Biometrika (2020), **XX**, X, *pp*. 1–24
Printed in Great Britain Advance Access publication on

Supplement to "Determining the Number of Factors in High-dimensional Generalized Latent Factor Models"

BY Y. CHEN

Department of Statistics, London School of Economics and Political Science, Houghton Street, London, WC2A 2AE, U.K.
y.chen186@lse.ac.uk

AND X. LI

School of Statistics, University of Minnesota
224 Church Street SE, Minneapolis, Minnesota, 55455, U.S.A.
lixx1766@umn.edu

A. PROOF OF THEORETICAL RESULTS

A.1. *Proof of Theorem* 1 *and Theorem* 2

We will present the proof of Theorem 2 first and then that of Theorem 1, because the former is more general than the latter. The proof of Theorem 2 is based on the following two lemmas, whose proof will be provided later in the supplementary material.

Let $G_i = (1, F_i^T)^T$ and $B_j = (d_j, A_j^T)^T$, then $m_{ij} = G_i^T B_j$. Define $\mathcal{M}_r = \{M = (m_{ij})_{1\le i\le N, 1\le j\le J} : m_{ij} = G_i^T B_j : G_i, B_j \in \mathbb{R}^r, \|G_i\| \le C, \|B_j\| \le C \text{ for all } 1 \le i \le N, 1 \le j \le N\}$, then $M^* \in \mathcal{M}_{K^*+1}$. Let $r^* = K^* + 1$ under Assumption 2. Also, let $l(M, Y, \Omega)$ denote the log-likelihood function where $\Omega = (\omega_{ij})_{1\le i\le N, 1\le j\le J}$.

LEMMA 1. *For all* $M \in \mathcal{M}_r$,

$$\begin{aligned}&\phi\{l(M, Y, \Omega) - l(M^*, Y, \Omega)\}\\ \le&(r + r^*)^{1/2}\big\{\|Z \circ \Omega\|_2 + 2\delta_{C^2}C^2\|Q\|_2\big\}\|M - M^*\|_F - \delta_{C^2}p_{\min}\|M - M^*\|_F^2\end{aligned} \tag{S.A.1}$$

where $Z = (z_{ij})_{1\le i\le N, 1\le j\le J}$, $z_{ij} = y_{ij} - b'(m^*_{ij})$, *and '*$\circ$*' denotes the matrix Hadamard product.*

LEMMA 2. *There is a universal constant* $c > 0$ *such that*

$$\Pr\Big(\|\Omega - P\|_2 \ge 4(\max_i n^*_{i\cdot})^{1/2} \vee (\max_j n^*_{\cdot j})^{1/2} + c\log^{1/2}(N + J)\Big) \le (N + J)^{-1}. \tag{S.A.2}$$

LEMMA 3. *Let* $V = (v_{ij})_{1\le i\le N, 1\le j\le J}$ *be a random matrix with independent and centered entries. In addition, assume* v_{ij}*s are sub-exponential random variables with parameters* $\nu, \alpha > 0$*. That is,* $E(e^{\lambda v_{ij}}) \le e^{\lambda^2\nu^2/2}$ *for all* $|\lambda| < 1/\alpha$*. Then, there exists a universal constant* $c > 0$ *such that with probability at least* $1 - (N + J)^{-1} - (n^*)^{-1}$,

$$\|V \circ \Omega\|_2 \le 4\max_{ij}\{E(z_{ij}^2)\}^{1/2}(\max_i n^*_{i\cdot})^{1/2} \vee (\max_j n^*_{\cdot j})^{1/2} + c(\alpha \vee \nu)\log n^* \log^{1/2}(N + J) \tag{S.A.3}$$

for all $N \ge 1, J \ge 1$*, and* $n^* \ge 6$*. In particular, under Assumptions 1 and 2,* $z_{ij} = y_{ij} - b'(m^*_{ij})$ *is sub-exponential with parameters* $\nu^2 = \phi\kappa_{2C^2} = \phi\sup_{|x|\le 2C^2} b''(x)$ *and* $\alpha = \phi/C^2$*, and there is a universal constant* $c > 0$ *such that with probability at least* $1 - (N + J)^{-1} - (n^*)^{-1}$,

$$\|Z \circ \Omega\|_2 \le 4(\phi\kappa_{2C^2})^{1/2}(\max_i n^*_{i\cdot})^{1/2} \vee (\max_j n^*_{\cdot j})^{1/2} + c\{(\phi/C^2) \vee (\phi\kappa_{2C^2})^{1/2}\}\log n^* \log^{1/2}(N + J) \tag{S.A.4}$$

for all $N \geq 1, J \geq 1$, *and* $n^* \geq 6$.

Remark 1. The constant 4 in the first term of the right-hand side of (S.A.3) can be improved to $2\sqrt{2}+\epsilon$ for any $\epsilon > 0$ with the constant c replaced by an ϵ-dependent constant c_ϵ. The logarithm term can be improved if Z is further assumed sub-Gaussian or bounded. We keep the current form which is sharp enough for our problem.

Proof of Theorem 2. By the definition of $\hat{M}^{(K)}$ and $K \geq K^*$, we have $\phi\{l(\hat{M}^{(K)}, Y, \Omega) - l(M^*, Y, \Omega)\} \geq 0$. Apply Lemma 1 with $M = \hat{M}^{(K)}$, $r = K + 1$ and combine it with $\phi\{l(\hat{M}^{(K)}, Y, \Omega) - l(M^*, Y, \Omega)\} \geq 0$. We obtain that for every $K \geq K^*$,

$$\|\hat{M}^{(K)} - M^*\|_F \leq p_{\min}^{-1}(K + K^* + 2)^{1/2}\big\{\delta_{C^2}^{-1}\|Z \circ \Omega\|_2 + 2C^2\|Q\|_2\big\}. \tag{S.A.5}$$

Thus,

$$\max_{K^* \leq K \leq K_{\max}} \big(\|\hat{M}^{(K)} - M^*\|_F\big) \leq 2p_{\min}^{-1}K_{\max}^{1/2}\big\{\delta_{C^2}^{-1}\|Z \circ \Omega\|_2 + 2C^2\|Q\|_2\big\}, \tag{S.A.6}$$

where we used the fact that $K + K^* + 2 \leq 2(K_{\max} + 1) \leq 4K_{\max}$ for $K_{\max} \geq 1$. Apply Lemma 2 and Lemma 3 to obtain an upper bound of the right-hand side of the above inequality and simplify it. We arrive at

$$\begin{aligned}
&\max_{K^* \leq K \leq K_{\max}} \big(\|\hat{M}^{(K)} - M^*\|_F\big) \\
\leq& 2p_{\min}^{-1}(K_{\max})^{1/2}\big[\{4\delta_{C^2}^{-1}(\phi\kappa_{2C^2})^{1/2} + 8C^2\}(\max_i n_{i\cdot}^*)^{1/2} \vee (\max_j n_{\cdot j}^*)^{1/2} \\
&\qquad + c\{(\phi/C^2) \vee (\phi\kappa_{2C^2})^{1/2}\log n^* + 2C^2\}\log^{1/2}(N+J)\big] \\
=& p_{\min}^{-1}(K_{\max})^{1/2}\{\kappa_{1,b,C,\phi}(\max_i n_{i\cdot}^*)^{1/2} \vee (\max_j n_{\cdot j}^*)^{1/2} + 2c(\kappa_{2,b,C,\phi}\log n^* + 2C^2)\log^{1/2}(N+J)\}
\end{aligned} \tag{S.A.7}$$

where we recall that $\kappa_{1,b,C,\phi} = 8\delta_{C^2}^{-1}(\phi\kappa_{2C^2})^{1/2} + 16C^2$ and $\kappa_{2,bC,\phi} = (\phi/C^2) \vee (\phi\kappa_{2C^2})^{1/2}$. This completes our proof. □

Proof of Theorem 1. Note that $\max_i n_{i\cdot}^* \leq p_{\max}J$ and $\max_j n_{\cdot j}^* \leq p_{\max}N$. Thus, (7) is simplified to

$$\max_{K^* \leq K \leq K_{\max}} \big(\|\hat{M}^{(K)} - M^*\|_F\big) \leq \kappa(K_{\max})^{1/2}\{p_{\min}^{-1/2}(N \vee J)^{1/2} + p_{\min}^{-1}\log n^* \log^{1/2}(N+J)\} \tag{S.A.8}$$

for some κ depending on C, b, ϕ and $p_{\max}/p_{\min}$. Because $p_{\min} = (p_{\min}/p_{\max})p_{\max} \geq (p_{\min}/p_{\max})n^*/(NJ)$, the above inequality implies

$$\begin{aligned}
&\max_{K^* \leq K \leq K_{\max}} \big(\|\hat{M}^{(K)} - M^*\|_F\big) \\
\leq& \kappa(K_{\max})^{1/2}\{(n^*/(NJ))^{-1/2}(N \vee J)^{1/2} + (n^*/(NJ))^{-1}\log(n^*)\log^{1/2}(N+J)\}
\end{aligned} \tag{S.A.9}$$

with a possibly different κ that also depends on C, b, and ϕ. Multiplying both sides by $(NJ)^{-1/2}$ and simplifying it, we arrive at

$$\begin{aligned}
&\max_{K^* \leq K \leq K_{\max}} \big\{(NJ)^{-1/2}\|\hat{M}^{(K)} - M^*\|_F\big\} \\
\leq& \kappa K_{\max}^{1/2}\Big[\big\{(N \vee J)/n^*\big\}^{1/2} + \{(NJ)^{1/2}\log^{1/2}(N+J)\}(n^*)^{-1}\log n^*\Big].
\end{aligned} \tag{S.A.10}$$

