Summary and discussion of: "Sparse Bayesian Methods for Low-Rank Matrix Estimation"

HKUST MATH5472 Final Project

WANG Gefei

1 Summary

This paper proposes a Variational Bayesian approach for low-rank matrices estimation [1].

For an observation matrix $\mathbf{Y} \in \mathbb{R}^{m\times n}$, we aim to an unknown low-rank matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$, $\operatorname{rank}(\mathbf{X}) = r \ll \min(m,n)$ such that $\mathbf{Y}$ is a function $f(\mathbf{X})$ of $\mathbf{X}$. For example, in matrix completion, $\mathbf{Y}$ is a projection of $\mathbf{X}$, where

$$\mathbf{Y}_{ij} = \mathcal{P}_\Omega(\mathbf{X})_{ij} = \begin{cases} \mathbf{X}_{ij}, & \text{if } (i,j) \in \Omega \\ 0, & \text{otherwise.} \end{cases}$$

In Robust Principal Component Analysis (RPCA), $\mathbf{Y} = \mathbf{X} + \mathbf{E}$, where $\mathbf{E}$ is a sparse error matrix.

The low-rank matrix $\mathbf{X}$ can be recoverd by solving

$$\begin{aligned} \text{minimize} \quad & \operatorname{rank}(\mathbf{X}) \\ \text{subject to} \quad & \mathbf{Y} = f(\mathbf{X}). \end{aligned}$$

To solve the problem more efficiently, an alternative approach is the convex relaxation of the original problem given by

$$\begin{aligned} \text{minimize} \quad & \|\mathbf{X}\|_* \\ \text{subject to} \quad & \mathbf{Y} = f(\mathbf{X}), \end{aligned}$$

where the nuclear norm $\|\cdot\|_*$ equals to the sum of the singular values of a matrix. Furthermore, the constraint can be relaxed, since the observation are often corrupted with dense noise. The relaxed problem becomes

$$\begin{aligned} \text{minimize} \quad & \|\mathbf{X}\|_* \\ \text{subject to} \quad & \|\mathbf{Y} - f(\mathbf{X})\|_F^2 < \epsilon, \end{aligned}$$

where $\|\cdot\|_F$ is the Frobenius norm. This formulation is equivalent to

$$\begin{aligned} \text{minimize} \quad & \|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2 \\ \text{subject to} \quad & \|\mathbf{Y} - f(\mathbf{A}\mathbf{B}^T)\|_F^2 < \epsilon, \end{aligned}$$

where $\mathbf{A} \in \mathbb{R}^{m \times r}$, $\mathbf{B} \in \mathbb{R}^{n \times r}$, $r \leq \min(m, n)$ and $\mathbf{X} = \mathbf{A}\mathbf{B}^T$, as long as r is large engough [2].

This paper adopts a Bayesian way to model this problem. They represent $\mathbf{X}$ as

$$\mathbf{X} = \mathbf{A}\mathbf{B}^T = \sum_{i=1}^{k} \mathbf{a}_{\cdot i}\mathbf{b}_{\cdot i}^T,$$

where $k \geq \text{rank}(\mathbf{X})$. To enforce most columns of $\mathbf{A}$ and $\mathbf{B}$ to zero, they associate the columns of $\mathbf{A}$ and $\mathbf{B}$ with Gaussian priors, parameterized with precisions $\{\gamma_i\}_{i=1}^k$:

$$p(\mathbf{A}|\boldsymbol{\gamma}) = \prod_{i=1}^{k} \mathcal{N}(\mathbf{a}_{\cdot i}|\mathbf{0}, \gamma_i^{-1}\mathbf{I}_m),$$
$$p(\mathbf{B}|\boldsymbol{\gamma}) = \prod_{i=1}^{k} \mathcal{N}(\mathbf{b}_{\cdot i}|\mathbf{0}, \gamma_i^{-1}\mathbf{I}_n).$$

If γ_i is very large, both the columns $\mathbf{a}_{\cdot i}$ and $\mathbf{b}_{\cdot i}$ will be set to zero.

