Supplement to "Personalized Prediction and Sparsity Pursuit in Latent Factor Models"

Yunzhang Zhu, Xiaotong Shen and Changqing Ye

We provide supplementary technical proofs for "Personalized Prediction and Sparsity Pursuit in Latent Factor Models". We use the prefix "PP" when referring to lemmas, equations etc., in the former paper, as in equation (PP-1) or Lemma PP-1.

Proof of Lemma PP-1 Note that $\hat{\boldsymbol{A}}\hat{\boldsymbol{B}}^T = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{B}}^T$. For the L_1-penalty function,

$$
\begin{aligned}
\|\tilde{\boldsymbol{A}}\|_1 + \|\tilde{\boldsymbol{B}}\|_1 &= \sum_{i=1}^{K}\left(\|\tilde{\boldsymbol{A}}_i\|_1 + \|\tilde{\boldsymbol{B}}_i\|_1\right) \geq 2\sum_{i=1}^{K}\sqrt{\|\boldsymbol{A}_i\|_1\|\boldsymbol{B}_i\|_1} \\
&= \sum_{i=1}^{K}\left(\|\hat{\boldsymbol{A}}_i\|_1 + \|\hat{\boldsymbol{B}}_i\|_1\right) = \|\hat{\boldsymbol{A}}\|_1 + \|\hat{\boldsymbol{B}}\|_1.
\end{aligned}
$$

Proof of Lemma PP-2: The proof for convergence of the algorithms follows the same argument as in Theorem 3.1 of [3], thus is omitted. To show that $(\hat{\boldsymbol{A}}^{L_1}, \hat{\boldsymbol{B}}^{L_1})$ satisfies (PP-3), note that if $\hat{\boldsymbol{b}}_i = \boldsymbol{0}$ then it follows from (PP-11) that for any $\boldsymbol{a}_j$

$$
\sum_{i\in R_{j.}} l(r_{ji}, \boldsymbol{x}_j^T\hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{\beta}}^T\boldsymbol{y}_i + \boldsymbol{a}_j^T\hat{\boldsymbol{b}}_i) + \lambda\|\boldsymbol{a}_j\|_1 \geq \sum_{i\in R_{j.}} l(r_{ji}, \boldsymbol{x}_j^T\hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{\beta}}^T\boldsymbol{y}_i + \boldsymbol{0}\hat{\boldsymbol{b}}_i) + \lambda\|\boldsymbol{0}\|_1,
$$

implying that the minimizer $\hat{\boldsymbol{a}}_j = \boldsymbol{0}$. The same holds for $\hat{\boldsymbol{b}}_i$ if $\hat{\boldsymbol{a}}_j = \boldsymbol{0}$. This completes the proof.

Proof of Lemma PP-3: Consider SVD of $\boldsymbol{\Theta}$ as follows: $\boldsymbol{\Theta} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$, where $\boldsymbol{D}$ is a diagonal matrix whose ith diagonal element is the ith singular value of $\boldsymbol{\Theta}$, and $\boldsymbol{U}$ and $\boldsymbol{V}$ are the left and right orthogonal matrices for the SVD. Note that $\|\boldsymbol{A}\|_2$ is rotation-invariant under any orthogonal-transformation. Hence it suffices to consider the constraint that $\boldsymbol{A}\boldsymbol{B}^T = \boldsymbol{D}$, equivalently, $\langle\boldsymbol{a}_j, \boldsymbol{b}_i\rangle = 0$ when $j \neq i$ and $\langle\boldsymbol{a}_i, \boldsymbol{b}_i\rangle = \sigma_i$; $i = 1, \cdots, K$. By the triangular inequality,

$$
\|\boldsymbol{A}\|_2^2 + \|\boldsymbol{B}\|_2^2 \geq 2\sum_{i=1}^{\min(U,M)}|\langle\boldsymbol{a}_i, \boldsymbol{b}_i\rangle| = 2\sum_{i=1}^{\min(U,M)}|\sigma_i| = 2\|\boldsymbol{\Theta}\|_*.
$$

implying that the minimal of (PP-5) is no smaller than that of (PP-18), where the equality holds when $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}^{1/2}$ and $\boldsymbol{B} = \boldsymbol{V}\boldsymbol{D}^{1/2}$. This establishes the equivalence. Finally,

uniqueness of the solution of (PP-18) follows from strict convexity of (PP-18) in $\boldsymbol{\Theta}$. This completes the proof.

Proof of Theorem PP-1: The proof uses a large deviation probability inequality of [4] to treat one-sided regularized log-likelihood ratios.