Note that for $n^*/(\log n^*)^2 \geq (N \wedge J)\log(N+J)$, $\big\{(N \vee J)/n^*\big\}^{1/2} \geq \{(NJ)^{1/2}\log^{1/2}(N+J)\}(n^*)^{-1}\log n^*$, and the above inequality is simplified as

$$\max_{K^* \leq K \leq K_{\max}} \big\{(NJ)^{-1/2}\|\hat{M}^{(K)} - M^*\|_F\big\} \leq 2\kappa\big\{K_{\max}(N \vee J)/n^*\big\}^{1/2}. \tag{S.A.11}$$

This completes the proof. □

A.2. *Proof of Theorem* 3*, Theorem* 4*, and Corollary* 2

The proofs of Theorem 3 and Theorem 4 are based on the following three supporting lemmas, whose proofs are given in the supplementary material. We start by recalling $u(n,N,J,K) = v(n,N,J,K) - v(n,N,J,K-1)$ and defining $R = 4(p_{\min}\delta_{C^2})^{-1}\left\{\|Z\circ\Omega\|_2 + 2\delta_{C^2}C^2\|Q\|_2\right\}^2$.

LEMMA 4. *If* $u(\cdot)$ *satisfies*

$$\lim_{N,J\to\infty}\Pr\Big(u(n,N,J,K^*+1) > 2\phi^{-1}(K^*+1)R\Big) = 1 \tag{S.A.12}$$

and

$$\lim_{N,J\to\infty}\Pr\Big(\inf_{K^*+2\le K\le K_{\max}} u(n,N,J,K) > 2\phi^{-1}R\Big) = 1, \tag{S.A.13}$$

then

$$\lim_{N,J\to\infty}\Pr(\hat{K} > K^*) = 0, \tag{S.A.14}$$

for $K_{\max}\ge K^*\ge 1$.

LEMMA 5. *If*

$$\lim_{N,J\to\infty}\Pr\Big(4(\delta_{C^2}p_{\min})^{-1}K^*R \le \sigma^2_{K^*+1}(M^*)\Big) = 1, \tag{S.A.15}$$

and $u(\cdot)$ *satisfies*

$$\lim_{N,J\to\infty}\Pr\Big(u(n,N,J,K) < \phi^{-1}\delta_{C^2}p_{\min}\sigma^2_{K+1}(M^*) \text{ for all } 1\le K\le K^*\Big) = 1 \tag{S.A.16}$$

then $\lim_{N,J\to\infty}\Pr(\hat{K} < K^*) = 0$ *for* $K^*\ge 1$.

LEMMA 6. *Under the asymptotic regime* (10)*,* $R = O_p(p_{\max}/p_{\min}(N\vee J))$.

In the rest of the section, we provide the proof of Theorem 4 first and then the proof of Theorem 3 because the former is more general than the latter.

Proof of Theorem 4. We will verify that conditions of Theorem 4 ensure conditions in Lemma 4 and Lemma 5. We start with verifying conditions in Lemma 4. According to the second line of (11),

$$\begin{aligned}
&\lim_{N,J\to\infty}\Pr\Big(u(n,N,J,K^*+1) > 2\phi^{-1}(K^*+1)R\Big)\\
\ge&\liminf_{N,J\to\infty}\Pr\Big(\xi_{N,J}(K^*+1)(p_{\max}/p_{\min})(N\vee J) > 2\phi^{-1}(K^*+1)R\Big)\\
\ge&\liminf_{N,J\to\infty}\Pr\Big(\xi_{N,J}(p_{\max}/p_{\min})(N\vee J) > 2\phi^{-1}R\Big)\\
=&1,
\end{aligned} \tag{S.A.17}$$

where the last line is obtained according to Lemma 6 and that $\xi_{N,J}\to\infty$ in probability. Similarly,

$$\begin{aligned}
\lim_{N,J\to\infty}\Pr\Big(\inf_{K^*+2\le K\le K_{\max}} u(n,N,J,K) > 2\phi^{-1}R\Big) &\ge \liminf_{N,J\to\infty}\Pr\Big(\xi_{N,J}\Big(\frac{p_{\max}}{p_{\min}}\Big)(N\vee J) > 2\phi^{-1}R\Big)\\
&=1.
\end{aligned} \tag{S.A.18}$$

Thus, conditions of Lemma 4 are verified and we obtain

$$\lim_{N,J\to\infty}\Pr(\hat{K} > K^*) = 0. \tag{S.A.19}$$

Next, we verify conditions of Lemma 5. According to Lemma 6 and the assumption $p_{\min}^{-2}p_{\max}K^*(N\vee J)=o(\sigma^2_{K^*+1}(M^*))$, we have

$$4(\delta_{C^2}p_{\min})^{-1}K^*R=O_p\big(p_{\min}^{-2}p_{\max}K^*(N\vee J)\big)=o_p\big(\sigma^2_{K^*}(M^*)\big). \tag{S.A.20}$$

Thus,

$$\lim_{N,J\to\infty}\Pr\Big(4(\delta_{C^2}p_{\min})^{-1}K^*R\le\sigma^2_{K^*}(M^*)\Big)=1. \tag{S.A.21}$$

In addition, according to the first line of (11),

$$\begin{aligned}&\lim_{N,J\to\infty}\Pr\Big(u(n,N,J,K)<\phi^{-1}\delta_{C^2}p_{\min}\sigma^2_{K+1}(M^*)\text{ for }1\le K\le K^*\Big)\\ \ge&\liminf_{N,J\to\infty}\Pr\Big(\xi_{N,J}^{-1}p_{\min}\sigma^2_{K^*+1}(M^*)<\phi^{-1}\delta_{C^2}p_{\min}\sigma^2_{K+1}(M^*)\text{ for all }K\\ \ge&\liminf_{N,J\to\infty}\Pr\Big(\xi_{N,J}^{-1}<\phi^{-1}\delta_{C^2}\Big)\\ =&1.\end{aligned} \tag{S.A.22}$$

From (S.A.21) and (S.A.22), conditions of Lemma 5 are verified and thus

$$\lim_{N,J\to\infty}\Pr(\hat{K}<K^*)=0. \tag{S.A.23}$$

We complete the proof by combining (S.A.19) and (S.A.23). □

Proof of Theorem 3. First note that the existence of u satisfying (9) implies $N\vee J=o(\sigma^2_{K^*+1}(M^*))$, which further implies $p_{\min}^{-2}p_{\max}K^*(N\vee J)=o(\sigma^2_{K^*+1}(M^*))$ under the asymptotic regime $p_{\min}^{-1}=O(1)$, $K^*=O(1)$. Thus, the assumption about the singular value of M^* in Theorem 4 is verified. Also, $p_{\min}^{-1}=O(1)$ implies that $(N\wedge J)\log(N+J)=o(n^*/(\log n^*)^2)$. Thus, (10) is verified.

We proceed to verify that u satisfies (11) in Theorem 4. We note that $p_{\min}^{-1}=O(1)$, $K^*=O(1)$ and u satisfies (9) implies that there exists $\xi_{N,J}\to\infty$ satisfying

$$u(n,N,J,K)\le\xi_{N,J}^{-1}p_{\min}\sigma^2_{K^*+1}(M^*)\text{ for all }K, \tag{S.A.24}$$

$$u(n,N,J,K)\ge\xi_{N,J}(p_{\max}/p_{\min})(N\vee J)\text{ for all }K, \tag{S.A.25}$$

and

$$u(n,N,J,K^*+1)\ge\xi_{N,J}(K^*+1)(p_{\max}/p_{\min})(N\vee J). \tag{S.A.26}$$

Note that $\sigma^2_{K+1}(M^*)\ge\sigma^2_{K^*+1}(M^*)$ for $K\le K^*$. Thus, (S.A.24) implies the first line of (11); (S.A.26) implies the second line of (11); (S.A.25) implies the last line of (11). It verifies (11) and completes the proof. □

Proof of Corollary 2. Under the asymptotic regime (8) and $N\vee J=o(\sigma^2_{K^*+1}(M^*))$, (10) and $p_{\min}^{-2}p_{\max}K^*(N\vee J)=o(\sigma^2_{K^*+1}(M^*))$ are verified in the proof of Theorem 3. We now verify (11).

From the conditions on $h(n,N,J)$, there exists a sequence $\xi_{N,J}$ (possibly depending on $h(n,N,J)$) such that $\xi_{N,J}\to\infty$ in probability and

$$\xi_{N,J}<h(n,N,J)(p_{\min}/p_{\max})(K^*+1)^{-1}\text{ and }\xi_{N,J}\le(h(n,N,J))^{-1}(N\vee J)^{-1}p_{\min}\sigma^2_{K^*+1}(M^*). \tag{S.A.27}$$

Also, note that $u(n,N,J,K)=v(n,N,J,K)-v(n,N,J,K-1)=(N\vee J)h(n,N,J)$. It is not hard to verify (S.A.27) implies (11), and, thus Theorem 4 applies.

