Sparse variable clustering models

Ilsang Ohn

February 11, 2019

Introduction: Factor models

The conventional factor model assumes that an observed variable $\mathbf{y} \equiv (y_1, \dots, y_p)$ is generated as

$$\mathbf{y}|\mathbf{z} \sim \mathrm{N}_p(\mathbf{Bz}, \phi\mathbf{I}), \quad \mathbf{z} \sim \mathrm{N}_K(\mathbf{0}, \mathbf{I}).$$

In the high dimensional setup, usually $\mathbf{B}$ is assumed to be sparse.
But if $p \asymp \mathrm{e}^n$, the sparse factor model assumes that most of the variables are just noise.

Sparse variable clustering models

Sparse variable clustering (SVC) model assumes that an observed variable $\mathbf{y} \equiv (y_1, \ldots, y_p)$ is generated as

$$\mathbf{y}|\mathbf{x} \sim \mathrm{N}_p(\mathbf{A}\mathbf{x}, \phi\mathbf{I}), \quad \mathbf{x} \sim \mathrm{N}_K(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{x}),$$

where $\mathbf{x} \equiv (x_1, \ldots, x_K)$ is an unobserved latent variable and $\mathbf{A}$ is a binary valued matrix such that

$$\mathbf{A} \in \mathcal{A}_{p,K} := \left\{\mathbf{A} \in \{0,1\}^{p \times K} : \left\|\mathbf{A}_{[j,:]}\right\|_0 = 1\right\}.$$

and $\boldsymbol{\Sigma}_\mathbf{x}$ is a sparse covariance matrix.

Notation For a $p_1 \times p_2$ matrix $\mathbf{A}$ we let $\mathbf{A}_{[i,j]}$ denote the (i,j)th entry of $\mathbf{A}$. For two index sets $I \subset [p_1]$ and $J \subset [p_2]$, let $\mathbf{A}_{[I,J]}$ denote the submatrix $(\mathbf{A}_{[i,j]})_{i \in I, j \in J}$. For notational convenience, we write $\mathbf{A}_{[i,J]} := \mathbf{A}_{[\{i\},J]}$, $\mathbf{A}_{[-i,J]} := \mathbf{A}_{[[p_1] \setminus \{i\},J]}$ and $\mathbf{A}_{[:,J]} := \mathbf{A}_{[[p_1],J]}$, and the similar notations are used for the column index.

Sparse variable clustering models

The sparse variable clustering model is equivalently written as

$$y_j = x_{\Bbbk(j)} + \epsilon_j$$

where $\epsilon_j \overset{\text{iid}}{\sim} \mathrm{N}(0, \phi)$, and $\Bbbk(j)$ denotes the index of 1 of the vector $\mathbf{a}_j$.

Sparse variable clustering models

We impose sparsity on the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{x}}$ by assuming that $\boldsymbol{\Sigma}_{\mathbf{x}}$ can be decomposed as

$$\boldsymbol{\Sigma}_{\mathbf{x}} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\top} + \boldsymbol{\Psi}$$

where $\boldsymbol{\Gamma}$ is a $K \times L$ sparse matrix and $\boldsymbol{\Psi}$ is a diagonal matrix with diagonal entries $\psi_1, \ldots, \psi_K$.

This is equivalent to say that the latent variable $\mathbf{x}$ is generated as

$$\mathbf{x}|\mathbf{z} \sim \mathrm{N}_K(\boldsymbol{\Gamma}\mathbf{z}, \boldsymbol{\Psi}), \quad \mathbf{z} \sim \mathrm{N}_L(\mathbf{0}, \mathbf{I}).$$

Covariance matrix under the sparse variable clustering models

The covariance matrix $\boldsymbol{\Sigma}$ of $\mathbf{y}$ is given by

$$\boldsymbol{\Sigma} = \mathbf{A}(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top + \boldsymbol{\Psi})\mathbf{A}^\top + \phi\mathbf{I}$$

where (j, j')th entry is given by

$$\mathrm{Cov}(y_j, y_{j'}) = \boldsymbol{\gamma}_{\Bbbk(j)}^\top \boldsymbol{\gamma}_{\Bbbk(j')} + \psi_{\Bbbk(j)} \mathbb{1}_{\{\Bbbk(j)=\Bbbk(j')\}} + \phi \mathbb{1}_{\{j=j'\}}.$$

Prior

- Data: $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)}$, which are assumed to be independently generated as

$$\mathbf{y}^{(i)} | \mathbf{x}^{(i)} \sim \mathrm{N}_p(\mathbf{A}\mathbf{x}^{(i)}, \phi\mathbf{I}), \quad \mathbf{x}^{(i)} | \mathbf{z}^{(i)} \sim \mathrm{N}_K(\boldsymbol{\Gamma}\mathbf{z}^{(i)}, \boldsymbol{\Psi}), \quad \mathbf{z}^{(i)} \sim \mathrm{N}_L(\mathbf{0}, \mathbf{I}).$$