Let $f(\boldsymbol{R}, \{z_{ji}\}, \boldsymbol{\Theta})$ be the probability density of $(\boldsymbol{R}, \{z_{ji}\})$ given $(\boldsymbol{x}_j, \boldsymbol{y}_i)$ at $\boldsymbol{\Theta}$. Let $\mathcal{F}_0(s) = \{(\boldsymbol{A}, \boldsymbol{B}) \in \mathcal{F} : \boldsymbol{\Theta} = \boldsymbol{A}\boldsymbol{B}^T, \|\boldsymbol{A}\|_0 + \|\boldsymbol{B}\|_0 \leq s\}$ be a L_0-constrained parameter space.

First we bound the bracketing u Hellinger metric entropy $H(u, \mathcal{F}_0(s))$ [6]. To define this quantity, for any $u > 0$, call a finite set of pairs of functions $\{(f_j^L, f_j^U), j = 1, \cdots, N\}$ a (Hellinger) u-bracketing of $\mathcal{F}$ if $\|(f_j^L)^{1/2} - (f_j^U)^{1/2}\|_2 \leq u$ for $j = 1, \cdots, N$, and for any $f \in \mathcal{F}$, there is a j such that $f_j^L \leq f \leq f_j^U$. The bracketing Hellinger metric entropy of $\mathcal{F}$, denoted by the function $H(\cdot, \mathcal{F})$, is defined by $H(u, \mathcal{F}) =$ logarithm of the cardinality of the u-bracketing (of $\mathcal{F}$) of the smallest size. For $\boldsymbol{\Theta} = \boldsymbol{A}\boldsymbol{B}^T$ with $(\boldsymbol{A}, \boldsymbol{B}) \in \mathcal{F}_0(s)$, let $B_\delta(\boldsymbol{A}, \boldsymbol{B}) = \{(\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{B}}) \in \mathcal{F}_0(s) : \sqrt{\|\tilde{\boldsymbol{A}} - \boldsymbol{A}\|_{F*}^2 + \|\tilde{\boldsymbol{B}} - \boldsymbol{B}\|_{F*}^2} \leq \delta\}$ be a ball centered at $(\boldsymbol{A}, \boldsymbol{B})$. Note that $0 \leq \delta_{ji} \leq 1$. For any $\tilde{\boldsymbol{\Theta}} = (\tilde{\theta}_{ji}) = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{B}}^T$ with $(\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{B}}) \in B_\delta(\boldsymbol{A}, \boldsymbol{B})$,

$$(MU)^{-1} \sum_{j=1}^{M} \sum_{i=1}^{U} \int \sup_{\{(\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{B}}) \in B_\delta(\boldsymbol{A}, \boldsymbol{B})\}} (f^{1/2}(r_{ji}, z_{ji}, \tilde{\theta}_{ji}) - f^{1/2}(r_{ji}, z_{ji}, \theta_{ji}))^2 d\mu(z_{ji})$$

$$\leq \frac{d_0^2}{MU} \sup_{\{(\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{B}}) \in B_\delta(\boldsymbol{A}, \boldsymbol{B})\}} 2(\|\tilde{\boldsymbol{A}} - \boldsymbol{A}\|_{F*}^2 \|\tilde{\boldsymbol{B}}\|_{F*}^2 + \|\tilde{\boldsymbol{B}} - \boldsymbol{B}\|_{F*}^2 \|\boldsymbol{A}\|_{F*}^2)$$

$$\leq \frac{d_0^2}{MU} \sup_{\{(\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{B}}) \in B_\delta(\boldsymbol{A}, \boldsymbol{B})\}} 2sL^2(\|\tilde{\boldsymbol{A}} - \boldsymbol{A}\|_{F*}^2 + \|\tilde{\boldsymbol{B}} - \boldsymbol{B}\|_{F*}^2),$$

where $\|\cdot\|_{F*}$ is the Frobenius-norm whose jith element is taken over $\sup_{(\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{B}}) \in B_\delta(\boldsymbol{A}, \boldsymbol{B})}$, and the fact that $\|\boldsymbol{A}\boldsymbol{B}^T\|_{F*}^2 \leq \|\boldsymbol{A}\|_{F*}^2 \|\boldsymbol{B}\|_{F*}^2$ has been used. By Lemma 1 of [5], it suffices to bound the entropy of $B_\delta(\boldsymbol{A}, \boldsymbol{B})$. Note that there are s nonzero elements of $(\boldsymbol{A}, \boldsymbol{B})$ with $\binom{(M+U)K}{s}$ possible locations. Then for $u \geq \varepsilon_{0,|\Omega|}^2$,

$$H(u, \mathcal{F}_0(s)) \leq \log \binom{(M+U)K}{s} + H_2(\frac{u}{Ld_0\sqrt{2s}(MU)^{-1/2}}, \mathcal{F}_{s,L})$$

$$\leq s \log \left(e\frac{(M+U)K}{s}\right) + s \log \left(\max\left(\frac{L^2 d_0 \sqrt{2s}}{u(MU)^{1/2}}, 1\right)\right), \tag{1}$$

where $\mathcal{F}_{s,L} = \{\boldsymbol{x} \in \mathbb{R}^s, \|\boldsymbol{x}\|_\infty \leq L\}$, $H_2(\cdot, \mathcal{F}_{s,L})$ is the ℓ_2-metric entropy $\mathcal{F}_{s,L}$ and inequality $\binom{n}{m} \leq \left(e\frac{n}{m}\right)^m$ has been used, c.f., Theorem 2.6 of [1].