We proceed to the proof of the 'in particular' part. Note that by definition $E(n)=n^*$ and $Var(n)=\sum_i\sum_j p_{ij}(1-p_{ij})\le\sum_i\sum_j p_{ij}=n^*$, which implies $\lim_{N,J\to\infty}\Pr(n>2n^*\text{ or }n<n^*/2)=0$ and further implies

$$\lim_{N,J\to\infty}\Pr\big(n/(N\vee J)\ge 2n^*/(N\vee J)\text{ or }n/(N\vee J)\le n^*/\{2(N\vee J)\}\big)=0.$$

Note that in this part, $h(n,N,J) = \log(n/(N \vee J))$. Also, $\log(n^*/\{2(N \vee J)\}) \to \infty$. Thus, $h(n,N,J) \to \infty$ in probability. In addition, on the event $n/(N \vee J) \leq 2n^*/(N \vee J)$, $(h(n,N,J))^{-1}(N \vee J)^{-1}\sigma^2_{K^*+1}(M^*) \geq \log\big(2n^*/(N \vee J)\big)(N \vee J)^{-1}\sigma^2_{K^*+1}(M^*)$. The right-hand-side of this inequality tend to infinity under the assumptions of the Corollary. This implies $(h(n,N,J))^{-1}(N \vee J)^{-1}\sigma^2_{K^*+1}(M^*) \to \infty$ in probability. □

B. Proof of Supporting Lemmas

Proof of Lemma 1. By definition,

$$\begin{aligned}
&\phi\{l(M,Y,\Omega) - l(M^*,Y,\Omega)\} \\
=&\sum_{ij} \omega_{ij}\big\{y_{ij}m_{ij} - b(m_{ij}) - y_{ij}m^*_{ij} + b(m^*_{ij})\big\} \\
=&\sum_{i,j}(y_{ij} - b'(m^*_{ij}))(m_{ij} - m^*_{ij})\omega_{ij} - \sum_{ij}\big\{b(m_{ij}) - b(m^*_{ij}) - b'(m^*_{ij})(m_{ij} - m^*_{ij})\big\}\omega_{ij}.
\end{aligned} \tag{S.B.1}$$

In the rest of the proof, we derive upper bounds for each term on the right-hand-side of the above display. For the first term $\sum_{i,j}(y_{ij} - b'(m^*_{ij}))(m_{ij} - m^*_{ij})\omega_{ij}$, we write it as

$$\sum_{i,j}(y_{ij} - b'(m^*_{ij}))(m_{ij} - m^*_{ij})\omega_{ij} = \langle Z \circ \Omega, M - M^*\rangle, \tag{S.B.2}$$

where $\langle A,B\rangle = tr(A^TB)$ denotes the matrix inner product. Recall the following inequality in linear algebra: $|\langle A,B\rangle| \leq \|A\|_2\|B\|_* \leq \sqrt{\text{rank}(B)}\|A\|_2\|B\|_F$ for any two matrices A and B. Applying this fact to the above display, we obtain

$$\big|\sum_{i,j}(y_{ij} - b'(m^*_{ij}))(m_{ij} - m^*_{ij})\omega_{ij}\big| \leq \{\text{rank}(M - M^*)\}^{1/2}\|Z \circ \Omega\|_2\|M - M^*\|_F. \tag{S.B.3}$$

Notice that $\text{rank}(M - M^*) \leq r + r^*$ for $M \in \mathcal{M}_r$. Thus, the above inequality implies

$$\big|\sum_{i,j}(y_{ij} - b'(m^*_{ij}))(m_{ij} - m^*_{ij})\omega_{ij}\big| \leq (r + r^*)^{1/2}\|Z \circ \Omega\|_2\|M - M^*\|_F. \tag{S.B.4}$$

We proceed to the analysis of the second term $\sum_{ij}\big\{b(m_{ij}) - b(m^*_{ij}) - b'(m^*_{ij})(m_{ij} - m^*_{ij})\big\}\omega_{ij}$. Note that for $M \in \mathcal{M}_r$, $|m_{ij}| \leq \|B_i\|\|G_j\| \leq C^2$. Similarly, $|m^*_{ij}| \leq C^2$. Thus, for any $\tilde{m}_{ij} = tm^*_{ij} + (1-t)m_{ij}$ and $t \in (0,1)$, $|\tilde{m}_{ij}| \leq C^2$. Recall the definition of $\delta_{C^2} = \inf_{|x| \leq C^2} b''(x)$. Then, $\frac{1}{2}b''(\tilde{m}_{ij}) \geq \delta_{C^2}$. This implies

$$\begin{aligned}
&\sum_{ij}\big\{b(m_{ij}) - b(m^*_{ij}) - b'(m^*_{ij})(m_{ij} - m^*_{ij})\big\}\omega_{ij} \\
=&\sum_{ij}\frac{1}{2}b''(\tilde{m}_{ij})(m_{ij} - m^*_{ij})^2\omega_{ij} \\
\geq&\delta_{C^2}\sum_{ij}(m_{ij} - m^*_{ij})^2\omega_{ij}.
\end{aligned} \tag{S.B.5}$$

Note that

$$\begin{aligned}
&\sum_{ij}(m_{ij} - m^*_{ij})^2\omega_{ij} \\
=&\sum_{ij}(m_{ij} - m^*_{ij})^2(\omega_{ij} - p_{ij}) + \sum_{ij}p_{ij}(m_{ij} - m^*_{ij})^2 \\
\geq&\langle (M - M^*) \circ (M - M^*), Q\rangle + p_{\min}\|M - M^*\|_F^2 \\
\geq&-\|(M - M^*) \circ (M - M^*)\|_*\|Q\|_2 + p_{\min}\|M - M^*\|_F^2.
\end{aligned} \tag{S.B.6}$$

where we define $Q = \Omega - P$ and $P = (p_{ij})_{1\le i\le N, 1\le j\le J}$. The next lemma is helpful for bounding matrix norms involving Hadamard products, whose proof is given later this section.

LEMMA 7. *For $M \in \mathcal{M}_r$, $\|(M - M^*) \circ (M - M^*)\|_* \le 2C^2(r + r^*)^{1/2}\|M - M^*\|_F$.*

Remark 2. The proof of Lemma 7 utilizes the property that $m_{ij} = B_i^T G_j$ with $\|B_i\|, \|G_j\| \le C$ and combine it with a result in Horn (1995). This improves the estimate in Chen et al. (2020) where $|m_{ij}| \le C^2$ is directly used to derive an upper bound $2C^2(r + r^*)\|M - M^*\|_F$. Comparing with this bound, the above lemma provide a sharper bound in the order of $r + r^*$.

Applying Lemma 7 to (S.B.6) and combine it with (S.B.5), we obtain

$$\begin{aligned}&\sum_{ij}\big\{b(m_{ij}) - b(m_{ij}^*) - b'(m_{ij}^*)(m_{ij} - m_{ij}^*)\big\}\omega_{ij}\\ \ge&\delta_{C^2}\big\{p_{\min}\|M - M^*\|_F^2 - 2C^2(r + r^*)^{1/2}\|M - M^*\|_F\|Q\|_2\big\}.\end{aligned} \tag{S.B.7}$$

We complete the proof by combining the above display with (S.B.1) and (S.B.4). □

Proof of Lemma 7. Let $\tilde{B}_i = (B_i^T, -(B_i^*)^T)^T$ and $\tilde{G}_j = (G_j^T, -(G_j^*)^T)^T$. Then, $\tilde{B}_i, \tilde{G}_j \in \mathbb{R}^{r+r^*}$, $\|\tilde{B}_i\|, \|\tilde{G}_j\| \le \sqrt{2}C$, and $m_{ij} - m_{ij}^* = \tilde{B}_i^T\tilde{G}_j$ for all i, j.

On the other hand, Theorem 2 in Horn (1995) states that, for any $m \times n$ matrices $A = (a_{ij}), B = (b_{ij})$, if $a_{ij} = g_j^T f_i$ for vectors g_j and f_is. Then,

$$\sum_{i=1}^k \sigma_i(A \circ B) \le \sum_{i=1}^k \|f_{[i]}\|\|g_{[i]}\|\sigma_i(B) \text{ for } k = 1, \cdots, m \wedge n, \tag{S.B.8}$$

where $\sigma_i(\cdot)$ denotes the ith largest singular value of a matrix, $\|f_{[1]}\| \ge \|f_{[2]}\| \ge \cdots \ge \|f_{[m]}\|$ and $\|g_{[1]}\| \ge \cdots \ge \|g_{[n]}\|$ denote the order statistics of $\{\|f_i\|\}_{i=1}^m$ and $\{\|g\|_j\}_{j=1}^n$. Now, we let $k = N \wedge J$, $A = M - M^*$, $B = A$, $f_i = \tilde{B}_i$, $g_j = \tilde{G}_j$ in the above result and note that $\|f_{[i]}\|, \|g_{[j]}\| \le \sqrt{2}C$ in this case, we obtain

$$\sum_{i=1}^{N\wedge J} \sigma_i((M - M^*) \circ (M - M^*)) \le \sum_{i=1}^{N\wedge J} 2C^2\sigma_i(M - M^*) = 2C^2\|M - M^*\|_*. \tag{S.B.9}$$

Noting the left-hand side of the above display equals $\|(M - M^*) \circ (M - M^*)\|_*$. Thus,

$$\|(M - M^*) \circ (M - M^*)\|_* \le 2C^2\|M - M^*\|_* \le 2C^2(r + r^*)^{1/2}\|M - M^*\|_F. \tag{S.B.10}$$

The proofs of Lemmas 2 and 3 are based on the next lemma that provides an upper tail bound for the spectral norm of a large class of random matrices. Its proof mainly combines standard symmetrization and truncation arguments with a recent result by Bandeira & Van Handel (2016) on the spectral norm of symmetric random matrices with independent, centered and symmetric entries.

