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

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions
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: 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 \(X_i\). Specifically, for Gaussian factor models

\[
(Y_i \mid X_i, \beta, \sigma^2) \sim N(X_i^T \beta, \sigma^2),
\]

with marginal predictor model

\[
X_i = B \psi_i + \nu_i, \quad \nu_i \sim N(0, \Psi)
\]

where \(B\) is a \(p \times k\) real-valued matrix. If \(\Psi\) is assumed diagonal, the elements of \(X_i\) are conditionally independent given the \(k\)-vector of common unobserved (latent) factors \(\psi_i\), so that \(\nu_i\) 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 \(\beta\) depend on \(B\) and \(\Psi\)? Two prevailing approaches represent extreme answers to this question. At one extreme, a pure linear regression model ignores the marginal distribution of the predictors, \(\pi(X)\), entirely, which is equivalent to setting \(\pi(\beta \mid B, \Psi) = \pi(\beta)\).

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 \(X_i\) and is conditionally independent of \(X_i\) given these factors: \(\pi(Y_i \mid X_i, \psi, \theta) = \pi(Y_i \mid \theta, \psi)\) and more specifically \(E(Y_i \mid \theta, \psi) = \theta^T \psi_i\), where \(\theta\) denotes a \(1 \times k\) vector of regression coefficients. This entails that \(\beta\) can be written as a deterministic function of \(B\) and \(\Psi\):

\[
\beta' = \theta B^T (BB^T + \Psi)^{-1}.
\]

This very strong prior can lead to poor estimates of \(\beta\) 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 \(B^T C (BB^T + \Psi)^{-1} X_i\), the projection of the predictors onto the \(k\)-dimensional subspace defined by \(B\) and \(\Psi\). 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 1933; 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 \(X\), even if our practical goal is to use the \(X_i\) 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):

\[
E(\beta' \mid B, \Psi, \theta) = \theta B (BB^T + \Psi)^{-1}.
\]

Permitting deviations from the pure factor model safeguards inference against misspecification in the sense that for any fixed values of \(B\), \(\Psi\), and \(\theta\), our prior on \(\beta\) 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 as...
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, $B$ is a $p \times k$ matrix and $\Psi$ is assumed diagonal. The matrix $B$ is referred to as a loadings matrix, the elements of $\Psi$ are referred to as idiosyncratic variances, and the $f_i$ are called as factor scores. Conditional on $B$ and $f_i$, the elements of each observation are independent. Integrating over $f_i$, we see

$$\text{cov}(X_i) \equiv \Sigma_X = BB' + \Psi. \quad (5)$$

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 $BB' + \Psi$, where $\text{rank}(B) \leq k$. Note that $BB' + \Psi$ has full rank whenever the idiosyncratic variances are strictly positive, while $B$, 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 $f_i$, we arrive at the Bayesian factor regression model:

$$Y_i = \theta f_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

$$\Sigma = \text{cov} \left( X_i, Y_i \right) = \begin{bmatrix} BB' + \Psi & V \\ V & \xi \end{bmatrix}, \quad (6)$$

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

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 Aguilera 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; Aguilera 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; Bhattacharyya 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

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

and idiosyncratic variances $\psi_{ij} = 0.2$ 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

$$A' = \begin{bmatrix} 0.0004 & -3.9967 & 0 & -7.9713 & -3.9967 \\ -5.9778 & 0.9990 & -0.9960 & 3.9967 & -0.0004 \end{bmatrix}$$

and idiosyncratic variances given by the vector

$$D = [1.2000, 0.1871, 0.2000, 1.5032, 0.1871, 1.3762, 0.1996, 1.2054, 0.1872, 1.2000].$$

Observe that the one-factor loadings matrix $A$ is very nearly equal to the first factor of $B$, 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 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
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:

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N(0, \Sigma)$$

$$\Sigma = \begin{bmatrix} BB' + \Psi & V' \\ V & \theta \xi \end{bmatrix}. \tag{7}$$

The difference between (6) and (7) is simply that in (7) $V$ is not required to exactly equal $\theta B'$. The matrix $B$ is still a $p \times k$ matrix with $k < p - 1$ so that the $p \times p$ predictor covariance matrix is constrained to the $BB' + \Psi$ form, but the full covariance matrix $\Sigma$ 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 $\Sigma X$.

