Partial factor modeling: predictor-dependent shrinkage for linear regression

Richard Hahn, Carlos Carvalho and Sayan Mukherjee

JASA 2013

Review by Esther Salazar
Duke University

December, 2013
Outline

- Motivation
- Partial factor regression
- Sparsity priors for variable selection
- Extensions
Factor regression framework

The factor regression framework may be written in two parts:

Linear regression for a scalar response $Y_i$:

$$(Y_i|X_i, \beta, \sigma^2) \sim \mathcal{N}(X_i^t \beta, \sigma^2)$$

Marginal model for a $p$-dimensional vector of predictor variables $X_i$

$$X_i = Bf_i + \nu_i, \quad \nu_i \sim \mathcal{N}(0, \Psi)$$

$$f_i \sim \mathcal{N}(0, I_k)$$

where $B \in \mathbb{R}^{p \times k}$ and $\Psi$ is a diagonal matrix
Factor regression framework

The factor regression framework may be written in two parts:

Linear regression for a scalar response $Y_i$:

$$
(Y_i | X_i, \beta, \sigma^2) \sim \mathcal{N}(X_i^t \beta, \sigma^2)
$$

Marginal model for a $p$-dimensional vector of predictor variables $X_i$

$$
X_i = B f_i + \nu_i, \quad \nu_i \sim \mathcal{N}(0, \Psi)
$$

$$
f_i \sim \mathcal{N}(0, I_k)
$$

where $B \in \mathbb{R}^{p \times k}$ and $\Psi$ is a diagonal matrix

- This work considers modification to a Gaussian factor model, suited for regression and variable selection.
- It differs from previous work on Bayesian variable selection in that it explicitly accounts for predictor correlation structure.
Gaussian factor model

Linear regression for a scalar response $Y_i$:

$$(Y_i \mid X_i, \beta, \sigma^2) \sim \mathcal{N}(X_i^t \beta, \sigma^2)$$

Marginal model for a $p$-dimensional vector of predictor variables $X_i$

$$X_i = Bf_i + \nu_i, \quad \nu_i \sim \mathcal{N}(0, \Psi)$$

$$f_i \sim \mathcal{N}(0, I_k)$$

This paper asks the question: how should the prior of $\beta$ depend on $B$ and $\Psi$?
Two extreme answers:

(1) Pure linear regression which ignores the marginal distribution the predictors $\pi(X)$

$$\pi(\beta | B, \Psi) = \pi(\beta)$$

(2) Pure factor model where $Y_i$ depends linearly on the same $k$ latent factors that captures the covariation in $X_i$

$$\pi(Y_i | X_i, f_i, \theta) = \pi(Y_i | \theta, f_i) \quad \text{and} \quad E(Y_i | \theta, f_i) = \theta f_i$$

Intuition:

$Y_i \sim N(\theta f_i, \sigma^2), \quad X_i \sim N(B f_i, \Psi), \quad f_i \sim N(0, I_k)$

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} BB^T + \Psi & B\theta^T \\ \theta B^T & \theta\theta^T + \sigma^2 \end{bmatrix} \right)$$

$$E(Y_i | X_i) = \theta B^T (BB^T + \Psi)^{-1} X_i$$

This entails that $\beta$ is a deterministic function $\beta^T = \theta B^T (BB^T + \Psi)^{-1}$

Also, $f_i = B^T (BB^T + \Psi)^{-1} X_i$, projection of $X_i$ onto a $k$-dimensional subspace
Bayesian linear factor model

Factor model for the predictors:

\[ X_i \sim \mathcal{N}(B f_i, \Psi), \quad f_i \sim \mathcal{N}(0, I_k) \]

Integrating over \( f_i \):

\[ \text{cov}(X_i) \equiv \Sigma_X = BB^T + \Psi \]

Assuming that the \( p \) predictors influence \( Y_i \) only through the \( k \)-dimensional vector \( f_i \):

\[
Y_i = \theta f_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)
\]

\[
\Sigma = \text{cov}(X_i, Y_i) = \begin{bmatrix} BB' + \Psi & V' \\ V & \xi \end{bmatrix},
\]

\[
V = \theta B', \quad \xi = \sigma^2 + \theta \theta'.
\]
Example 1. Consider the 10-dimensional two-factor Gaussian model with loadings matrix

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

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

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

and idiosyncratic variances given by the vector

\[ D = [1.2000 \ 0.1871 \ 0.2000 \ 1.5032 \ 0.1871 \ 1.3762 \ 0.1996 \ 1.2054 \ 0.1872 \ 1.2000]. \]
By the likelihood criterion the two models are nearly identical. In terms of predicting $X_{10}$ the two-factor model is nearly always the best.
Partial factor regression

Idea: Relax the assumption that the latent factors capturing the predictor covariance $\Sigma_X$ are sufficient for predicting the response $Y_i$.