LEMMA 8. *Let $X = (x_{ij})_{1\le i\le N, 1\le j\le J}$ be an $N \times J$ matrix with $E(x_{ij}) = 0$ and $E(x_{ij}^2) < \infty$. Then, there is a universal constant $c > 0$ such that for all $t, \lambda \ge 0$*

$$\Pr\Big(\|X\|_2 \ge 4(\sigma_1 \vee \sigma_2) + t\Big) \le (N + J)e^{-t^2/(c\lambda^2)} + \sum_{i=1}^N\sum_{j=1}^J \Pr\big(|x_{ij} - x'_{ij}| > \lambda\big), \tag{S.B.11}$$

where we define $\sigma_1 = \max_{1\le i\le N}\{\sum_{j=1}^J E(x_{ij}^2)\}^{1/2}$, $\sigma_2 = \max_{1\le j\le J}\{\sum_{i=1}^N E(x_{ij}^2)\}^{1/2}$, and x'_{ij} is an independent copy of x_{ij}.

Proof of Lemma 8. Let $X' = (x'_{ij})$ which is an independent copy of X and let $\tilde{X} = (\tilde{x}_{ij}) = X - X'$. Then, $\tilde{x}_{ij}$s have symmetric distribution and are independent. Let $Z = (z_{ij}) = \begin{pmatrix} 0 & \tilde{X} \\ \tilde{X}^T & 0 \end{pmatrix}$. Z is a symmetric $(N + J) \times (N + J)$ random matrix whose entries are independent and symmetric random variables.

Define a random matrix $Z(\lambda)$ as the truncated Z,

$$Z(\lambda) = (z_{ij}(\lambda))_{1\le i\le N, 1\le j\le J} = (z_{ij} I(|z_{ij}| \le \lambda))_{1\le i\le N, 1\le j\le J}. \tag{S.B.12}$$

Then, entries of $Z(\lambda)$ are independent, symmetric random variables and are bounded by λ. Apply Corollary 3.12 in Bandeira & Van Handel (2016) to $Z(\lambda)$, then there exists a universal constant $c > 0$ such that

$$\Pr\Big(\|Z(\lambda)\|_2 \ge 2^{3/2} \max_{1\le i\le (N+J)} \big[\sum_{j=1}^{N+J} E\{z_{ij}^2(\lambda)\}\big]^{1/2} + t\Big) \le (N+J)e^{-t^2/(c\lambda^2)} \tag{S.B.13}$$

Note that

$$\begin{aligned}\max_{1\le i\le (N+J)} \big[\sum_{j=1}^{N+J} E\{z_{ij}^2(\lambda)\}\big]^{1/2} &\le \max_{1\le i\le (N+J)} \Big\{\sum_{j=1}^{N+J} E(z_{ij}^2)\Big\}^{1/2} \\ &= \max\big[\max_{1\le i\le N}\{\sum_{j=1}^{J} E(\tilde{x}_{ij}^2)\}^{1/2}, \max_{1\le j\le J}\{\sum_{i=1}^{N} E(\tilde{x}_{ij}^2)\}^{1/2}\big] \\ &= \sqrt{2}(\sigma_1 \vee \sigma_2).\end{aligned} \tag{S.B.14}$$

Thus,

$$\Pr\Big(\|Z(\lambda)\|_2 \ge 4(\sigma_1 \vee \sigma_2) + t\Big) \le (N+J)e^{-t^2/(c\lambda^2)}. \tag{S.B.15}$$

On the other hand,

$$\Pr\big(\|Z\|_2 \ge 4(\sigma_1 \vee \sigma_2) + t\big) \le \Pr\big(\|Z(\lambda)\|_2 \ge 4(\sigma_1 \vee \sigma_2) + t\big) + \Pr\big(\max_{1\le i,j\le N+J} |z_{ij}| > \lambda\big). \tag{S.B.16}$$

The above two inequalities together imply

$$\Pr\big(\|Z\|_2 \ge 4(\sigma_1 \vee \sigma_2) + t\big) \le (N+J)e^{-t^2/(c\lambda^2)} + \Pr(\max_{1\le i,j\le N+J} |z_{ij}| > \lambda). \tag{S.B.17}$$

Note that $\|Z\|_2 = \|\tilde{X}\|_2$ and $\max_{1\le i,j\le N+J} |z_{ij}| = \max_{1\le i\le N, 1\le j\le J} |\tilde{x}_{ij}|$. From the above inequality, we obtain

$$\Pr\big(\|\tilde{X}\|_2 \ge 4(\sigma_1 \vee \sigma_2) + t\big) \le (N+J)e^{-t^2/(c\lambda^2)} + \Pr(\max_{1\le i\le N, 1\le j\le J} |\tilde{x}_{ij}| > \lambda). \tag{S.B.18}$$

With a union bound, we further get

$$\Pr(\|\tilde{X}\|_2 \ge 4(\sigma_1 \vee \sigma_2) + t) \le (N+J)e^{-t^2/(c\lambda^2)} + \sum_{1\le i\le N, 1\le j\le J} \Pr(|\tilde{x}_{ij}| > \lambda). \tag{S.B.19}$$

Recall $\tilde{X} = X - X'$ and the function $I(\|X - X'\|_2 \ge 4(\sigma_1 \vee \sigma_2) + t)$ is convex in X'. Thus, by Jensen's inequality,

$$\Pr(\|X\|_2 \ge 4(\sigma_1 \vee \sigma_2) + t) \le \Pr(\|X - X'\|_2 \ge 4(\sigma_1 \vee \sigma_2) + t) = \Pr(\|\tilde{X}\|_2 \ge 4(\sigma_1 \vee \sigma_2) + t). \tag{S.B.20}$$

This, together with (S.B.19) completes the proof. □

Proof of Lemma 2. Let ω'_{ij} be an independent copy of ω_{ij}, then $|\omega'_{ij} - p_{ij} - (\omega_{ij} - p_{ij})| \le 1$. In addition, $E(\omega_{ij} - p_{ij})^2 = p_{ij}(1 - p_{ij}) \le p_{ij}$. Thus, $\max_i\{\sum_j E(\omega_{ij} - p_{ij})^2\}^{1/2} \le \max_i(\sum_j p_{ij})^{1/2} = (\max_i n^*_{i\cdot})^{1/2}$ and $\max_j\{\sum_i E(\omega_{ij} - p_{ij})^2\}^{1/2} \le (\max_j n^*_{\cdot j})^{1/2}$.

Choose $\lambda = 1$ and apply Lemma 8 to $\Omega - P$, we obtain that for all $t \ge 0$,

$$\Pr\big(\|\Omega - P\|_2 \ge 4(\max_i n^*_{i\cdot} \vee \max_j n^*_{\cdot j})^{1/2} + t\big) \le (N+J)e^{-t^2/c}. \tag{S.B.21}$$

Let $t = (2c\log(N+J))^{1/2}$ in the above inequality, we obtain

$$\Pr\left(\|\Omega - P\|_2 \geq 4(\max_i n_{i\cdot}^* \vee \max_j n_{\cdot j}^*)^{1/2} + (2c\log(N+J))^{1/2}\right) \leq (N+J)^{-1}. \tag{S.B.22}$$

We complete the proof by noting that $(2c)^{1/2}$ is still a universal constant. □

Proof of Lemma 3. Apply Lemma 8 to $V \circ \Omega$, we obtain that for all $t, \lambda \geq 0$,

$$\Pr\left(\|V \circ \Omega\|_2 \geq 4(\sigma_1 \vee \sigma_2) + t\right) \leq (N+J)e^{-t^2/(c\lambda^2)} + \sum_{ij} \Pr(|v_{ij}\omega_{ij} - v'_{ij}\omega'_{ij}| \geq \lambda), \tag{S.B.23}$$

where (v'_{ij}, ω'_{ij}) is an independent copy of (v_{ij}, ω_{ij}), $\sigma_1 = \max_i\{\sum_j E(v_{ij}^2\omega_{ij}^2)\}^{1/2}$ and $\sigma_2 = \max_j\{\sum_i E(v_{ij}^2\omega_{ij}^2)\}^{1/2}$. We proceed to a detailed analysis of σ_1, σ_2 and the probability $\Pr(|v_{ij}\omega_{ij} - v'_{ij}\omega'_{ij}| \geq \lambda)$. First, a direct calculation gives

$$\sigma_1 = \max_i\Big\{\sum_j p_{ij}E(v_{ij}^2)\Big\}^{1/2} \leq (\max_i n_{i\cdot}^*)^{1/2}\max_{ij}\{E(v_{ij}^2)\}^{1/2}. \tag{S.B.24}$$