Additionally, they use a conjugate Gamma hyperprior for the precisions $\{\gamma_i\}_{i=1}^k$:

$$p(\gamma_i) = \text{Gamma}(a, \frac{1}{b}) \propto \gamma_i^{a-1} \exp(-b\gamma_i),$$

where a and b are set to small values (10^{-6}) to make the hyperprior broad.

1.1 Matrix Completion

For the matrix completion the observation is modeled as

$$\mathbf{Y} = \mathcal{P}_\Omega(\mathbf{X} + \mathbf{N}),$$

where $\mathbf{N}$ is the dense error. They assume the noise is white Gaussian with precision $\beta = \frac{1}{\epsilon}$, then

$$p(\mathbf{Y}|\mathbf{A}, \mathbf{B}, \beta) = \prod_{(i,j)\in\Omega} \mathcal{N}(Y_{ij}|X_{ij}, \beta^{-1}).$$

The precision β has a non-informative Jeffrey's prior

$$p(\beta) = \beta^{-1}.$$

The joint distribution is

$$p(\mathbf{Y}, \mathbf{A}, \mathbf{B}, \boldsymbol{\gamma}, \beta) = p(\mathbf{Y}|\mathbf{A}, \mathbf{B}, \beta)p(\mathbf{A}|\boldsymbol{\gamma})p(\mathbf{B}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})p(\beta).$$

Exact posterior of the latent variables $\mathbf{z} = (\mathbf{A}, \mathbf{B}, \boldsymbol{\gamma}, \beta)$ is intractable because the marginal $p(\mathbf{Y})$ is intractable. Instead, the authors adopt the mean-field variational inference method to estimate the posterior.

The complete data log-likelihood is

$$\begin{aligned}
&\log p(\mathbf{Y}, \mathbf{A}, \mathbf{B}, \boldsymbol{\gamma}, \beta) \\
=&\sum_{(i,j)\in\Omega} \frac{1}{2}\left[\log\beta - \beta(Y_{ij} - X_{ij})^2\right] + \sum_{i=1}^{k} \frac{1}{2}\left[(m+n)\log\gamma_i - \gamma_i(\|\mathbf{a}_{\cdot i}\|_2^2 + \|\mathbf{b}_{\cdot i}\|_2^2)\right] \\
&+ (a-1)\sum_{i=1}^{k}\log\gamma_i - b\sum_{i=1}^{k}\gamma_i - \log\beta + \text{const.}
\end{aligned}$$

The optimal posterior approximation of $\mathbf{a}_{i\cdot}$) is obtained by taking expection of the complete data log-likelihood w.r.t. $q(\mathbf{z}\backslash\mathbf{a}_{i\cdot})$. The result is still quadratic, thus

$$q(\mathbf{a}_{i\cdot}^T) = \mathcal{N}(\boldsymbol{\mu}_i^a, \boldsymbol{\Sigma}_i^a),$$

where

$$\begin{aligned}
\boldsymbol{\mu}_i^a =& \mathbb{E}[\beta]\boldsymbol{\Sigma}_i^a\mathbb{E}[\mathbf{B}_i]^T\mathbf{y}_{i\cdot}^T, \\
\boldsymbol{\Sigma}_i^a =& (\mathbb{E}[\beta]\mathbb{E}[\mathbf{B}_i^T\mathbf{B}_i] + \boldsymbol{\Gamma})^{-1},
\end{aligned}$$

where $\mathbf{B}_i$ contains only the j-th rows of $\mathbf{B}$ such that $(i,j)\in\Omega$,

$$\mathbb{E}[\mathbf{B}_i^T\mathbf{B}_i] = \sum_{j:(i,j)\in\Omega}\mathbb{E}[\mathbf{b}_{j\cdot}^T\mathbf{b}_{j\cdot}] = \sum_{j:(i,j)\in\Omega}\left[\mathbb{E}[\mathbf{b}_{j\cdot}]^T\mathbb{E}[\mathbf{b}_{j\cdot}] + \boldsymbol{\Sigma}_j^b\right] = \mathbb{E}[\mathbf{B}_i]^T\mathbb{E}[\mathbf{B}_i] + \sum_{j:(i,j)\in\Omega}\boldsymbol{\Sigma}_j^b,$$