Just as crucially, the prior on $V$ may be conditioned on $\Sigma X$. 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 $X_i, Y_i$, and the $k$ latent factors, which have covariance:

$$\text{cov} \begin{pmatrix} X_i \\ f_i \\ Y_i \end{pmatrix} = \begin{bmatrix} BB' + \Psi & B & \theta \xi \\ B & I_k & \theta \xi \\ V & \theta \xi & \xi \end{bmatrix}. \tag{8}$$

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

$$E(Y_i | f_i, X_i) = \theta f_i + \{V - \theta B')\Psi^{-1}\{\Psi^{-1}(X_i - Bf_i)\} \tag{9}$$

$$\text{var}(Y_i | f_i, X_i) = \xi - [V \theta] \Sigma^{-1}_{X,f} [V \theta]' \equiv \sigma^2. \tag{10}$$

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

$$v_j \sim N(\theta B'j, \omega^2 w_j^2 \psi_j^2), \tag{11}$$

for element $j$, implying that a priori the error piece plays no role in the regression. Here, $\omega^2$ 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); $\psi_j^2$ is the $j$th diagonal element of $\Psi$. This effect is perhaps easiest to see via the following reparameterization: define

$$\Lambda = (V - \theta B')\Psi^{-1/2}. \tag{12}$$

Then

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

is equivalent to (11).

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
Using this formulation (along with some additional hyperparameters), the model may be expressed as
\[
X_i \mid B, f, \Psi \sim N(Bf_i, \Psi)
\]
\[
Y_i \mid X_i, B, \theta, A, f, \Psi, \sigma^2 \sim N(\theta f_i + A \{\Psi^{-1}(X_i - Bf_i)\}, \sigma^2)
\]
\[
\lambda_j \sim N(0, \omega^2 w_j^2), \quad f_i \sim N(0, I_k)
\]
\[
\theta_h \sim N(0, \tau^2 q_h^2)
\]
\[
b_{jh} \sim N(0, \tau^2 t_{jh}^2), \quad h = 1, \ldots, k, \quad j = 1, \ldots, p.
\]

Independent half-Cauchy priors are placed over \(\tau\) and \(\omega\) and the individual elements of the vectors \(t\), \(w\), and \(q\). This corresponds to the so-called horseshoe priors (Carvalho, Polson, and Scott 2010) over the elements of \(B\), \(\theta\), and \(A\), respectively. The residual standard deviations \(\sigma\) and each element of \(\Psi^{1/2}\) 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 \(A\) affect posterior estimates of \(\beta\).

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 when 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 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
\[
\text{MSPE}(\beta_{est}) = \frac{E((Y_{n+1} - X_{n+1}\beta_{est})^2)}{E((Y_{n+1} - X_{n+1}\beta_{true})^2)}
\]
where the expectation is taken over \((Y_{n+1}, X_{n+1})\). 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 \(\text{Pr(optimal)}\).

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 \(X\), 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, \(\beta\). 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>
<thead>
<tr>
<th>Method</th>
<th>MSPE</th>
<th>Pr(optimal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partial factor regression</td>
<td>1.31</td>
<td>0.33</td>
</tr>
<tr>
<td>Bhattacharya et al.</td>
<td>1.33</td>
<td>0.58</td>
</tr>
<tr>
<td>Carvalho et al.</td>
<td>1.49</td>
<td>0.09</td>
</tr>
</tbody>
</table>

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.
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

<table>
<thead>
<tr>
<th>Method</th>
<th>MSPE</th>
<th>Pr(optimal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partial factor regression</td>
<td>1.59</td>
<td>0.54</td>
</tr>
<tr>
<td>Bhattacharya et al.</td>
<td>5.86</td>
<td>0.41</td>
</tr>
<tr>
<td>Carvalho et al.</td>
<td>1.84</td>
<td>0.05</td>
</tr>
</tbody>
</table>

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 $\Sigma_{X,Y}^{-1}$, 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 $\beta$ 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 $f_i$) follows as

$$Y_i = \theta f_i + \Lambda \Psi^{-\frac{1}{2}}(X_i - BF_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2). \quad (17)$$