Similarly, $\sigma_2 \leq (\max_j n_{\cdot j}^*)^{1/2}\max_{ij}\{E(v_{ij}^2)\}^{1/2}$. Now we find an upper bound of $\Pr(|v_{ij}\omega_{ij} - v'_{ij}\omega'_{ij}| \geq \lambda)$. Note that

$$\begin{aligned}
&\Pr(|v_{ij}\omega_{ij} - v'_{ij}\omega'_{ij}| \geq \lambda) \\
=&p_{ij}^2\Pr(|v_{ij} - v'_{ij}| \geq \lambda) + 2p_{ij}(1-p_{ij})\Pr(|v_{ij}| \geq \lambda) \\
\leq&3p_{ij}\Pr(|v_{ij} - v'_{ij}| \geq \lambda) \vee \Pr(|v_{ij}| \geq \lambda).
\end{aligned} \tag{S.B.25}$$

For $\Pr(|v_{ij}| \geq \lambda)$, we use a tail bound for sub-exponential variables

$$\Pr(|v_{ij}| \geq \lambda) \leq 2e^{-\lambda^2/(2\nu^2)} \vee e^{-\lambda/(2\alpha)}. \tag{S.B.26}$$

Similarly, noting that $v_{ij} - v'_{ij}$ is also sub-exponential with parameters $2\nu^2, \alpha$, we have

$$\Pr(|v_{ij} - v'_{ij}| \geq \lambda) \leq 2e^{-\lambda^2/(4\nu^2)} \vee e^{-\lambda/(2\alpha)}. \tag{S.B.27}$$

Combining the above two inequalities with (S.B.25), we have

$$\Pr(|v_{ij}\omega_{ij} - v'_{ij}\omega'_{ij}| \geq \lambda) \leq 6p_{ij}e^{-\lambda^2/(4\nu^2)} \vee e^{-\lambda/(2\alpha)}. \tag{S.B.28}$$

Combining the above inequality with (S.B.23), we arrive at

$$\begin{aligned}
&\Pr\left(\|V \circ \Omega\|_2 \geq 4\max_{ij}\{E(v_{ij}^2)\}^{1/2}(\max_i n_{i\cdot}^*)^{1/2} \vee (\max_j n_{\cdot j}^*)^{1/2} + t\right) \\
\leq&(N+J)e^{-t^2/(c\lambda^2)} + 6e^{-\lambda^2/(4\nu^2)} \vee e^{-\lambda/(2\alpha)}n^*.
\end{aligned} \tag{S.B.29}$$

Let $\lambda = 4(\alpha \vee \nu)\log n^*$. It is not hard to verify that $6e^{-\lambda^2/(4\nu^2)} \vee e^{-\lambda/(2\alpha)}n^* \leq (n^*)^{-1}$ for $n^* \geq 6$. Let $t = \lambda\{2c\log(N+J)\}^{1/2}$, we obtain $(N+J)e^{-t^2/(c\lambda^2)} \leq (N+J)^{-1}$. Combining the above inequalities with (S.B.29), we obtain that with probability at least $1 - (N+J)^{-1} - (n^*)^{-1}$,

$$\begin{aligned}
&\|V \circ \Omega\|_2 \\
\leq&4\max_{ij}\{E(v_{ij}^2)\}^{1/2}(\max_i n_{i\cdot}^*)^{1/2} \vee (\max_j n_{\cdot j}^*)^{1/2} + 4\sqrt{2}c^{1/2}(\alpha \vee \nu)\log n^* \log^{1/2}(N+J).
\end{aligned} \tag{S.B.30}$$

This completes the proof of inequality (S.A.3) (note that $4\sqrt{2}c^{1/2}$ is also a universal constant). We proceed to prove the 'in particular' part of the lemma. For each $z_{ij} = y_{ij} - b'(m_{ij}^*)$, its second moment is $E(z_{ij}^2) = \phi b''(m_{ij}^*) \leq \phi\kappa_{2C^2}$. In addition, its moment generating function is $E(e^{\lambda z_{ij}}) = \exp[\phi^{-1}\{b(m_{ij}^* + \lambda\phi) - b(m_{ij}^*)\} - \lambda b'(m_{ij}^*)] = \exp\{\phi b''(m_{ij}^* + \tilde{\lambda}\phi)\lambda^2/2\}$ for some $|\tilde{\lambda}| \leq |\lambda|$. Since $|m_{ij}^*| \leq C^2$ by assumption, we can see that for $|\lambda| \leq C^2/\phi$, $|m_{ij}^* + \tilde{\lambda}\phi| \leq 2C^2$ and thus $E(e^{\lambda z_{ij}}) \leq \exp\{\kappa_{2C^2}\phi\lambda^2/2\}$ for all $|\lambda| \leq C^2/\phi$. This implies that z_{ij} is sub-exponential with the parameters

$\nu^2 = \phi\kappa_{2C^2}$ and $\alpha = \phi/C^2$. We complete the proof by applying (S.A.3) with the above parameters for Z. □

Proof of Lemma 4. For each $K^* + 1 \le K \le K_{\max}$, we first derive an upper bound for $\phi\{l(\hat{M}^{(K)}, Y, \Omega) - l(\hat{M}^{(K^*)}, Y, \Omega)\} - (v(n, N, J, K) - v(n, N, J, K^*))$. According to Lemma 1,

$$\begin{aligned}
&\phi\{l(\hat{M}^{(K)}, Y, \Omega) - l(M^*, Y, \Omega)\} \\
\le&(K + K^* + 2)^{1/2}\big(\|Z \circ \Omega\|_2 + 2\delta_{C^2}C^2\|Q\|_2\big)\|\hat{M}^{(K)} - M^*\|_F \qquad \text{(S.B.31)} \\
\le&2K^{1/2}\big(\|Z \circ \Omega\|_2 + 2\delta_{C^2}C^2\|Q\|_2\big)\|\hat{M}^{(K)} - M^*\|_F.
\end{aligned}$$

Combining this with (S.A.5) gives

$$\phi\{l(\hat{M}^{(K)}, Y, \Omega) - l(M^*, Y, \Omega)\} \le 4p_{\min}^{-1}K\big\{\|Z \circ \Omega\|_2 + 2\delta_{C^2}C^2\|Q\|_2\big\}^2 = KR. \qquad \text{(S.B.32)}$$

Thus, the penalized log-likelihood satisfies

$$\begin{aligned}
&\max_{K^*+1\le K\le K_{\max}}\Big[-2l(\hat{M}^{(K)}, Y, \Omega) + v(n, N, J, K) - \{-2l(M^*, Y, \Omega) + v(n, N, J, K^*)\}\Big] \\
\ge&\max_{K^*+1\le K\le K_{\max}}\Big[-2\phi^{-1}KR + \sum_{l=K^*+1}^{K} u(n, N, J, K)\Big]
\end{aligned} \qquad \text{(S.B.33)}$$

It is easy to see that, if the events $u(n, N, J, K^* + 1) > 2\phi^{-1}(K^* + 1)R$ and $u(n, N, J, l) > 2\phi^{-1}R$ happen at the same time for all $K^* + 2 \le l \le K_{\max}$, then the right-hand side of the above inequality is strictly greater than zero. Thus,

$$\begin{aligned}
&\Pr\Big(\hat{K} \le K^*\Big) \\
\ge&\Pr\Big(\max_{K^*+1\le K\le K_{\max}}\Big[-2l(\hat{M}^{(K)}, Y, \Omega) + v(n, N, J, K) - \{-2l(M^*, Y, \Omega) + v(n, N, J, K^*)\}\Big] > 0\Big) \\
\ge&\Pr\Big(u(n, N, J, K^* + 1) > 2\phi^{-1}(K^* + 1)R, \text{ and } \inf_{K^*+2\le l\le K_{\max}} u(n, N, J, K) > 2\phi^{-1}R\Big).
\end{aligned} \qquad \text{(S.B.34)}$$

We complete the proof by noting the right-hand side of the above inequality tend to one under the assumptions of the lemma. □

The proof of Lemma 5 requires the next lemma.

LEMMA 9. *If* $4(\delta_{C^2}p_{\min})^{-1}K^*R \le \sigma^2_{K^*+1}(M^*)$, *then*

$$\phi\{l(\hat{M}^{(K)}, Y, \Omega) - l(\hat{M}^{(K^*)}, Y, \Omega)\} \le -\frac{1}{2}\delta_{C^2}p_{\min}\Big\{\sum_{l=K+2}^{K^*+1}\sigma_l^2(M^*)\Big\} \qquad \text{(S.B.35)}$$

for $0 \le K \le K^* - 1$.