$\mathbf{y}_{i\cdot}$ contains only the observed entries of the i-th row of $\mathbf{Y}$, and $\boldsymbol{\Gamma} = \text{diag}(\mathbb{E}[\gamma_1], \cdots, \mathbb{E}[\gamma_k])$. Similarly,

$$q(\mathbf{b}_{j\cdot}^T) = \mathcal{N}(\boldsymbol{\mu}_j^b, \boldsymbol{\Sigma}_j^b),$$

where

$$\begin{aligned}
\boldsymbol{\mu}_j^b =& \mathbb{E}[\beta]\boldsymbol{\Sigma}_j^b\mathbb{E}[\mathbf{A}_j]^T\mathbf{y}_{\cdot j}, \\
\boldsymbol{\Sigma}_j^b =& (\mathbb{E}[\beta]\mathbb{E}[\mathbf{A}_j^T\mathbf{A}_j] + \boldsymbol{\Gamma})^{-1}.
\end{aligned}$$

The posterior of γ_i is a Gamma distribution with mean

$$\begin{aligned}
\mathbb{E}[\gamma_i] =& \frac{2a+m+n}{2b + \mathbb{E}[\|\mathbf{a}_{\cdot i}\|^2] + \mathbb{E}[\|\mathbf{b}_{\cdot i}\|^2]} \\
=& \frac{2a+m+n}{2b + \mathbb{E}[\mathbf{a}_{\cdot i}]^T\mathbb{E}[\mathbf{a}_{\cdot i}] + \sum_{j=1}^m(\boldsymbol{\Sigma}_j^a)_{ii} + \mathbb{E}[\mathbf{b}_{\cdot i}]^T\mathbb{E}[\mathbf{b}_{\cdot i}] + \sum_{j=1}^n(\boldsymbol{\Sigma}_j^b)_{ii}}.
\end{aligned}$$

The posterior of β is also a Gamma distribution with mean

$$\mathbb{E}[\beta] = \frac{N}{\mathbb{E}[\|\mathbf{Y} - \mathcal{P}_\Omega(\mathbf{A}\mathbf{B}^T)\|_F^2]} \approx \frac{pmn}{\mathbb{E}[\|\mathbf{Y} - \mathcal{P}_\Omega(\mathbf{A}\mathbf{B}^T)\|_F^2]},$$

where N is the number of observed entries in $\mathbf{Y}$, and p is the fraction of observed entries in $\mathbf{Y}$. Then, the posterior means of the latent variables can be estimated iteratively. Columns of $\mathbf{A}$ and $\mathbf{B}$ will be removed if γ_i is large enough during the iterations.

1.2 Robust PCA

For the RPCA the observation is modeled as

$$\mathbf{Y} = \mathbf{X} + \mathbf{E} + \mathbf{N},$$

where $\mathbf{E}$ is the sparse error and $\mathbf{N}$ is the dense error. The conditional distribution for the observation is given by

$$\begin{aligned} p(\mathbf{Y}|\mathbf{A},\mathbf{B},\mathbf{E},\beta) =& \mathcal{N}(\mathbf{Y}|\mathbf{A}\mathbf{B}^T + \mathbf{E}, \beta^{-1}\mathbf{I}_{mn}) \\ \propto& \exp\left(\frac{\beta}{2}\|\mathbf{Y} - \mathbf{A}\mathbf{B}^T - \mathbf{E}\|_F^2\right). \end{aligned}$$

The sparse component $\mathbf{E}$ is modeled with independent Gaussian priors on each entry

$$p(\mathbf{E}|\boldsymbol{\alpha}) = \prod_{i=1}^{m}\prod_{j=1}^{n}\mathcal{N}(E_{ij}|0,\alpha_{ij}^{-1}).$$