- Prior

$$\begin{aligned}
\pi(\mathbf{A}) &= \mathrm{CRP}(\mathbf{A}; \alpha) \\
\mathrm{K} &= \mathrm{ncol}(\mathbf{A}) \\
\pi\left(\mathbf{U}|\mathrm{K}\right) &= \mathrm{IBP}_{\mathrm{K}}(\mathbf{U}; \kappa) \\
\pi\left(\boldsymbol{\Gamma}|\mathbf{U}\right) &= \prod_{k=1}^{\mathrm{K}} \prod_{l=1}^{\infty} \{(1 - u_{k,l})\delta_0(\gamma_{k,l}) + u_{k,l}\mathrm{Lap}(\gamma_{k,l}; 1)\} \\
\pi(\boldsymbol{\Psi}|\mathrm{K}) &= \prod_{k=1}^{\mathrm{K}} \mathrm{IG}(\psi_k; a_\psi, b_\psi) \\
\pi(\phi) &= \mathrm{IG}(\phi; a_\phi, b_\phi)
\end{aligned}$$

MCMC sampler

For the sampling of **A** from the posterior distribution, we use the approximated distribution of the CRP by truncating the number of the "breaks" of the stick breaking representation of CRP. Recall that the stick breaking representation of the CRP is given by

$$\pi\left((\mathbb{k}(j))_{j\in[p]} | (v_h)_{h\in\mathbb{N}}\right) = \prod_{j=1}^{p} \mathrm{Mult}\left(\mathbb{k}(j); 1, \left(v_k \prod_{h=1}^{k-1}(1-v_h)\right)_{k\in\mathbb{N}}\right)$$

$$\pi\left((v_h)_{h\in\mathbb{N}}\right) = \prod_{h=1}^{\infty} \mathrm{Beta}(v_h; 1, \alpha).$$

where $\mathrm{Mult}(\cdot; n, \mathbf{p})$ denotes the multinomal distribution with a number of trials n and event probabilities $\mathbf{p} \equiv (p_k)_{k\in\mathbb{N}}$ with $\sum_{k=1}^{\infty} p_k = 1$.

For the sampling from the posterior distribution under the Laplace distribution prior, we use the fact that $\gamma \sim \mathrm{Lap}(1)$ if and only if $\gamma|\tau \sim \mathrm{N}(0,\tau)$, $\tau \sim \mathrm{Exp}(1/2)$.

MCMC sampler

- Sampling $\mathbf{a}_j$ for $j \in [p]$.
  Let K^* be the specified upper bound of the number of clusters. That is, we set $v_{K^*} = 1$ so that the prior probability that the each variable belongs to the $K^*, K^* + 1, \dots,$-th clusters.
  We sample $\Bbbk(j) \in \{1, \dots, K^*\}$ and let $\mathbf{a}_j = (\mathbb{1}_{\{\Bbbk(j)=1\}}, \dots, \mathbb{1}_{\{\Bbbk(j)=K^*\}})$. We update $\Bbbk(j)$ by multinomial sampling with

$$\pi(\Bbbk(j) = k|-) \propto \pi(\Bbbk(j) = k) \exp\left\{-(2\phi)^{-1} \sum_{i=1}^{n} \left(y_j^{(i)} - x_k^{(i)}\right)^2\right\}$$

  where $\pi(\Bbbk(j) = k) = v_k \prod_{h<k}(1 - v_h)$.
- Sampling v_k for $k \in [K^*]$.
  We update the stick-breaking weight v_k for $k \in [K^* - 1]$ by the sampling

$$v_k|- \sim \mathrm{Beta}\left(1 + np_k, \alpha + n \sum_{h=k+1}^{K^*} p_h\right)$$

  where $p_k := |\{j \in [p] : \Bbbk(j) = k\}|$.

MCMC sampler

- Sampling $u_{k,l}$ for $k \in [K]$ and $l \in \mathbb{N}$
  Let L^* be the number of columns of $\boldsymbol{\Gamma}$. Let $K_{-k,l} = \sum_{k' \neq k} u_{k',l}$. We first update $u_{k,l}$ for $k \in [K]$ and $l \in [L^*]$ by binary sampling with probability

$$\frac{\Pi(u_{k,l} = 1|-)}{\Pi(u_{k,l} = 0|-)} = \frac{K_{-k,l}}{K^* - K_{-k,l}} \sqrt{\frac{\hat{\tau}_{k,l}}{\tau_{k,l}}} \exp\left(\frac{1}{2\hat{\tau}_{k,l}} \hat{\gamma}_{k,l}^2\right)$$

  where

$$\hat{\tau}_{k,l} := \left(\psi_k^{-1} \sum_{i=1}^{n} \left(z_k^{(i)}\right)^2 + \tau_{k,l}^{-1}\right)^{-1}$$

$$\hat{\gamma}_{k,l} := \hat{\tau}_{k,l} \left\{\psi_k^{-1} \sum_{i=1}^{n} z_l^{(i)} \left(x_k^{(i)} - \sum_{l' \neq l} \gamma_{k,l'} z_{l'}^{(i)}\right)\right\}.$$