Proof of Lemma 9. First, according to Lemma 1, $\hat{M}^{(K)} \in \mathcal{M}_{K+1}$ and $K + 1 \le K^*$, we have

$$\begin{aligned}
&\phi\{l(\hat{M}^{(K)}, Y, \Omega) - l(M^*, Y, \Omega)\} \\
\le&(K + K^* + 2)^{1/2}\big\{\|Z \circ \Omega\|_2 + 2\delta_{C^2}C^2\|Q\|_2\big\}\|\hat{M}^{(K)} - M^*\|_F - \delta_{C^2}p_{\min}\|\hat{M}^{(K)} - M^*\|_F^2 \\
\le&\sup_{M\in\mathcal{M}_{K+1}}\Big[2(K^*)^{1/2}\big\{\|Z \circ \Omega\|_2 + 2\delta_{C^2}C^2\|Q\|_2\big\}\|M - M^*\|_F - \delta_{C^2}p_{\min}\|M - M^*\|_F^2\Big].
\end{aligned} \qquad \text{(S.B.36)}$$

Note that the expression inside 'sup' is a quadratic function in $\|M - M^*\|_F$. Let $d(M^*, \mathcal{M}_{K+1}) = \inf_{M\in\mathcal{M}_{K+1}} \|M - M^*\|_F$. From properties of a quadratic function, if $d(M^*, \mathcal{M}_{K+1}) \geq 2(\delta_{C^2}p_{\min})^{-1} \cdot 2(K^*)^{1/2}\big\{\|Z\circ\Omega\|_2 + 2\delta_{C^2}C^2\|Q\|_2\big\}$

$$\begin{aligned}
&\phi\{l(\hat{M}^{(K)}, Y, \Omega) - l(M^*, Y, \Omega)\}\\
\leq&(2K^*)^{1/2}\big\{\|Z\circ\Omega\|_2 + 2\delta_{C^2}C^2\|Q\|_2\big\}d(M^*,{}_{K+1}) - \delta_{C^2}p_{\min}d^2(M^*,{}_{K+1})\\
\leq&-\frac{1}{2}\delta_{C^2}p_{\min}d^2(M^*, \mathcal{M}_{K+1}).
\end{aligned} \tag{S.B.37}$$

Note that $\phi\{l(\hat{M}^{(K^*)}, Y, \Omega) - l(M^*, Y, \Omega)\} \geq 0$. Thus, the above inequality implies that on the event $d(M^*, \mathcal{M}_K) \geq 4(\delta_{C^2}p_{\min})^{-1}(K^*)^{1/2}\big\{\|Z\circ\Omega\|_2 + 2\delta_{C^2}C^2\|Q\|_2\big\}$,

$$\phi\{l(\hat{M}^{(K)}, Y, \Omega) - l(\hat{M}^{(K^*)}, Y, \Omega)\} \leq -\frac{1}{2}\delta_{C^2}p_{\min}d^2(M^*, \mathcal{M}_K). \tag{S.B.38}$$

Now we proceed to a lower bound for $d(M^*, \mathcal{M}_{K+1})$. Recall the well-known fact that $\inf_{M \text{ has a rank } K+1} \|M^* - M\|_F^2 = \sum_{l=K+2}^{K^*+1}\sigma_l^2(M^*)$ where $\sigma_1(M^*) \geq \cdots \geq \sigma_{K^*+1}(M^*)$ denotes the non-zero singular values of M^*. Thus, $d(M^*, \mathcal{M}_{K+1}) \geq \{\sum_{l=K+2}^{K^*+1}\sigma_l^2(M^*)\}^{1/2} \geq \sigma_{K^*+1}(M^*)$. Combine this with (S.B.38), we have

$$\phi\big\{l(\hat{M}^{(K)}, Y, \Omega) - l(\hat{M}^{(K^*)}, Y, \Omega)\big\} \leq -\frac{1}{2}\delta_{C^2}p_{\min}\Big\{\sum_{l=K+2}^{K^*+1}\sigma_l^2(M^*)\Big\}, \tag{S.B.39}$$

if $4(\delta_{C^2}p_{\min})^{-1}(K^*)^{1/2}\big\{\|Z\circ\Omega\|_2 + 2\delta_{C^2}C^2\|Q\|_2\big\} \leq \sigma_{K^*+1}(M^*)$ and $K \leq K^* - 1$. We complete the proof by noting that $4(\delta_{C^2}p_{\min})^{-1}(K^*)^{1/2}\big\{\|Z\circ\Omega\|_2 + 2\delta_{C^2}C^2\|Q\|_2\big\} \leq \sigma_{K^*+1}(M^*)$ is equivalent to $4(\delta_{C^2}p_{\min})^{-1}K^*R \leq \sigma_{K^*+1}^2(M^*)$. □

Proof of Lemma 5. According to Lemma 9, for each $0 \leq K \leq K^* - 1$,

$$\begin{aligned}
&-2l(\hat{M}^{(K)}, Y, \Omega) + v(n, N, J, K) - \big\{-2l(\hat{M}^{(K^*)}, Y, \Omega) + v(n, N, J, K^*)\big\}\\
\geq&\phi^{-1}\delta_{C^2}p_{\min}\Big\{\sum_{l=K+2}^{K^*+1}\sigma_l^2(M^*)\Big\} - \sum_{l=K+1}^{K^*}u(n, N, J, l),
\end{aligned} \tag{S.B.40}$$

if $4(\delta_{C^2}p_{\min})^{-1}K^*R \leq \sigma_{K^*}^2(M^*)$. Clearly, right-hand-side of the above inequality is strictly greater than zero if $u(n, N, J, l) < \phi^{-1}\delta_{C^2}p_{\min}\sigma_{l+1}^2(M^*)$ for all $1 \leq l \leq K^*$. Thus,

$$\begin{aligned}
&\Pr\big(\hat{K} \geq K^*\big)\\
\geq&\Pr\Big(\max_{1\leq K\leq K^*}\big[-2l(\hat{M}^{(K)}, Y, \Omega) + v(n, N, J, K) - \big\{-2l(\hat{M}^{(K^*)}, Y, \Omega) + v(n, N, J, K^*)\big\}\big] > 0\Big)\\
\geq&\Pr\big(4(\delta_{C^2}p_{\min})^{-1}K^*R \leq \sigma_{K^*+1}^2(M^*) \text{ and } u(n, N, J, K) < \phi^{-1}\delta_{C^2}p_{\min}\sigma_{K+1}^2(M^*) \text{ for all } 1 \leq K \leq K^*\big)
\end{aligned} \tag{S.B.41}$$

The right-hand-side of the above inequality tend to one under the assumptions of the Lemma. This completes the proof. □

Proof of Lemma 6. According to Lemma 3, there is a universal constant c such that with probability least $1 - (N + J)^{-1} - (n^*)^{-1}$,

$$\|Z\circ\Omega\|_2 \leq 4(\phi\kappa_{2C^2})^{1/2}(\max_i n_{i\cdot}^*)^{1/2} \vee (\max_j n_{\cdot j}^*)^{1/2} + c\{(\phi/C^2) \vee (\phi\kappa_{2C^2})^{1/2}\}\log n^* \log^{1/2}(N + J). \tag{S.B.42}$$

Under the asymptotic regime (10), we have $4(\phi\kappa_{2C^2})^{1/2} = O(1)$, $\max_i n_{i\cdot}^* = O(p_{\max}J)$, $\max_j n_{\cdot j}^* = O(p_{\max}N)$, $c\{(\phi/C^2) \vee (\phi\kappa_{2C^2})^{1/2}\} = O(1)$, and $\log n^* \log^{1/2}(N + J) = O\big((N \wedge$

$J)^{-1/2}(n^*)^{1/2}\big) = O\big(\{p_{\max}(N \vee J)\}^{1/2}\big)$. Thus, the right-hand-side of (S.B.42) is of the order $O(\{p_{\max}(N \vee J)\}^{1/2})$ and

$$\|Z \circ \Omega\|_2 = O_p(\{p_{\max}(N \vee J)\}^{1/2}) \tag{S.B.43}$$

as $N, J \to \infty$. Similarly, according to Lemma 2,

$$\|Q\|_2 \leq 4(\max_i n_{i\cdot}^*)^{1/2} \vee (\max_j n_{\cdot j}^*)^{1/2} + c\log^{1/2}(N+J) \tag{S.B.44}$$

with probability at least $1-(N+J)^{-1}$. Under the the asymptotic regime (10), the right-hand-side of the above inequality is of the order $O(\{p_{\max}(N \vee J)\}^{1/2})$, and thus

$$\|Q\|_2 = O_p(\{p_{\max}(N \vee J)\}^{1/2}). \tag{S.B.45}$$

We complete the proof by combining (S.B.43), (S.B.45), and the definition of R. □

C. Proof of Proposition 1

Without loss of generality, assume $J \geq N$, N/K^* is an integer, and $\phi = 1$. The proof can be easily extended to the other cases.

To prove the lower bound for the minimax risk, we use a local Fano's method, which is a standard tool for proving lower error bounds (Tsybakov, 2008). Throughout the proof we use the notation $F = (F_1, \cdots, F_N)^T$, $d = (d_1, \cdots, d_J)^T$ and $A = (A_1, \cdots, A_J)^T$.