Fixing $\Lambda$ 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 $\Lambda^*, \theta$, and $B$ implies positive probability of zero elements in $\beta$. For instance, a point-mass prior at zero (sometimes called a two-groups model or a spike-and-slab model) on the $j$th element of $\Lambda$ can be written as

$$\pi(\lambda_j \mid \alpha_j) = \alpha_j N(\lambda_j; 0, \omega^2 w^2_j) + (1 - \alpha_j) \delta(\lambda_j), \quad (18)$$

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>n</th>
<th>p</th>
<th>PFR</th>
<th>RIDGE</th>
<th>PL$S$</th>
<th>PCR</th>
<th>LASSO</th>
<th>PCR</th>
<th>BFR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nutrimouse</td>
<td>40</td>
<td>120</td>
<td>377.3 (27%)</td>
<td>296.2</td>
<td>418.5 (41%)</td>
<td>492.3 (66%)</td>
<td>391.2 (32%)</td>
<td>517.73 (75%)</td>
<td></td>
</tr>
<tr>
<td>Cereal</td>
<td>15</td>
<td>145</td>
<td>31.8</td>
<td>41.86 (32%)</td>
<td>51.30 (61%)</td>
<td>42.97 (35%)</td>
<td>45.01 (42%)</td>
<td>62.34 (96%)</td>
<td></td>
</tr>
<tr>
<td>Yarn</td>
<td>28</td>
<td>268</td>
<td>0.29</td>
<td>0.50 (72%)</td>
<td>0.37 (28%)</td>
<td>0.30 (3%)</td>
<td>0.42 (45%)</td>
<td>7.80 (260%)</td>
<td></td>
</tr>
<tr>
<td>Gasoline</td>
<td>60</td>
<td>401</td>
<td>0.54</td>
<td>0.68 (26%)</td>
<td>0.74 (37%)</td>
<td>0.81 (50%)</td>
<td>0.70 (30%)</td>
<td>0.71 (31%)</td>
<td></td>
</tr>
<tr>
<td>Multidrug</td>
<td>60</td>
<td>853</td>
<td>183.0 (18%)</td>
<td>154.0</td>
<td>164.0 (6%)</td>
<td>220.4 (43%)</td>
<td>170.6 (11%)</td>
<td>212.13 (38%)</td>
<td></td>
</tr>
</tbody>
</table>

NOTE: Percentages shown are amount worse than the best method, reported in bold type.
where $\delta_0$ denotes a point mass at zero. Analogous priors are placed on the elements of $\mathbf{B}$ and $\theta$, with corresponding hyperparameters $\alpha_\theta$ and $\alpha_\beta$:

$$
\pi(\theta_h \mid \alpha_\theta) = \alpha_\theta N(\theta_h; 0, \tau^2 q^2_h) + (1 - \alpha_\theta) \delta_0(\theta_h),
$$

$$
\pi(\beta_{jh} \mid \alpha_\beta) = \alpha_\beta N(\beta_{jh}; 0, \tau^2 \tilde{r}^2_{jh}) + (1 - \alpha_\beta) \delta_0(\beta_{jh}).
$$

(19)

Sparsity of $\beta$ is then induced via the identity

$$
\beta' = (\theta - \Lambda \Psi^{-\frac{1}{2}} \mathbf{B}) \mathbf{B}' (\mathbf{BB}' + \Psi)^{-1} + \Lambda \Psi^{-\frac{1}{2}}.
$$

(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 $\alpha_\theta = 0$ and $\alpha_\beta = 0$, and for our simulation study we adopt the convenient lower-triangular restriction of Geweke and Zhou (1996).

Zeros in $\beta$ arise only when $\Sigma_X$ 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

$$
\mathbf{B} =
\begin{bmatrix}
\mathbf{B}_1 & 0 & 0 & \cdots \\
0 & \mathbf{B}_2 & 0 & \cdots \\
0 & 0 & \ddots & \cdots \\
\vdots & \vdots & \ddots & \mathbf{B}_d
\end{bmatrix},
$$

and $\Lambda = (\Lambda_1 \ \Lambda_2 \ \cdots \ \Lambda_d)$. $\theta = (\theta_1 \ \theta_2 \ \cdots \ \theta_d)$ represent conformable partitions. From these definitions it follows straightforwardly that $\beta$ also partitions as $\beta' = (\beta'_1 \ \beta'_2 \ \cdots \ \beta'_d)$, with each group being defined as in (20):

$$
\beta'_i = (\theta_i - \Lambda_i \Psi^{-\frac{1}{2}} \mathbf{B}_i) \mathbf{B}'_i (\mathbf{B}_i \mathbf{B}'_i + \Psi)^{-1} + \Lambda_i \Psi^{-\frac{1}{2}}.
$$

(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 $\beta_i$ is a zero vector precisely when $\Lambda_i$ and $\theta_i$ 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 ($\theta_i = 0$) and is also unrelated to that predictor residually ($\Lambda_i = 0$).