To apply Theorem 1 of [4], we verify the required entropy condition there for $\mathcal{F}$:

$$\sup_{\{s \geq s_0\}} \psi_1(\varepsilon, s) \leq c_2 |\Omega|^{1/2}, \tag{2}$$

where $\psi_1(x, s) = \int_x^{x^{1/2}} H^{1/2}(u, \mathcal{F}_0(s)) du/x$ with $x = (c_1 \varepsilon^2 + \lambda(s-1))$. Using the entropy bound in (2), we obtain $\psi_1(x, s) = \frac{\varepsilon_{0,|\Omega|}(1-x^{1/2})}{x^{1/2}}$ for $0 < x \leq 1$. Here we only focus our

2

attention to the case of $x \leq 1$ because (2) is met automatically when $x > 1$. Note that $\psi_1(x, s)$ is nonincreasing in s for any fixed $x > 0$. Then $\sup_{\{s \geq s_0\}} \psi_1(\varepsilon, s) = \psi_1(\varepsilon, s_0)$. Solving (2) yields that $c_1 \varepsilon^2 + \lambda(s_0 - 1) = \varepsilon_{0,|\Omega|}^2$, hence that $\varepsilon^2 = \frac{1}{2c_1} \varepsilon_{0,|\Omega|}^2$ and $\lambda(s_0 - 1) \leq \frac{1}{2} \varepsilon_{0,|\Omega|}^2$. The result then follows from Theorem 1 of [4]. This yields (PP-19), thus the rate of convergence in P when letting $|\Omega|, M, U \to \infty$. For the corresponding risk result, note that $h(\cdot, \cdot) \leq 1$, and

$$Eh^q(\hat{\mathbf{\Theta}}^{L_0}, \mathbf{\Theta}_0) \;=\; \varepsilon_{0,|\Omega|}^q + \int_{\varepsilon_{0,|\Omega|}}^1 P(h(\hat{\mathbf{\Theta}}^{L_0}, \mathbf{\Theta}_0) \geq x^{1/q}) dx \leq \varepsilon_{0,|\Omega|}^q + \exp\big(-c_1 |\Omega| \varepsilon_{0,|\Omega|}^{2/q}\big).$$

The desired result then follows. This completes the proof.

Proof of Lemma PP-4: Using the SVD of $\mathbf{\Theta}_0$, we obtain $\mathbf{\Theta}_0 = \bar{\mathbf{A}} \bar{\mathbf{B}}^T$, where $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$ are $U \times r_0$ and $M \times r_0$. Assume, without loss of generality, that $\bar{\mathbf{A}}^T = (\bar{\mathbf{A}}_1, \bar{\mathbf{A}}_2)$ where $\bar{\mathbf{A}}_1$ is a $r_0 \times r_0$ nonsingular matrix. Now define $\mathbf{A}_0^T = (\mathbf{I}_{r_0 \times r_0}, \bar{\mathbf{A}}_1^{-1} \bar{\mathbf{A}}_2)$ and $\mathbf{B}_0^T = \bar{\mathbf{A}}_1^T \bar{\mathbf{B}}^T$. By construction, $\mathbf{A}_0 \mathbf{B}_0^T = \bar{\mathbf{A}}_1 \bar{\mathbf{B}}_1^T$ with $\|\mathbf{A}_0\|_0 + \|\mathbf{B}_0\|_0 = (M + U - r_0 + 1) r_0$. This completes the proof.

Proof of Theorem PP-2: The proof is essentially the same except the entropy calculations.

Let $\mathcal{F}_1(s) = \{(\mathbf{A}, \mathbf{B}) \in \mathcal{F} : \mathbf{\Theta} = \mathbf{A}\mathbf{B}^T, \|\mathbf{A}\|_1 + \|\mathbf{B}\|_1 \leq s\}$ for the corresponding constrained space for the L_1-method. Similarly, define $\mathcal{F}_2(s)$ as $\{(\mathbf{A}, \mathbf{B}) \in \mathcal{F} : \mathbf{\Theta} = \mathbf{A}\mathbf{B}^T, \|\mathbf{A}\|_2^2 + \|\mathbf{B}\|_2^2 \leq s\}$ for the corresponding constrained space for the L_2-method.