We start with constructing a local packing of as follows. First, let $F^{(0)} = C(I_{K^*}, \cdots, I_{K^*})^T$. Note that here we used the assumption that N/K^* is an integer. Also, let $d^{(0)} = (0, \cdots, 0)^T$. Next, according to the (Gilbert-Varshanmov bound) (Gilbert, 1952), there exists a set $\mathcal{B} = \{B^{(l)} : l = 1, \cdots L\} \subset \{1, -1\}^{J \times K^*}$ satisfying $L \geq \exp(JK^*/8)$ and

$$\sum_{j=1}^{J}\sum_{k=1}^{K^*} I(b_{jk} \neq b'_{jk}) \geq JK^*/4 \tag{S.C.1}$$

for any $B, B' \in \mathcal{B}$ and $B \neq B'$. Then, we construct a set $\mathcal{A} = \{A = \gamma B : B \in \mathcal{B}\}$ for some γ specified in the sequel. Now define

$$\begin{aligned}\mathcal{M}^* &= \{M = (m_{ij}) : m_{ij} = d_j + F_i^T A_j \text{ for all } i \text{ and } j \text{ where } d = d^{(0)}, F = F^{(0)} \text{ and } A \in \mathcal{A}\} \\ &= \{M = F^{(0)}A^T : A \in \mathcal{A}\}\end{aligned} \tag{S.C.2}$$

The set $\mathcal{M}^*$ defined above has the following properties.

(a) $|\mathcal{M}^*| = L \geq \exp(JK^*/8)$.
(b) The Kullback-Leibler divergence $\max_{M,M' \in \mathcal{M}^*} KL(P_M \| P_{M'}) \leq \kappa n^* \gamma^2$ for some constant κ, where P_M denotes the probability measure for $(Y_{ij}, \omega_{ij})_{1 \leq i \leq N, 1 \leq j \leq J}$ when the true parameter is M.
(c) $\|M - M'\|_F^2 \geq C^2 NJ\gamma^2$ for $M, M' \in \mathcal{M}^*$ and $M \neq M'$.

Property (a) holds obviously. Property (b) holds because of the following inequalities

$$\begin{aligned}KL(P_M \| P_{M'}) &= \sum_i \sum_j p_{ij}\{b'(m_{ij})(m_{ij} - m'_{ij}) - (b(m_{ij}) - b(m'_{ij})\} \\ &\leq p_{\max} \sum_i \sum_j \{b'(m_{ij})(m_{ij} - m'_{ij}) - (b(m_{ij}) - b(m'_{ij})\} \\ &\leq \kappa n^*/(NJ)\|M - M'\|_F^2 \\ &= \kappa C^2 n^*/(NJ) \cdot (N/K^*) \sum_j \sum_k (a_{jk} - a'_{jk})^2,\end{aligned} \tag{S.C.3}$$

where we used the construction $M = F^{(0)}A^T$ in the last equation. Note that $|a_{jk}| = |\gamma b_{jk}| = \gamma$ for $A \in \mathcal{A}$. Thus, $(a_{jk} - a'_{jk})^2 \le 4\gamma^2$, which leads to property (b) of the set $\mathcal{M}^*$ for a possibly different κ. Property (c) holds for the following reasons. By construction, for $M, M' \in \mathcal{M}^*$ and $M \neq M'$

$$\begin{aligned} \|M - M'\|_F^2 &= C^2 N/K^* \cdot \sum_j \sum_k (a_{jk} - a'_{jk})^2 \\ &= C^2 (N/K^*)\gamma^2 \sum_j \sum_k (b_{jk} - b'_{jk})^2 \\ &= C^2 (N/K^*)\gamma^2 \cdot 4 \sum_j \sum_k I(b_{jk} \neq b'_{jk}) \\ &\ge C^2 (N/K^*)\gamma^2 JK^*, \end{aligned} \tag{S.C.4}$$

where the last inequality is due to (S.C.1). Thus, property (c) holds.

Now, for an arbitrary estimator $\bar{M}$, define a new estimator $\tilde{\bar{M}} = \arg\min_{W \in \mathcal{M}^*} \|W - \bar{M}\|_F$. It is easy to see that for $M \in \mathcal{M}^*$,$\|\tilde{\bar{M}} - M\|_F \le 2\|\bar{M} - M\|_F$. By a version of Fano's inequality, we have

$$\max_{M^* \in \mathcal{M}^*} P_{M^*}\left(\tilde{\bar{M}} \neq M^*\right) \ge 1 - \frac{\kappa n^* \gamma^2 + 1}{\log |\mathcal{M}^*|} \ge 1 - \frac{\kappa n^* \gamma^2 + 1}{JK/8}. \tag{S.C.5}$$

Choose $\gamma = \kappa^{-1}(JK/n^*)^{1/2}$ for a possibly different κ, then for $JK \ge 64$, we have

$$\max_{M^* \in \mathcal{W}_M} P_{M^*}\left(\tilde{\bar{M}} \neq M^*\right) \ge \frac{1}{2}. \tag{S.C.6}$$

Furthermore, we have

$$\begin{aligned} &\max_{M^* \in \mathcal{M}^*} P_{M^*}\left(\|\bar{M} - M^*\|_F \ge (1/2) \cdot C(NJ)^{1/2}\gamma\right) \\ \ge &\max_{M^* \in \mathcal{M}^*} P_{M^*}\left(\|\tilde{\bar{M}} - M^*\|_F \ge C(NJ)^{1/2}\gamma\right) \\ \ge &\max_{M^* \in \mathcal{M}^*} P_{M^*}\left(\tilde{\bar{M}} \neq M^*\right) \\ \ge &\frac{1}{2}. \end{aligned} \tag{S.C.7}$$

Simplifying the term $(1/2) \cdot C(NJ)^{1/2}\gamma$, we arrive at

$$\max_{M^* \in \mathcal{M}^*} P_{M^*}\left((NJ)^{-1/2}\|\bar{M} - M^*\|_F \ge 2^{-1}\kappa^{-1} \cdot (JK/n^*)^{1/2}\right) \ge \frac{1}{2}. \tag{S.C.8}$$

Note that for $A \in \mathcal{A}$, $\|A_j\| \le \gamma\sqrt{K^*} = \kappa^{-1}(J(K^*)^2/n^*)^{1/2}$. Thus, for a possibly larger constant κ, we have $\|A_j\| \le C$ under the assumption $(K^*)^2(J+N) \le n^*$. Thus, $\mathcal{M}^*$ is a subset of the parameter space of interest. That is,

$$\mathcal{M}^* \subset \mathcal{G} := \{M = (m_{ij}) : m_{ij} = d_j + F_i^T A_j, \text{ and } (\|F_i\|^2 + 1)^{\frac{1}{2}} \le C, ((d_j)^2 + \|A_j\|^2)^{\frac{1}{2}} \le C, \text{ for all } i\} \tag{S.C.9}$$

This further implies

$$\max_{M^* \in \mathcal{G}} P_{M^*}\left(\frac{1}{\sqrt{NJ}}\|\bar{M} - M^*\|_F \ge 2^{-1}\kappa^{-1} \cdot (JK^*/n^*)^{1/2}\right) \ge \frac{1}{2}. \tag{S.C.10}$$

This completes our proof.

D. ON OPTIMIZATION FOR JOINT LIKELIHOOD

We provide some discussions on the optimization problem (3) for the constrained joint maximum likelihood estimator. The two reasons below explain why the solution given by an alternating maximization

algorithm typically performs well, even though (3) is a non-convex optimization problem. First, according to the proofs of Theorems 1 through 4, Theorems 1 and 2 hold as long as the estimates satisfy

$$l_K(\hat{F}_1, ..., \hat{F}_N, \hat{A}_1, \hat{d}_1, ..., \hat{A}_J, \hat{d}_J) \geq l_{K^*}(F_1^*, ..., F_N^*, A_1^*, d_1^*, ..., A_J^*, d_J^*)$$

when $K \geq K^*$. In addition, for Theorems 3 and 4 to hold, we only need

$$l_{K^*}(\hat{F}_1, ..., \hat{F}_N, \hat{A}_1, \hat{d}_1, ..., \hat{A}_J, \hat{d}_J) \geq l_{K^*}(F_1^*, ..., F_N^*, A_1^*, d_1^*, ..., A_J^*, d_J^*). \tag{S.D.1}$$

It means that the number of factors can be consistently selected even if our estimate is not a global solution to (3) as long as (S.D.1) holds.

Second, we use good starting points when solving the optimization (3). Specifically, under the logistic factor model for binary data, a singular-value-decomposition-based algorithm is proposed by Zhang et al. (2020) that is guaranteed to give a consistent estimator of the model parameters. Although this estimator is statistically less efficient than the joint-likelihood-based estimator (thus cannot be directly plugged into the likelihood to construct an information criterion), it can serve as a good starting point when solving the optimization (3). For other models, similar singular-value-decomposition-based algorithms can also be developed.

We also discuss the choice of constraint constant C which needs to be specified when computing the constrained joint maximum likelihood estimator. First of all, we point out that it is standard to impose such a constraint for low-rank matrix estimation under nonlinear models. For example, in the work of Cai & Zhou (2013) on 1-bit matrix completion, it is required that the max norm (i.e., the maximum value of the absolute values of entries) of underlying low-rank matrix is smaller than a constant, which plays essentially the same role as the constant C in the current work. Second, according to our simulation study, the estimation of the model parameters and the performance of the proposed information criteria are not sensitive to the choice of C, as long as it is set to be sufficiently large. Given a specific dataset, we suggest to run the estimator under different values of C to check its sensitivity. In practice, we suggest to start with a sufficiently large C, followed by a sensitivity analysis to check whether the estimator is sensitive to the current choice of C.