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 $\beta$ 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.

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

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

$$
\mathbf{B}' = \begin{bmatrix}
-0.25 & 0.1 & 0.1 & c & 0 & 0 \\
0 & 0 & 0 & -0.8 & 0.2 & -2.0 \\
\end{bmatrix}, \ \Psi = 0.5\mathbf{I}
$$

$$
\Lambda = (0 \ 0.1 \ 0 \ 0 \ 0 \ 0), \ \theta = -0.5 \ 0.
$$

(23)

Now consider how $\beta$ changes as a function of $c$, noting that when $c = 0$ we have the desired block independence of $\mathbf{X}$.

Consider the following partitioning: $\theta = (\theta_1 \ \theta_2)$, $\Lambda = (\Lambda_1 \ \Lambda_2)$, and $\beta = (\beta_1 \ \beta_2)$. Specifically $\Lambda_1 = (\Lambda_1 \ \Lambda_2 \ \Lambda_3)$, $\Lambda_2 = (\Lambda_4 \ \Lambda_5 \ \Lambda_6)$ and $\beta_1 = (\beta_1 \ \beta_2 \ \beta_3)$, $\beta_2 = (\beta_4 \ \beta_5 \ \beta_6)$. We see that because $\theta_2 = 0$ and $\Lambda_2 = 0$, the only way elements of $\beta_2$ become nonzero is via nonzero $c$. When $c = 0$, all the elements of $\beta$ 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 $\Lambda = (0 \ 0.1 \ 0 \ 0 \ 0 \ 0.25)$, the picture changes dramatically, as shown in Figure 3. This perhaps is counterintuitive because $\lambda_6$ 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 $\Lambda_6$ 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 $\lambda_j$ is nonzero, the posterior probability that $\beta_j$ 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...
is a two-groups regression model with an independent prior on $\beta$:

$$
\pi(\beta_j | v_j, v) = \frac{1}{2} \delta_0(\beta_j) + \frac{1}{2} N(0, \omega^2 w_j^2),
$$

$$
\omega \sim C^+(0, 1),
$$

$$
w_j \sim C^+(0, 1).
$$

(24)

This prior is an amalgam of the usual two-groups model (with normal density for the continuous portion) and a shrinkage prior on the nonzero regression coefficients:

$$
\pi(\beta_M | X, M) = N(0, g^{-1}(X_M X_M^t)^{-1}).
$$

(25)

where $M$ 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 $\alpha$ 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 $\theta$, $B$, and $\Lambda$ are given horseshoe point-mass priors as in (24). However, different prior probabilities are chosen (fixed): $a_0 = 0.5$, $\alpha_0 = 0.1$, and $\sigma_0 = 0.9$ (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: $p^{-1} \sum_j (\hat{\beta}_j - \beta_j)^2$,

- the average misclassified error (ME): $p^{-1} \sum_j 1\{\beta_j \neq 0\} - \beta_j > \tfrac{1}{2}\}^2$,

- and the squared difference between the posterior inclusion probability and the true model indicator vector, or the "probability estimate" (PE) error: $p^{-1} \sum_j (1\{\beta_j \neq 0\} - \tilde{\alpha}_\beta_j)^2$.

where $\tilde{\alpha}_\beta_j$ 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.