For the L_1-method, we bound the bracketing Hellinger entropy $H(u, \mathcal{F}_1(s))$. To this end, let $\tilde{B}_\delta(\mathbf{A}, \mathbf{B}) = \{(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}) \in \mathcal{F}_1(s) : \sqrt{\|\tilde{\mathbf{A}} - \mathbf{A}\|_{F^*}^2 + \|\tilde{\mathbf{B}} - \mathbf{B}\|_{F^*}^2} \leq \delta\}$. Note that $0 \leq \delta_{ji} \leq 1$. By Assumption A, the triangular inequality and boundedness of $\|\mathbf{A}_j\|_\infty$ and $\|\mathbf{B}_j\|_\infty$, For any $\tilde{\mathbf{\Theta}} = (\tilde{\theta}_{ji}) = \tilde{\mathbf{A}} \tilde{\mathbf{B}}^T$ with $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}) \in \mathcal{F}_0(s)$,

$$(MU)^{-1} \sum_{j=1}^M \sum_{i=1}^U \int \sup_{\{(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}) \in \tilde{B}_\delta(\mathbf{A}, \mathbf{B})\}} (f^{1/2}(r_{ji}, z_{ji}, \tilde{\theta}_{ji}) - f^{1/2}(r_{ji}, z_{ji}, \theta_{ji}))^2 d\mu(z_{ji})$$

$$\leq \;\; \frac{d_0^2}{MU} \sup_{\{(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}) \in \tilde{B}_\delta(\mathbf{A}, \mathbf{B})\}} 2(\|\tilde{\mathbf{A}} - \mathbf{A}\|_{F^*}^2 \|\tilde{\mathbf{B}}\|_{F^*}^2 + \|\tilde{\mathbf{B}} - \mathbf{B}\|_{F^*}^2 \|\mathbf{A}\|_{F^*}^2)$$

$$\leq \;\; \frac{d_0^2}{MU} \sup_{\{(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}) \in \tilde{B}_\delta(\mathbf{A}, \mathbf{B})\}} 2(M + U) K L^2 (\|\tilde{\mathbf{A}} - \mathbf{A}\|_{F^*}^2 + \|\tilde{\mathbf{B}} - \mathbf{B}\|_{F^*}^2)$$

Similarly, it follows from Lemma 3 of [2] with $q = 1$ that

$$H(u, \mathcal{F}_1(s)) \;\; \leq \;\; cs^2 \Big(\min \big(\frac{L\sqrt{2} d_0 (K(M+U))^{1/2}}{u(MU)^{1/2}}, 1\big)\Big)^2 \log((M + U)K),$$

for some constant $c > 0$.

For the L_2-method, $H(u, \mathcal{F}_2(s))$ can be bounded similarly. Note that

$$H(u, \mathcal{F}_2(s)) \;\; \leq \;\; c \log\Big(\min(\frac{\sqrt{s} L \sqrt{2} d_0 (K(M+U))^{1/2}}{u(MU)^{1/2}}, 1)\Big)(M + U)K,$$

for some constant $c > 0$. The rest of the proof proceeds as that of Theorem PP-1. This completes the proof.

Proof of Lemma PP-5: Note that $s_1 \geq \tilde{s}_0 c_{\min}$, where $\tilde{s}_0$ is the number of nonzero element for the best L_1-factorization. The result then follows from the fact that $\tilde{s}_0 \geq s_0$ by definition. Moreover, minimization of $\|\boldsymbol{A}\|_1 + \|\boldsymbol{B}\|_1$ subject to $\boldsymbol{\Theta}_0 = \boldsymbol{A}\boldsymbol{B}$ implies that the elements of its minimizer $\boldsymbol{A}$ and $\boldsymbol{B}$ are bounded away from zero provided that those of $\boldsymbol{\Theta}_0$. Similarly, $s_2 \geq s_1 c_{\min}$ can be proved. This completes the proof.

References

[1] Stanica, P., and Montgomery, A.P. (2001). Good lower and upper bounds on binomial coefficients. *J. Ineq. in Pure. Appl. Math.*, **2**, art 30.

[2] Raskutii, G., and Wainwright, M. J. (2009). Minimax rates of estimation for high-dimensional linear regression over l_q-balls. Tech report, UC-Berkeley.

[3] Chen, B., He, S., Li, Z., and Zhang, S. (2012). Maximum block improvement and polynomial optimization. *SIAM J. on Optimi.*, 22, 87-107.

[4] Shen, X. (1998). On the method of penalization. *Statistica Sinica*, **8**, 337-357.

[5] Ossiander, M. (1987). A central limit theorem under metric entropy with L_2 bracketing. *Ann. Proba.*, **15**, 897-919.

[6] Wong, W. H., and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.*, **23**, 339-362.