two-groups model does not account for these patterns and so performs worse. The g prior model incorporates these patterns via the prior covariance used for $\boldsymbol{\beta}$, but potentially underregularizes low-variance directions in the observed data; this point is discussed further in Section 4.1. This high variability similarly causes the SE of the g prior model to be somewhat (10%) higher than the independent two-groups model. Meanwhile, the elastic net model outperforms the g prior and the two-groups model on misclassification error, but is by far the worst in terms of mean SE.

4. CONNECTIONS AND EXTENSIONS

4.1 Relation to Ridge Regression and Zellner's g Prior

The partial factor model offers robust model-based regularized linear regression. By biasing a full p-dimensional regression toward a low-dimensional factor model, the partial factor regression framework represents a principled compromise between least-square regression and full factor modeling. Indeed, from the viewpoint of the implied prior over the regression

parameters $\boldsymbol{\beta}$, there arise interesting connections to ridge regression and to Zellner's g prior (Zellner 1986; Liang et al. 2008), two classical linear regression methods. For ease of comparison, in what follows set the prior variances $q_h^2\tau^2 = \omega^2 w_j^2$ for all h, fix $w_j = 1$ for all j, fix the residual variance at $\sigma^2 = 1$, and assume $\mathbf{X}\mathbf{X}^t$ is invertible (similar expressions arise using a generalized inverse). Then, from expression (20) one finds that the prior variance of $\boldsymbol{\beta}$ in terms of $\mathbf{B}$ and $\boldsymbol{\Psi}$ is

$$\begin{aligned} \operatorname{cov}(\boldsymbol{\beta}) &= \omega^2(\mathbf{B}\mathbf{B}^t + \boldsymbol{\Psi})^{-1}, \\ &= \omega^2 \boldsymbol{\Sigma}_X^{-1}, \end{aligned} \tag{26}$$

from which the posterior mean may be expressed as

$$\mathrm{E}(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{X}, \boldsymbol{\Sigma}_X) = \left(\omega^{-2}\mathbf{I}_p + \boldsymbol{\Sigma}_x^{-1}\mathbf{X}\mathbf{X}^t\right)^{-1} \boldsymbol{\Sigma}_X^{-1}\mathbf{X}\mathbf{Y}. \tag{27}$$

By comparison, the normal prior used in ridge regression has prior variance $\operatorname{cov}(\boldsymbol{\beta}) = \omega^2\mathbf{I}_p$ and yields the estimator

$$\mathrm{E}_{\text{ridge}}(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{X}) = (\mathbf{X}\mathbf{X}^t + \omega^{-2}\mathbf{I}_p)^{-1}\mathbf{X}\mathbf{Y}. \tag{28}$$

And finally, the g prior is a normal prior with $\operatorname{cov}(\boldsymbol{\beta}) = g^{-1}(\mathbf{X}\mathbf{X}^t)^{-1}$, yielding the posterior estimator

$$\mathrm{E}_{\text{Zellner}}(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{X}) = (1+g)^{-1}(\mathbf{X}\mathbf{X}^t)^{-1}\mathbf{X}\mathbf{Y}. \tag{29}$$

The differences between these expressions are instructive. It is straightforward to show that the ridge estimator downweights the contribution of the directions in (observed) predictor space with lower sample variance, from which one may argue that (Hastie, Tibshirani, and Friedman 2001):

> ridge regression protects against the potentially high variance of gradients estimated in the short directions. The implicit assumption is that the response will tend to vary most in the directions of high variance in the inputs.

The g prior, by contrast, shrinks $\boldsymbol{\beta}$ more in directions of high sample variance in the predictor space a priori, which has the net effect of shrinking the orthogonal directions of the design space equally regardless of whether the directions are long or short. This reflects the substantive belief that higher variance directions in predictor space need not influence the response variable more than the directions of lower variance.

However, these rationales conflate the observed design space with the pattern of stochastic covariation characterizing the random predictor variable. That is, we want the effects of ridge regression to "[protect] against the potentially high variance of gradients estimated in the short directions" but we would like to do so without having to assume that "the response will tend to vary most in the directions of high variance in the inputs." Teasing apart these two aspects of the problem can be done by conditioning on $\mathbf{X}$ and $\boldsymbol{\Sigma}_x \equiv \operatorname{cov}(\mathbf{X}_i)$ individually, which is what the partial factor model does. This teasing apart may be observed directly from the form of (27). Because $\boldsymbol{\Sigma}_X$ and $\mathbf{X}\mathbf{X}^t/n$ are not in general identical, we still get shrinkage in different directions, thus combatting the "high variance of gradients estimated in short directions" while not having to assume that any direction in predictor space is more or less important a priori.