This is achieved by using the covariance structure

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

$$
\Sigma = \begin{bmatrix}
BB^T + \Psi & V^T \\
V & \xi
\end{bmatrix}
$$

where $V = (v_1, \ldots, v_p)$, the $1 \times p$ row vector, is not exactly equal to $\theta B^T$

Novelty: Prior for $V$, conditional on $\theta$, $B$ and $\Psi$

$$
v_j \sim N (\{\theta B^T\}_j, \omega^2 w_j^2 \psi_j^2)
$$

where $\omega^2$ is a global variance, $w_j^2$ is a predictor-specific variance (Carvalho, Polson, Scott 2010), $\psi_j^2$ is the diagonal element of $\Psi$.
The hierarchical specification arises from the jointly normal distribution between $X_i$, $Y_i$, and the $k$ latent factors, which have covariance:

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

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

$$\begin{aligned}
E(Y_i | f_i, X_i) &= \theta f_i + (V - \theta B') \Psi^{-1/2} \{\Psi^{-1/2} (X_i - B f_i)\} \\
\text{var}(Y_i | f_i, X_i) &= \xi - [V \theta] \Sigma_{X_i f_i}^{-1} [V \theta]' = \sigma^2. \quad (10)
\end{aligned}$$

Let $\Lambda = (V - \theta B^T) \Psi^{-1/2}$. Considering the prior $v_j \sim \mathcal{N} \left( \{\theta B^T\}_j, \omega^2 w_j^2 \psi_j^2 \right)$ and the reparameterization given by $\Lambda$, we have that

$$\lambda_j \sim \mathcal{N}(0, \omega^2 w_j^2)$$
Hierarchical specification

\[
X_i \mid B, f_i, \Psi \sim N(Bf_i, \Psi)
\]

\[
Y_i \mid X_i, B, \theta, \Lambda, f_i, \Psi, \sigma^2 \sim N(\theta f_i + \Lambda \{ \Psi^{-\frac{1}{2}}(X_i - Bf_i) \}, \sigma^2)
\]

\[
\lambda_j \sim N(0, \omega^2 w_j^2),
\]

\[
f_i \sim N(0, I_k)
\]

\[
\theta_h \sim N(0, \tau^2 q_h^2)
\]

\[
b_{jh} \sim N(0, \tau^2 t_{jh}^2), \quad h = 1, \ldots, k,
\]

\[
j = 1, \ldots, p.
\]

- Priors for \(\tau, \omega\), and the individual elements of \(w, q, t\): half-Cauchy
- This corresponds to the horseshoe priors (Carvalho et al, 2010) over the elements of \(B, \theta\) and \(\Lambda\)
- **Posterior inference**: the model can be fit using a Gibbs sampling approach
They compare partial factor regression to five other methods: (1) ridge regression, (2) partial least square, (3) lasso regression, (4) principal component regression and (5) Bayesian factor model using the model selection prior of Bhattacharya and Dunson (2011).

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>PLS</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>
</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>
</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>
</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>
</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>
</tr>
</tbody>
</table>

NOTE: Percentages shown are amount worse than the best method, reported in bold type.
Sparsity priors for variable selection

**Problem:** With the assumption that the predictors $X$ and the data $Y$ come from a joint normal distribution, the variable selection problem is related to infer exactly zero entries in the precision matrix $\Sigma_{X,Y}^{-1}$.

From the partial factor model

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

We can use a spike-and-slab prior over $\Lambda$ such that

$$\pi(\lambda_j | \alpha_{\lambda}) = \alpha_{\lambda} N(\lambda_j; 0, \omega^2 w_j^2) + (1 - \alpha_{\lambda}) \delta_0(\lambda_j),$$

Analogous priors are placed on the elements of $B$ and $\theta$

$$\pi(\theta_h | \alpha_{\theta}) = \alpha_{\theta} N(\theta_h; 0, \tau^2 q_h^2) + (1 - \alpha_{\theta}) \delta_0(\theta_h),$$

$$\pi(\beta_{jh} | \alpha_{\beta}) = \alpha_{\beta} N(\beta_{jh}; 0, \tau^2 t_{jh}^2) + (1 - \alpha_{\beta}) \delta_0(\beta_{jh})$$

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

$$\beta^t = (\theta - \Lambda \Psi^{-\frac{1}{2}} B) B'(BB^t + \Psi)^{-1} + \Lambda \Psi^{-\frac{1}{2}}.$$
Extensions: beyond the linear model

- 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.

- For instance: if $Z_i$ is binary

$$Z_i = 1(Y_i < 0)$$

where $Y_i$ follows the partial factor model called \textit{partial factor probit model}.