The precision $\boldsymbol{\alpha}$ also has a non-informative Jeffrey's prior

$$p(\alpha_{ij}) = \alpha_{ij}^{-1}.$$

The joint distribution is

$$p(\mathbf{Y},\mathbf{A},\mathbf{B},\mathbf{E},\boldsymbol{\gamma},\boldsymbol{\alpha},\beta) = p(\mathbf{Y}|\mathbf{A},\mathbf{B},\mathbf{E},\beta)p(\mathbf{A}|\boldsymbol{\gamma})p(\mathbf{B}|\boldsymbol{\gamma})p(\mathbf{E}|\boldsymbol{\alpha})p(\boldsymbol{\gamma})p(\boldsymbol{\alpha})p(\beta).$$

Exact posterior of the latent variables $\mathbf{z} = (\mathbf{Y},\mathbf{A},\mathbf{B},\mathbf{E},\boldsymbol{\gamma},\boldsymbol{\alpha},\beta)$ is again intractable.

The complete data log-likelihood is

$$\begin{aligned} &\log p(\mathbf{Y},\mathbf{A},\mathbf{B},\mathbf{E},\boldsymbol{\gamma},\boldsymbol{\alpha},\beta) \\ =& \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1}{2}\left[\log\beta - \beta(Y_{ij} - \mathbf{a}_{i\cdot}\mathbf{b}_{j\cdot}^T - E_{ij})^2\right] + \sum_{i=1}^{k}\frac{1}{2}\left[(m+n)\log\gamma_i - \gamma_i(\|\mathbf{a}_{\cdot i}\|_2^2 + \|\mathbf{b}_{\cdot i}\|_2^2)\right] \\ &+ \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1}{2}\left[\log\alpha_{ij} - \alpha_{ij}E_{ij}^2\right] - \sum_{i=1}^{m}\sum_{j=1}^{n}\log\alpha_{ij} + (a-1)\sum_{i=1}^{k}\log\gamma_i - b\sum_{i=1}^{k}\gamma_i - \log\beta + \text{const.} \end{aligned}$$

The optimal posterior approximation of $\mathbf{a}_{i\cdot}$ is

$$q(\mathbf{a}_{i\cdot}^T) = \mathcal{N}(\boldsymbol{\mu}_i^a, \boldsymbol{\Sigma}^A),$$

where

$$\begin{aligned} \boldsymbol{\mu}_i^a =& \mathbb{E}[\beta]\boldsymbol{\Sigma}^A\mathbb{E}[\mathbf{B}]^T(\mathbf{y}_{i\cdot} - \mathbf{e}_{i\cdot})^T, \\ \boldsymbol{\Sigma}^A =& (\mathbb{E}[\beta]\mathbb{E}[\mathbf{B}^T\mathbf{B}] + \boldsymbol{\Gamma})^{-1}. \end{aligned}$$

Similarly,

$$q(\mathbf{b}_{j\cdot}^T) = \mathcal{N}(\boldsymbol{\mu}_j^b, \boldsymbol{\Sigma}^B),$$

where

$$\begin{aligned}\boldsymbol{\mu}_j^b =& \mathbb{E}[\beta]\boldsymbol{\Sigma}^B\mathbb{E}[\mathbf{A}]^T(\mathbf{y}_{\cdot j} - \mathbf{e}_{\cdot j}),\\ \boldsymbol{\Sigma}^B =& (\mathbb{E}[\beta]\mathbb{E}[\mathbf{A}^T\mathbf{A}] + \boldsymbol{\Gamma})^{-1}.\end{aligned}$$

$$q(E_{ij}) = \mathcal{N}(\mu_{ij}^E, (\sigma_{ij}^E)^2),$$

where

$$\begin{aligned}\mu_{ij}^E =& \mathbb{E}[\beta](\sigma_{ij}^E)^2(Y_{ij} - (\boldsymbol{\mu}_i^a)^T\boldsymbol{\mu}_j^b),\\ (\sigma_{ij}^E)^2 =& \frac{1}{\mathbb{E}[\beta] + \mathbb{E}[\alpha_{ij}]}.\end{aligned}$$