Put another way, one may consider the g prior an approximate version of the partial factor prior, which uses $\mathbf{X}\mathbf{X}^t$ as a plug-in estimate of Σ_X. From this vantage, the benefit of the partial factor model becomes clear, in that it properly handles the not-inconsiderable uncertainty associated with this often high-dimensional parameter.

4.2 Beyond the Linear Model

Inference and prediction can often be improved by making structural simplifications to a statistical model. In a Bayesian framework, this can be accomplished by positing lower-dimensional latent variables that govern the joint distribution between predictors and the response variable, facilitating "borrowing information." An immediate downside to this approach is that specifying high-dimensional joint distributions and priors is difficult, particularly in terms of modulating the degree of regularization implied for a given conditional density. The partial factor model addresses this problem by parameterizing the joint sampling model using a compositional form, which allows the conditional regression to be handled independently of the marginal predictor distribution. Specifically, this formulation of the joint distribution realizes borrowing of information via a hierarchical prior rather than through a fixed structure.

The partial factor model applies these ideas in the classic setting of a joint normal distribution for the purpose of regularized linear regression, but the conceptual underpinnings are readily extended. For instance, it is straightforward to extend the method to a binary or categorical response variable Z_i by treating the continuous response Y_i as an additional latent variable (Albert and Chib 1993). So, if Z_i is a binary response variable, one can write

$$Z_i = \mathbb{1}(Y_i < 0),$$

where Y_i is modeled as in (17), inducing a partial factor probit model for Z_i conditional on the vector of predictors $\mathbf{X}_i$.

In fact, the idea of using a compositional representation in conjunction with a hierarchical prior can be profitably extended to many joint distributions by specifying the compositional form explicitly at the initial stages of the modeling process. For example, the conditional expectation of the response under the partial factor model given in (17) suggests the following nonlinear generalization:

$$\mathrm{E}(Y_i \mid \mathbf{f}_i, \mathbf{X}_i) = \phi\left(\mathbf{f}_i, \mathbf{X}_i - \mathrm{E}(\mathbf{X}_i \mid \mathbf{f}_i)\right) \tag{30}$$

for an unknown function ϕ. A more modest nonlinear generalization would be to assume additivity:

$$\mathrm{E}(Y_i \mid \mathbf{f}_i, \mathbf{X}_i) = \phi(\mathbf{f}_i) + \varphi(\mathbf{X}_i - \mathrm{E}(\mathbf{X}_i \mid \mathbf{f}_i)), \tag{31}$$

where ϕ and φ denote smooth functions to be inferred from the data. Crucially, the smoothness assumptions for ϕ and φ can be chosen individually. In particular, it would be interesting to consider simple parametric forms for ϕ while allowing φ to be a smooth nonparametric function.

APPENDIX A: COMPUTATIONAL IMPLEMENTATION

The strategy for sampling from the posterior distribution is essentially a Gibbs sampler; the full conditional sampling steps are given below. We use the convention that a dash to the right of the conditioning bar should be read as "everything else."

1. Sample the latent factors: $(\mathbf{F} \mid -)$. Using the joint normal distribution of $\mathbf{f}_i$, $\mathbf{X}_i$, and Y_i, we draw $\mathbf{f}_i \sim \mathrm{N}(\mu_i, S)$, where

$$\mu_i = (\mathbf{B}^t \quad \boldsymbol{\theta}^t)\Sigma_{X,Y}^{-1}(\mathbf{X}_i^t \quad Y_i)^t$$
$$S = \mathbf{I}_k - (\mathbf{B}^t \ \boldsymbol{\theta}^t)\Sigma_{X,Y}^{-1}(\mathbf{B}^t \ \boldsymbol{\theta}^t)^t.$$

This form comes directly from a block partition of the covariance matrix in Equation (8); applications of the Sherman–Woodbury–Morrison identity yield a more cumbersome expression that is amenable to efficient inversion.