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 $\Lambda$ are drawn from independent normal distributions with mean zero and variance 0.25, and the elements of $\theta$ 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 classification 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 $\beta$. 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 $\beta$, 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 $\beta$, 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 $q_h^2 = \omega^2 w_j^2$ for all $h$, fix $w_j = 1$ for all $j$, fix the residual variance at $\sigma^2 = 1$, and assume $XX'$ is invertible (similar expressions arise using a generalized inverse). Then, from expression (20) one finds that the prior variance of $\beta$ in terms of $B$ and $\Psi$ is
\[
\text{cov}(\beta) = \omega^2 (BB' + \Psi)^{-1},
\]
\[
= \omega^2 \Sigma_X^{-1},
\]
from which the posterior mean may be expressed as
\[
E(\beta \mid Y, X, \Sigma_X) = (\omega^{-2}I_p + \Sigma_X^{-1}XX')^{-1} \Sigma_X^{-1}XY.
\] (27)

By comparison, the normal prior used in ridge regression has prior variance $\text{cov}(\beta) = \omega^2 I_p$ and yields the estimator
\[
E_{\text{ridge}}(\beta \mid Y, X) = (XX' + \omega^{-2}I_p)^{-1}XY.
\] (28)

And finally, the $g$ prior is a normal prior with $\text{cov}(\beta) = g^{-1}(XX')^{-1}$, yielding the posterior estimator
\[
E_{\text{Zellner}}(\beta \mid Y, X) = (1 + g)^{-1}(XX')^{-1}XY.
\] (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 $\beta$ 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 $X$ and $\Sigma_X = \text{cov}(X_i)$ individually, which is what the partial factor model does. This teasing apart may be observed directly from the form of (27). Because $\Sigma_X$ and $XX'/n$ 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 $XX'$ as a plug-in estimate of $\Sigma_X$. 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
\[
Z_i = \mathbb{I}(Y_i < 0),
\]
where $Y_i$ is modeled as in (17), inducing a partial factor probit model for $Z_i$ conditional on the vector of predictors $X_i$.

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:
\[
E(Y_i \mid f_i, X_i) = \phi(f_i, X_i - E(X_i \mid f_i))
\] (30)
for an unknown function $\phi$. A more modest nonlinear generalization would be to assume additivity:
\[
E(Y_i \mid f_i, X_i) = \phi(f_i) + \varphi(X_i - E(X_i \mid f_i)),
\] (31)
where $\phi$ and $\varphi$ denote smooth functions to be inferred from the data. Crucially, the smoothness assumptions for $\phi$ and $\varphi$ can be chosen individually. In particular, it would be interesting to consider simple parametric forms for $\phi$ while allowing $\varphi$ 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: \((F \mid -)\). Using the joint normal distribution of \(f_1, X_1, \) and \(Y_1, \) we draw \(f_1 \sim N(\mu_1, S)\), where
\[
\mu_1 = (B' \theta') \Sigma_{X,Y}^{-1} (X_1 \ Y_1 \ Y_1^t) \\
S = I_l - (B' \theta') \Sigma_{X,Y}^{-1} (B' \theta')^t.
\]
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
\[
\cdot \text{Draw } (u \mid \eta) \sim \text{Uniform}(0, (1 + \eta)^{-1/2}) \\
\cdot \text{Draw } (\eta \mid r, u) \sim \text{Gamma}((m + 1)/2, u(1 + \eta)^{-1/2}) \\
\cdot \text{Sample } (\sigma^2 \mid -). \text{ Let } \lambda = \lambda_i/\eta. (a) \text{ Sample } (\Psi \mid -). \text{ Let } m = n \text{ and for each dimension } j = 1, \ldots, p, \text{ define } r_1(j) = X_j - B f_1. \\
\cdot \text{Sample } \sigma^2 \mid -). \text{ Let } m = n \text{ and define } r_i = Y_i - \theta f_i - \Lambda \Psi^{-1/2} (X_i - Bf_i). \\
\cdot \text{Sample } (\omega \mid -). \text{ Let } \Lambda \text{ be the vector of nonzero elements } A \text{ and } \omega \text{ be the corresponding elements of } \omega. \text{ Then let } m \text{ be the length of } \Lambda \text{ and define } r_i = \lambda_i/\omega. \\
\cdot \text{Sample } (\tau \mid -). \text{ Let } B \text{ be a vector of the nonzero elements of } (B, \theta) \text{ and } f_i \text{ be a vector of the corresponding elements of } (t, q). \text{ Then let } m \text{ be the length of } B \text{ and define } r_i = b_i/t_i.
\]