The posterior of γ_i is a Gamma distribution with mean

$$\begin{aligned}\mathbb{E}[\gamma_i] =& \frac{2a + m + n}{2b + \mathbb{E}[\|\mathbf{a}_{\cdot i}\|^2] + \mathbb{E}[\|\mathbf{b}_{\cdot i}\|^2]}\\ =& \frac{2a + m + n}{2b + \mathbb{E}[\mathbf{a}_{\cdot i}]^T\mathbb{E}[\mathbf{a}_{\cdot i}] + m(\boldsymbol{\Sigma}^A)_{ii} + \mathbb{E}[\mathbf{b}_{\cdot i}]^T\mathbb{E}[\mathbf{b}_{\cdot i}] + n(\boldsymbol{\Sigma}^B)_{ii}}.\end{aligned}$$

The posterior of $\boldsymbol{\alpha}$ is also a Gamma distribution with mean

$$\mathbb{E}[\alpha_{ij}] = \frac{1}{\mathbb{E}[E_{ij}]^2 + \Sigma_{ij}^E}.$$

The posterior of β is also a Gamma distribution with mean

$$\mathbb{E}[\beta] = \frac{mn}{\mathbb{E}[\|\mathbf{Y} - \mathbf{A}\mathbf{B}^T - \mathbf{E}\|_F^2]},$$

where

$$\begin{aligned}\mathbb{E}[\|\mathbf{Y} - \mathbf{A}\mathbf{B}^T - \mathbf{E}\|_F^2] =& \|\mathbf{Y} - \mathbb{E}[\mathbf{A}]\mathbb{E}[\mathbf{B}]^T - \mathbb{E}[\mathbf{E}]\|_F^2 + \mathrm{Tr}(n\mathbb{E}[\mathbf{A}]^T\mathbb{E}[\mathbf{A}]\boldsymbol{\Sigma}^B)\\ &+ \mathrm{Tr}(m\mathbb{E}[\mathbf{B}]^T\mathbb{E}[\mathbf{B}]\boldsymbol{\Sigma}^A) + \mathrm{Tr}(mn\boldsymbol{\Sigma}^A\boldsymbol{\Sigma}^B) + \sum_{i=1}^m\sum_{j=1}^n \Sigma_{ij}^E.\end{aligned}$$

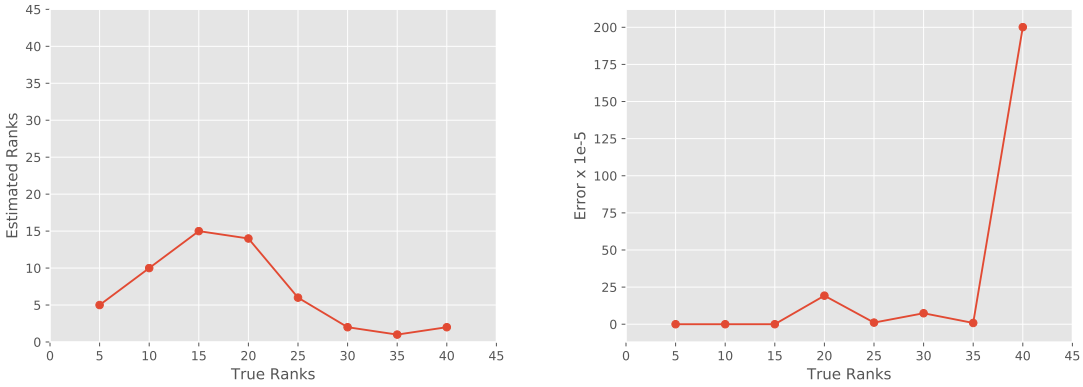
2 Result and Discussion

We implement both the algorithms in Python. Our code for the experiments is available at `https://github.com/gefeiwang/MATH5472_Projects/tree/master/Project1`.