2. Sample variance components. All of the variance components have the same update, differing only in how we calculate the "residuals." This step is based on the slice sampler described in Damien, Wakefield, and Walker (1999). Each of these updates is described in terms of random variables r_l, $l = 1, \ldots, m$, which are distributed independently as $\mathrm{N}(0, s^2)$ with prior density on the variance given by $p(s) \propto s^{2a-1}(1+s^2)^{-(a+1/2)}$. For $a = 1/2$, this is a half-Cauchy density and corresponds to the horseshoe prior on the r_l; for $a = 1$ it corresponds a Strawdeman–Berger prior on r_l. Define $\eta = 1/s^2$. Then we sample s as follows.

- Draw $(u \mid \eta) \sim \mathrm{Uniform}(0, (1+\eta)^{-(a+1/2)})$.
- Draw $(\eta \mid r, u) \sim \mathrm{Gamma}((m+1)/2, \sum_{l=1}^m r_i^2/2)$ restricted to be below $u^{-(1/(a+1/2))} - 1$.
- Set $s = \eta^{-1/2}$.

(a) Sample $(\boldsymbol{\Psi} \mid -)$. Let $m = n$ and for each dimension $j = 1, \ldots, p$, define $r_l(j) = X_{jl} - \mathbf{b}_j \mathbf{f}_l$.
(b) Sample $(\sigma \mid -)$. Let $m = n$ and define $r_l = Y_l - \boldsymbol{\theta}\mathbf{f}_l - \boldsymbol{\Lambda}\boldsymbol{\Psi}^{-\frac{1}{2}}(\mathbf{X}_l - \mathbf{B}\mathbf{f}_l)$.
(c) Sample $(\omega \mid -)$. Let $\tilde{\boldsymbol{\Lambda}}$ be the vector of nonzero elements $\boldsymbol{\Lambda}$ and $\tilde{\mathbf{w}}$ be the corresponding elements of $\mathbf{w}$. Then let m be the length of $\tilde{\boldsymbol{\Lambda}}$ and define $r_l = \tilde{\lambda}_l / w_l$.
(d) Sample $(\mathbf{w} \mid -)$. For each w_j, $j = 1, \ldots, p$, let $m = 1$ and define $r = \lambda_j / \omega$.
(e) Sample $(\tau \mid -)$. Let $\tilde{\mathbf{B}}$ be a vector of the nonzero elements of $\{\mathbf{B}, \boldsymbol{\theta}\}$ and $\tilde{\mathbf{t}}$ be a vector of the corresponding elements of $\{\mathbf{t}, \mathbf{q}\}$. Then let m be the length of $\tilde{\mathbf{B}}$ and define $r_l = \tilde{b}_l / t_l$.
(f) Sample $(\mathbf{t} \mid -)$. For each t_{jh}, $j = 1, \ldots, p$, $h = 1, \ldots, k$, let $m = 1$ and define $r = b_{jg}/\tau$.
(g) Sample $(\mathbf{q} \mid -)$. For each q_h, $h = 1, \ldots, k$, let $m = 1$ and define $r = \theta_h / \tau$.

3. Sample the residual regression coefficients: $(\boldsymbol{\Lambda} \mid -)$. Define $Y_i^* = Y_i - \boldsymbol{\theta}\mathbf{f}_i$ and $\mathbf{X}_i^* = \boldsymbol{\Psi}^{-1/2}(\mathbf{X}_i - \mathbf{B}\mathbf{f}_i)$. Sequentially for each $j = 1, \ldots, p$, define $\tilde{Y}_i = Y_i^* - \boldsymbol{\Lambda}_{-j}\mathbf{X}_{-j}^*$ and $\tilde{X}_{ji} = X_{ji}^*$ and first draw $\lambda_j \sim \mathrm{N}(\mu, s)$ with

$$\mu = s\tilde{\mathbf{X}}\tilde{\mathbf{Y}}/\sigma^2,$$
$$s = \left(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^t/\sigma^2 + \omega^{-2}w_j^{-2}\right)^{-1}, \tag{A.1}$$

and then set to exactly zero with probability

$$\frac{(1-\alpha_\lambda)\phi(0 \mid \mu, s)}{(1-\alpha_\lambda)\phi(0 \mid \mu, s) + \alpha_\lambda \phi(0 \mid 0, \omega^2 w_j^2)},$$

where $\phi(\cdot \mid m, s)$ denotes the normal density function with mean m and variance s.