3. Sample the residual regression coefficients: \((A \mid -)\). Define \(Y_i^* = Y_i - A \Psi^{-1/2} (X_i - B f_i)\). Sequentially for each \(j = 1, \ldots, p\), define \(\tilde{Y}_j = Y_j - A \Psi^{-1/2} \tilde{X}_j\) and \(\tilde{X}_j = X_j'f_1\) and first draw \(\lambda_j \sim N(\mu, s)\) with
\[
\begin{align*}
\mu &= s \tilde{X} \tilde{Y} / \sigma^2, \\
\tau &= (\tilde{X} \tilde{X}^{-1} + \omega^2 \tilde{w}^{-2})^{-1},
\end{align*}
\]
and then set to exactly zero with probability
\[
\frac{(1 - \alpha_\mu) \phi(0 \mid \mu, s)}{1 - \alpha_\mu} \phi(0 \mid \mu, s) + \alpha_\mu \phi(0 \mid 0, \omega^2 w_i^2),
\]
where \(\phi(-, m, s)\) denotes the normal density function with mean \(m\) and variance \(s\).

4. Sample the factor regression coefficients: \((\theta \mid -)\). Define \(Y_i^* = Y_i - A \Psi^{-1/2} (X_i - B f_i)\). Sequentially for each \(j = 1, \ldots, k\), define \(\tilde{Y}_j = Y_j - \theta \omega f_j\) and first draw \(\theta_j \sim N(\mu, s)\) with
\[
\begin{align*}
\mu &= s \tilde{f}_j \tilde{Y} / \sigma^2, \\
\tau &= (\tilde{f}_j \tilde{f}_j + \tau^{-2} q_h^{-2})^{-1},
\end{align*}
\]
and then set to exactly zero with probability
\[
\frac{(1 - \alpha_\mu) \phi(0 \mid \mu, s)}{1 - \alpha_\mu} \phi(0 \mid \mu, s) + \alpha_\mu \phi(0 \mid 0, \tau^2 q_h^2),
\]

5. Sample the factor loadings: \((B \mid -)\). The strategy here is to use a Metropolis-adjusted Gibbs update. Specifically, we use the posterior distribution of \(B\) disregarding \(Y\) as a proposal distribution: that is, we draw from
\[
\pi(B \mid X) = \frac{f(X \mid B) \pi(B)}{f(X \mid B) \pi(B) d B}
\]
Because \(p(X \mid B)\) and \(\pi(B)\) are shared with the true posterior \(p(B \mid X, Y)\), the probability of transitioning from \(B\) to \(B'\) is expressed as
\[
\min \left( 1, \sum_{i=1}^n \phi(Y_i \mid B' ; X, -) / \sum_{i=1}^n \phi(Y_i \mid B, X, -) \right).
\]
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 \(\hat{X}_h = X_h - B_h f_{h,j}\) and first draw \(b_{h,j} \sim N(\mu, s)\) with
\[
\mu = s f_{h,j} \hat{X}_h / \psi_j^2. \\
s = (f_{h,j} f_{h,j} + \tau^{-2} q_{h,j}^{-2})^{-1}
\]
and then to exactly zero with probability
\[
\frac{(1 - \alpha_\mu) \phi(0 \mid \mu, s)}{1 - \alpha_\mu} \phi(0 \mid \mu, s) + \alpha_\mu \phi(0 \mid 0, \tau^2 q_h^2).
\]

**APPENDIX B: SENSITIVITY STUDY**

This section seeks to briefly characterize the quantitative behavior of the partial factor model as the prior variance of \(A\) is varied. To do so, consider \(w_i^2 = \omega = c\), 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 \(\omega \) and \(w_i\) 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 \(\theta_i \sim N(0, 1)\) for \(i = 1, \ldots, k\). We use the same heuristic described in Section 2.1 to choose 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 \(A\) 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.

<table>
<thead>
<tr>
<th>Horseshoe Metric</th>
<th>hyperprior</th>
<th>c = 0.0001</th>
<th>c = 0.05</th>
<th>c = 0.1</th>
<th>c = 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSPE</td>
<td>1.38</td>
<td>5.27</td>
<td>1.53</td>
<td>1.76</td>
<td>2.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AI</td>
<td></td>
<td></td>
<td>0.42</td>
</tr>
</tbody>
</table>

**REFERENCES**