MCMC sampler

- Sampling $u_{k,l}$ for $k \in [K]$ and $l \in \mathbb{N}$ (cont'd)
  For sampling new columns, we propose $\tilde{L}_k \in \mathbb{N}_0$ and $\tilde{\gamma}_k \in \mathbb{R}^{\tilde{L}_k}$ from the prior distribution as
  $$(\tilde{L}_k, \mathbf{M}_k) \sim \mathrm{Pois}(\kappa/(K^* - 1)) \left\{\mathrm{Lap}(1)\right\}^{\tilde{L}_k}$$
  and accept the proposal with probability
  $$\max\left\{1, |2\pi\mathbf{M}_k|^{-\frac{K^*}{2}} \exp\left(\left(\frac{1}{2}e_{k,l}^2\right)\tilde{\gamma}_k^\top \mathbf{M}_k^{-1}\tilde{\gamma}_k\right)\right\}$$
  where $\mathbf{M}_k = \psi_k^{-1}\tilde{\gamma}_k\tilde{\gamma}_k^\top + \mathbf{I}$ and $e_{k,l} = \psi_k^{-1}\sum_{i=1}^n \left(x_k^{(i)} - \boldsymbol{\Gamma}_{[k,:]}^\top \mathbf{z}^{(i)}\right)$. If the proposal is accepted, update the current $\boldsymbol{\Gamma}$ by setting $\boldsymbol{\Gamma}_{[k,[L^*+\tilde{L}_k]\setminus[L^*]]} = \tilde{\gamma}_k$ and L^* by $L^* + \tilde{L}_k$.

MCMC sampler

- Sampling $\gamma_{k,l}$ for for $k \in [K]$ and $l \in [L^*]$
  If $u_{k,l} = 1$, update $\gamma_{k,l}$ by sampling
  $$\gamma_{k,l} | - \sim \mathrm{N}\left(\hat{\gamma}_{k,l}, \hat{\tau}_{k,l}\right)$$
  and sampling $\gamma_{k,l}|- \sim \mathrm{N}(0, \tau_{k,l})$ otherwise
- Sampling $\tau_{k,l}$ for for $k \in [K]$ and $l \in [L^*]$
  If $u_{k,l} = 1$, update $\tau_{k,l}$ by sampling from the following distribution
  $$\pi(\tau_{k,l}|-) \propto \tau_{k,l}^{-1/2} \exp\left(-\frac{1}{2}\left(\tau_{k,l} + \frac{\gamma_{k,l}^2}{\tau_{k,l}}\right)\right)$$
  which is equivalent to sampling from the generalized inverse Gaussian distribution and sampling $\tau_{k,l} \sim \mathrm{Exp}(1/2)$ otherwise.

MCMC sampler

- Sampling $\mathbf{x}^{(i)}$ for $i \in [n]$

$$\mathbf{x}^{(i)}|- \sim \mathrm{N}\left(\left(\phi^{-1}\mathbf{A}^\top\mathbf{A} + \boldsymbol{\Psi}^{-1}\right)^{-1}\left(\boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma}\mathbf{z}^{(i)} + \phi^{-1}\mathbf{A}^\top\mathbf{y}^{(i)}\right), \left(\phi^{-1}\mathbf{A}^\top\mathbf{A} + \boldsymbol{\Psi}^{-1}\right)^{-1}\right)$$

- Sampling $\mathbf{z}^{(i)}$ for $i \in [n]$

$$\mathbf{z}^{(i)}|- \sim \mathrm{N}\left(\boldsymbol{\Psi}^{-1}\left(\boldsymbol{\Gamma}\boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma} + \mathbf{I}\right)^{-1}\boldsymbol{\Gamma}^\top\mathbf{x}^{(i)}, \left(\boldsymbol{\Gamma}\boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma} + \mathbf{I}\right)^{-1}\right).$$

- Sampling ψ_k for $k \in [K^*]$

$$\psi_k|- \sim \mathrm{IG}\left(a_\psi + \frac{n}{2}, b_\psi + \frac{1}{2}\sum_{i=1}^{n}\left(x_k^{(i)} - \boldsymbol{\Gamma}_{[k,:]}^\top\mathbf{z}^{(i)}\right)^2\right).$$

- Sampling ϕ

$$\phi|- \sim \mathrm{IG}\left(a_\phi + \frac{np}{2}, b_\phi + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(y_j^{(i)} - \mathbf{a}_j^\top\mathbf{x}^{(i)}\right)^2\right)$$