4. Sample the factor regression coefficients: $(\boldsymbol{\theta} \mid -)$. Define $Y_i^* = Y_i - \boldsymbol{\Lambda}\boldsymbol{\Psi}^{-1/2}(\mathbf{X}_i - \mathbf{B}\mathbf{f}_i)$. Sequentially for each $h = 1, \ldots, k$, define $\tilde{Y}_i = Y_i^* - \boldsymbol{\theta}_{-h}\mathbf{f}_{-h,i}$ and first draw $\theta_h \sim \mathrm{N}(\mu, s)$ with

$$\mu = s\mathbf{f}_h\tilde{\mathbf{Y}}/\sigma^2,$$
$$s = \left(\mathbf{f}_h^t\mathbf{f}_h/\sigma^2 + \tau^{-2}q_h^{-2}\right)^{-1}. \tag{A.2}$$

and then set to exactly zero with probability

$$\frac{(1-\alpha_\theta)\phi(0 \mid \mu, s)}{(1-\alpha_\theta)\phi(0 \mid \mu, s) + \alpha_\theta \phi(0 \mid 0, \tau^2 q_h^2)}.$$

Table A.1. As the variance of the prior on $\boldsymbol{\Lambda}$ is relaxed, the norm gets bigger and prediction and inference improves to a point, but eventually declines. When the prior variance is learned via a hyperprior, one gets estimation results near the optimal setting among the fixed values

Metric	Horseshoe hyperprior	$c = 0.0001$	$c = 0.05$	$c = 0.1$	$c = 0.5$
MSPE	1.38	5.27	1.53	1.76	2.10
$\mathrm{E}\|\|\boldsymbol{\Lambda}\|\|$	0.42	0.00	0.13	0.38	6.14

5. Sample the factor loadings: $(\mathbf{B} \mid -)$. The strategy here is to use a Metropolis-adjusted Gibbs update. Specifically, we use the posterior distribution of $\mathbf{B}$ *disregarding* $\mathbf{Y}$ as a proposal distribution: that is, we draw from

$$\pi(\mathbf{B} \mid \mathbf{X}) = \frac{f(\mathbf{X} \mid \mathbf{B})\pi(\mathbf{B})}{\int f(\mathbf{X} \mid \mathbf{B})\pi(\mathbf{B})d\mathbf{B}}.$$

Because $p(\mathbf{X} \mid \mathbf{B})$ and $\pi(\mathbf{B})$ are shared with the true posterior $\pi(\mathbf{B} \mid \mathbf{X}, \mathbf{Y})$, the probability of transitioning from $\mathbf{B}$ to $\mathbf{B}'$ is expressed as

$$\min\left(1, \frac{\prod_{i=1}^n \phi(Y_i \mid \mathbf{B}', \mathbf{X}, -)}{\prod_{i=1}^n \phi(Y_i \mid \mathbf{B}, \mathbf{X}, -)}\right).$$

Drawing from this proposal is in turn done with a Gibbs sampler similar to the previous two steps. Sequentially for each $h = 1, \ldots, k$ and each $j = 1, \ldots, p$, define $\tilde{X}_{ji} = X_{ji} - \mathbf{b}_{j,-h}\mathbf{f}_{-h,i}$ and first draw $b_{jh} \sim \mathrm{N}(\mu, s)$ with

$$\mu = s\mathbf{f}_h\tilde{\mathbf{X}}_j^t/\psi_j^2,$$
$$s = \left(\mathbf{f}_h^t\mathbf{f}_h/\psi_j^2 + \tau^{-2}t_{jh}^{-2}\right)^{-1}. \tag{A.3}$$

and then set to exactly zero with probability

$$\frac{(1-\alpha_b)\phi(0 \mid \mu, s)}{(1-\alpha_b)\phi(0 \mid \mu, s) + \alpha_b\phi\left(0 \mid 0, \tau^2 t_{jh}^2\right)}.$$

APPENDIX B: SENSITIVITY STUDY

This section seeks to briefly characterize the quantitative behavior of the partial factor model as the prior variance of $\boldsymbol{\Lambda}$ is varied. To do so, consider $w_j^2 = \omega = c$, as c ranges over the set $\{0.0001, 0.05, 0.1, 0.5\}$. Additionally, we consider the "learned" model using the default global-local horseshoe hyperpriors over ω and w_j described in (14). For this demonstration, we let the true model be a $k = 5$ factor model with $p = 50$ and $n = 40$ drawn as in (15) with $\theta_g \sim \mathrm{N}(0, 1)$ for $g = 1, \ldots, k$. We use the same heuristic described in Section 2.1 to chose the number of factors in the factor model.