2.1 Matrix Completion

We test the model with simulated data. The underlying low-rank matrices $\mathbf{X}$ of size 300×300 of ranks $r = 5, 10, 15, 20$ is calculated by $\mathbf{X} = \mathbf{A}\mathbf{B}^T$, where $\mathbf{A}$ and $\mathbf{B}$ are $300 \times r$ matrices with entries sampled from $\mathcal{N}(0, 1)$. The fraction of observed entries is $p = 0.2$, and the dense error $\mathbf{N}$ is sampled from $\mathcal{N}(0, 0.05)$. We initial the estimated rank to be $r = 50$.

The estimated ranks and reconstruction errors of $\mathbf{X}$ (take average over entries) are shown in Figure 1. When the rank of the underlying matrices is small enough, the algorithm in section 1.1, named VBMC, estimates the ranks and recovers the low-rank matrices $\mathbf{X}$ accurately. However, when the rank increases, the model tends to severely under-estimate the rank, and the reconstruction error becomes very large.

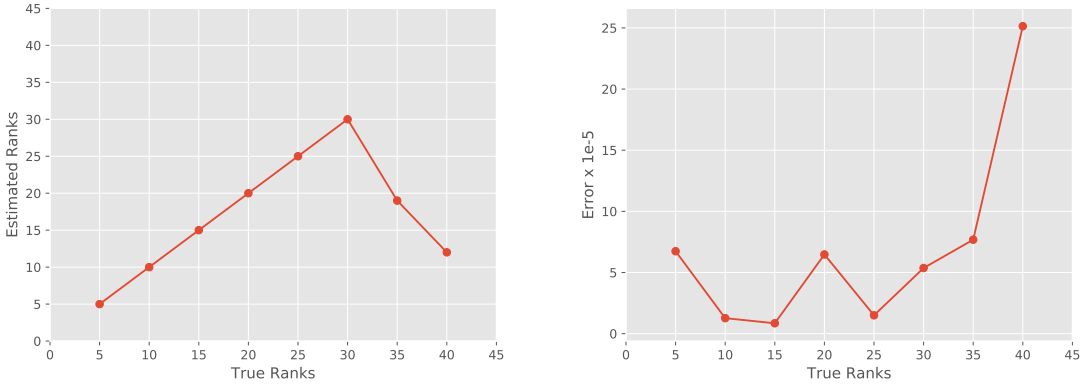


(a) (b)

Figure 1: Simulation studies for matrix completion.

2.2 Robust PCA

We use simulated data similar to previous setting. The underlying low-rank matrices $\mathbf{X}$ of size 300×300 of ranks $r = 5, 10, 15, 20$ is calculated by $\mathbf{X} = \mathbf{A}\mathbf{B}^T$, where $\mathbf{A}$ and $\mathbf{B}$ are $300 \times r$ matrices with entries sampled from $\mathcal{N}(0, 1)$. The fraction of non-zero entries of the sparse error term $\mathbf{E}$ is $p = 0.05$, and the non-zero entries are samples from $U(-10, 10)$. The dense error $\mathbf{N}$ is again sampled from $\mathcal{N}(0, 0.05)$. We initial the estimated rank to be $r = 50$. The estimated ranks and reconstruction errors of $\mathbf{X}$ (take average over entries) are shown in Figure 2. When the rank of the underlying matrices is small enough, the algorithm in section 1.2, named VBRPCA, estimates the ranks and recovers the low-rank matrices $\mathbf{X}$ accurately. However, like the previous experiment, when the rank increases, the model also tends to under-estimate the rank.



(a) (b)

Figure 2: Simulation studies for matrix completion.

3 Conclusion

In this report, we discuss about the Variational Bayesian approach for low-rank matrices estimation with detailed implementations in matrix completion and RPCA, introduced in [1]. Simulation studies show that this model is powerful when the matrices $\mathbf{X}$ are indeed low-rank with respect to the matrices size.

References

[1] S. D. Babacan, M. Luessi, R. Molina, and A. K. Katsaggelos. Sparse bayesian methods for low-rank matrix estimation. *IEEE Transactions on Signal Processing*, 60(8):3964–3977, 2012.

[2] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.