We examine two quantities, the MSPE (as a fraction of the best possible) as in (16) and the posterior mean size of $\boldsymbol{\Lambda}$ as measured by the 2-norm. Results are recorded in Table A1.

These results are intuitive: as the prior is tighter about zero, posterior estimates look more and more like those from a factor model, while if the prior is loosened estimates get closer and closer to those of a pure regression model. Similar patterns persist over independent trials of this exercise.

[Received July 2011. Revised November 2012.]

REFERENCES

Aguilar, O., and West, M. (2000), "Bayesian Dynamic Factor Models and Variance Matrix Discounting for Portfolio Allocation," *Journal of Business and Economic Statistics*, 18, 338–357. [999,1000]

Albert, J., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679. [1006]

Bai, J. (2003), "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71, 135–171. [1000]

Bartholomew, D., and Moustaki, I. (2011), *Latent Variable Models and Factor Analysis: A Unified Approach*, (3rd ed.) New York: Wiley. [1000]

Bhattacharya, A., and Dunson, D. B. (2011), "Sparse Bayesian Infinite Factor Models," *Biometrika*, 98, 291–306. [1000,1001,1002,1003]

Cao, K.-A. L., Gonzalez, I., and Dejean, S. (2009), "IntegrOmics: An R Package to Unravel Relationships Between Two Omics Data Sets," *Bioinformatics*, 25, 2855–2856. [1003]

Carvalho, C. M., Lucas, J., Wang, Q., Nevins, J., and West, M. (2008), "High-Dimensional Sparse Factor Modelling: Applications in Gene Expression Genomics," *Journal of the American Statistical Association*, 103, 1438–1456. [1000]

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010), "The Horseshoe Estimator for Sparse Signals," *Biometrika*, 97, 465–480. [1001,1002,1005]

Chamberlain, G. (1983), "Funds, Factors and Diversification in Arbitrage Pricing Theory," *Econometrica*, 51, 1305–1323. [1000]

Chamberlain, G., and Rothschild, M. (1983), "Arbitrage, Factor Structure and Mean-Variance Analysis on Large Asset Markets," *Econometrica*, 51, 1281–1304. [1000]

Clyde, M., and George, E. I. (2004), "Model Uncertainty," *Statistical Science*, 19, 81–94. [999,1003]

Clyde, M. A., Ghosh, J., and Littman, M. L. (2011), "Bayesian Adaptive Sampling for Variable Selection and Model Averaging," *Journal of Computational and Graphical Statistics*, 20, 80–101. [1005]

Cox, D. (1968), "Notes on Some Aspects of Regression Analysis," *Journal of the Royal Statistical Society,* Series A, 131, 265–279. [999]

Damien, P., Wakefield, J., and Walker, S. (1999), "Gibbs Sampling for Bayesian Non-Conjugate and Hierarchical Models by Using Auxiliary Variables," *Journal of the Royal Statistical Society,* Series B, 61, 331–344. [1007]

Dawid, A. P., and Lauritzen, S. L. (1993), "Hyper-Markov Laws in the Statistical Analysis of Decomposable Graphical Models," *The Annals of Statistics*, 3, 1272–1317. [1003]

Dempster, A. (1972), "Covariance Selection," *Biometrics*, 28, 157–175. [1003]

Fama, E., and French, K. (1992), "The Cross-Section of Expected Stock Returns," *Journal of Finance*, 47, 427–465. [1000]

——— (1993), "Common Risk Factors in the Returns on Stocks and Bonds," *Journal of Financial Economics*, 33, 3–56. [1000]

Fan, J., Fan, Y., and Lv, J. (2008), "High Dimensional Covariance Matrix Estimation Using a Factor Model," *Journal of Econometrics*, 147, 186–197. [1000]

Fruhwirth-Schnatter, S., and Lopes, H. (2012), "Parsimonious Bayesian Factor Analysis When the Number of Factors is Unknown," Technical Report, University of Chicago Booth School of Business. [1000,1004]

George, E. I., and Foster, D. P. (2000), "Calibration and Empirical Bayes Variable Selection," *Biometrika*, 87, 731–747. [1003]

George, E. I., and McCulloch, R. E. (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373. [999,1003]

George, E. I., and Oman, S. D. (1996), "Multiple-Shrinkage Principal Component Regression," *Journal of the Royal Statistical Society*, Series D, 45, 111–124. [1002]

Geweke, J., and Zhou, G. (1996), "Measuring the Pricing Error of the Arbitrage Pricing Theory," *The Review of Financial Studies*, 9, 557–587. [1000,1004]

Ghosh, J., and Dunson, D. B. (2009), "Default Prior Distributions and Efficient Posterior Computation in Bayesian Factor Analysis," *Journal of Computational and Graphical Statistics*, 18, 306–320. [1000]

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning* (Springer Series in Statistics), New York: Springer. [1000,1006]

Hotelling, H. (1957), "The Relationship of the Newer Multivariate Statistical Methods to Factor Analysis," *British Journal of Statistical Psychology*, 10, 69–79. [999]

Jeng, X. J., and Daye, Z. J. (2011), "Sparse Covariance Thresholding for High-Dimensional Variable Selection," *Statistica Sinica*, 21, 625–657. [1003]

Jolliffe, I. T. (1982), "A Note on the Use of Principal Components in Regression," *Journal of the Royal Statistical Society*, Series C, 31, 300–303. [999]

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008), "Mixtures of g Priors for Bayesian Variable Selection," *Journal of the American Statistical Association*, 103, 410–423. Available at *http://ideas.repec.org/a/bes/jnlasa/v103y2008mmarchp410-423.html*. [1003,1005]

Lopes, H. (2003), "Factor Models: An Annotated Bibliography," *Bulletin of the International Society for Bayesian Analysis*, 10, 7–10. [999,1000]

Lopes, H., and Carvalho, C. M. (2007), "Factor Stochastic Volatility With Time Varying Loadings and Markov Switching Regimes," *Journal of Statistical Planning and Inference*, 137, 3082–3091. [1000]

Lopes, H., and West, M. (2004), "Bayesian Model Assessment in Factor Analysis," *Statistica Sinica*, 14, 41–67. [1000]

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J., and West, M. (2012), "Sparse Statistical Modelling in Gene Expression Genomics," in *Bayesian Inference for Gene Expression and Proteomics*, eds. M., Vannucci, K.-A. Do, and P. Müller, Cambridge: Cambridge University Press, chapter 8, pp. 155–176. [1000]

Merl, D., Chen, J. L.-Y., Chi, J.-T., and West, M. (2009), "An Integrative Analysis of Cancer Gene Expression Studies Using Bayesian Latent Factor Modeling," *Annals of Applied Statistics*, 3, 1675–1694. [1000]

Mevik, B. H., and Wehrens, R. (2007), "The `pls` Package: Principal Component and Partial Least Squares Regression in R," *Journal of Statistical Software*, 18, 1–24. [1003]

Mitchell, T., and Beauchamp, J. (1988), "Bayesian Variable Selection in Linear Regression" (with discussion), *Journal of the American Statistical Association*, 83, 1023–1036. [999]

Press, S. (1982), *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference* (2nd ed.), New York: Krieger. [1000]

Spearman, C. (1904), "General Intelligence, Objectively Determined and Measured," *American Journal of Psychology*, 15, 201–293. [1000]

Speed, T., and Kiiveri, H. (1986), "Gaussian Markov Distributions Over Finite Graphs," *The Annals of Statistics*, 14, 138–150. [1003]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [1000,1003]

Varmuza, K., and Filzmoser, P. (2009), *Introduction to Multivariate Statistical Analysis in Chemometrics*, Boca Raton, FL: CRC Press. [1003]

West, M. (2003), "Bayesian Factor Regression Models in the 'Large p, Small n' Paradigm," in *Bayesian Statistics 7*, eds. M. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, Oxford: Oxford University Press, pp. 723–732. [999,1000]

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, New York: Wiley. [1003]

——— (1986), "On Assessing Prior Distributions and Bayesian Regression Analysis With g-Prior Distributions," in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, ed. P. K. Goel, Amsterdam: Elsevier, pp. 233–243. [1006]

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society*, Series B, 67, 301–320. [1005]