E. Information Criteria based on Marginal Likelihood

We provide some discussion on the behavior of the maximum marginal likelihood when both N and J grow to infinity. To simplify the discussion, we assume there is no missing value and the dispersion parameter is 1, but this discussion can be generalized to the case when there are missing data and the dispersion parameter needs to be estimated. Consider a model with K factors. The marginal likelihood approach assumes that the factors F_1, ..., F_N are i.i.d. samples from a known distribution h. Then the marginal likelihood function takes the form

$$m_K(A, D) = \sum_{i=1}^{N} \log\Big(\int \exp\left(l_i(x, A, D)\right) h(x) dx\Big),$$

where $l_i(x, A, D) = \sum_{j=1}^{J} \log g(y_{ij}|A_j, d_j, x, 1)$, $A = (A_j : j = 1, ..., J)$, $D = (d_j : j = 1, ..., J)$, and $x \in \mathbb{R}^K$. Let $(\hat{A}, \hat{D})$ be the estimator based on the marginal likelihood, i.e., $(\hat{A}, \hat{D}) \in \arg\max m_K(A, D)$. Furthermore, let

$$\hat{F}_i = \arg\max_x \; l_i(x, \hat{A}, \hat{D}) \log(h(x)).$$

Then by the Laplace approximation (Huber et al., 2004) and under suitable regularity conditions, we should be able to establish

$$m_K(\hat{A},\hat{D}) = \sum_{i=1}^{N}\sum_{j=1}^{J}\log g(y_{ij}|\hat{A}_j,\hat{d}_j,\hat{F}_i,1) + \sum_{i=1}^{N}\log(h(\hat{F}_i)) + \frac{NK}{2}\log(2\pi/J) - \frac{1}{2}\sum_{i=1}^{N}\log\Big(det(H(\hat{F}_i,\hat{A},\hat{D}))\Big) + R_{N,J}, \tag{S.E.1}$$

where $H(\hat{F}_i,\hat{A},\hat{D})$ is the Hessian matrix of $L_i(x) = l_i(x,\hat{A},\hat{D})$ evaluated at $\hat{F}_i$ and the $R_{N,J}$ term comes from the remainder term of Laplace approximation. Note that the first term in (S.E.1) is the dominant term that takes the same form as the joint likelihood, though $\hat{A}_j, \hat{d}_j$, and $\hat{F}_i$ are obtained from the marginal likelihood. The remainder $R_{N,J}$ is a term with a smaller asymptotic order. Moreover, we believe that the error bound established in Theorem 1 can be extended to $\hat{M}_K = (\hat{d}_j + \hat{F}_i^T\hat{A}_j)_{N\times J}$ when $K^* \le K \le K_{max}$. Therefore, the development in this article will also be useful when developing marginal-likelihood-based information criteria for generalized latent factor models under a high-dimensional setting.

F. COMPARISON WITH SOME RELATED WORKS

As discussed in Remark 2, the error bound (5) improves several recent results on low-rank matrix estimation and completion. We now summarize the comparison in Remark 2 using Table F.1 below. This comparison focuses on the error bound (5) when $K_{max} = K^*$ and data entries are binary and are uniformly missing.

	Key setting on M	K^*	Error bound
Current	$\text{Rank}(M) \le K^*$	Can diverge	$O_p\Big[\Big\{\frac{K^*(N\vee J)}{n^*}\Big\}^{1/2}\Big]$
Chen et al. (2020)	$\text{Rank}(M) \le K^*$	Fixed	$O_p\Big[\Big\{\frac{(N\vee J)}{n^*} + \frac{NJ}{(n^*)^{3/2}}\Big\}^{1/2}\Big]$
Bhaskar & Javanmard (2015)	$\text{Rank}(M) \le K^*$	Can diverge	$O_p\Big[\frac{K^*(N\vee J)^{1/2}}{(n^*)^{1/2}} + \frac{(N\vee J)^3(N\wedge J)^{1/2}(K^*)^{3/2}}{(n^*)^2}\Big]$
Ni & Gu (2016)	$\text{Rank}(M) \le K^*$	Can diverge	$O_p\Big[\Big\{\frac{K^*(N\vee J)\log(N+J)}{n^*}\Big\}^{1/2}\Big]$
Cai & Zhou (2013)	$\|M\|_* \le \alpha\sqrt{K^*NJ}$	Can diverge	$O_p\Big[\Big\{\frac{K^*(N\vee J)}{n^*}\Big\}^{1/4}\Big]$
Davenport et al. (2014)	$\|M\|_* \le \alpha\sqrt{K^*NJ}$	Can diverge	$O_p\Big[\Big\{\frac{K^*(N\vee J)}{n^*}\Big\}^{1/4}\Big]$

Table F.1: Comparison with existing results on the recovery of M. Here, $\|\cdot\|_*$ denotes the matrix nuclear norm and α is a positive constant.

We further compare the current development with Chen et al. (2019) and Chen et al. (2020) that also concern likelihood-based analysis of generalized latent factor models. We discuss the similarities and differences below.

1. Model: Chen et al. (2020) and the current work consider the same generalized latent factor model (1) and Chen et al. (2019) consider the special case for binary data as given in Example 1.
2. Estimation versus selection: The current work establishes results on both the estimation of generalized latent factor model and information criteria for the selection of factors. In contrast, Chen et al. (2019) and Chen et al. (2020) focus on the estimation problem.
3. Confirmatory versus exploratory setting: Chen et al. (2019) and the current work consider an exploratory factor analysis setting, for which no prior knowledge is assumed on the factor structure. Chen et al. (2020) focus on a confirmatory factor analysis setting though its results are also generally applicable under an exploratory setting.
4. Setting on missingness: The current work considers a flexible setting for the missingness of data entries that allows the entries to be non-uniformly missing. In contrast, Chen et al. (2019) and Chen et al.

	$N = J$						$N = 5J$					
	S1			S2			S1			S2		
Average time	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
$J = 100$	6	9	11	5	9	11	36	48	73	33	44	67
$J = 200$	21	25	37	17	22	36	217	195	314	201	175	296
$J = 300$	53	58	86	46	49	78	509	522	785	483	486	714
$J = 400$	106	108	161	92	97	142	966	1144	1423	869	1131	1329

Table G.1: The average computation time (in seconds) for running one independent replication for each of the 48 simulation settings.

(2020) consider a uniformly missing setting which can be viewed as a special case of the current setting.

5. Optimality: Both the current work and Chen et al. (2020) establish minimax optimality results on the estimation of generalized latent factor models. The current optimality result, which is established under a more general setting, can be viewed as an extension of that of Chen et al. (2020). Minimax optimality is not considered in Chen et al. (2019).

In summary, the new contribution of the current paper is of twofold. First, we propose information criteria for selecting the number of factors in high-dimensional generalized latent factor models and establish conditions under which selection consistency is guaranteed. Second, we substantially extend the results on the estimation of generalized latent factor models under a general setting where the data entries can be non-uniformly missing and the number of factors can also grow to infinity.

G. Additional Simulation Results

G.1. *Additional Results for Simulation in Section 4.1*

The average running time for one independent replication for each of the 48 simulation settings is given in Table G.1, where the computation is run on a computer with an Intel(R) Xeon(R) CPU 2.30GHz. The computation code for our simulations can be found on the author's Github page: `https://github.com/yunxiaochen/JML_IC`.

G.2. *Simulation under Poisson Factor Model*

We further provide a simulation study under the Poisson factor model as given in Example 2. Similar to the simulation study in Section 4.1, we consider the same factor strength settings S1 and S2, and the same missing data settings M1-M3. Again, we consider two relationships between N and J, including $N = J$ and $N = 5J$. We consider $J = 100, 200, 300, 400$. Again, we let $K^* = 3$ and the true model parameters be generated the similarly as the simulation study in Section 4.1. More precisely, under the setting S1, the true parameters d_j^*, a_{j1}^*, ..., a_{j3}^* are generated by sampling independently from the uniform distribution over the interval $[-1, 1]$ and the true factor values are generated f_{i1}^*, ..., f_{i3}^* are generated by sampling independently from the uniform distribution over the interval $[-1, 1]$. Under the setting S2, f_{i3}^* is generated from the uniform distribution over the interval $[-0.4, 0.4]$ and the rest of the parameters are generated the same as those in S1. We use the proposed JIC to select K from the candidate set $\{1, 2, 3, 4, 5\}$ and the constraint constant C in (3) is set to be 3. The true model parameters satisfy this constraint. There are 48 simulation settings in total and 100 independent replications are run for each setting. Figure G.1 below shows the value of $\max_{3\le K\le 5}\{(NJ)^{-1/2}\|\hat{M}^{(K)} - M^*\|_F\}$ under different settings. Table G.2 shows the accuracy on determining the number of factors. Finally, Table G.3 gives the average average running time for one independent replication for each of the 48 simulation settings.

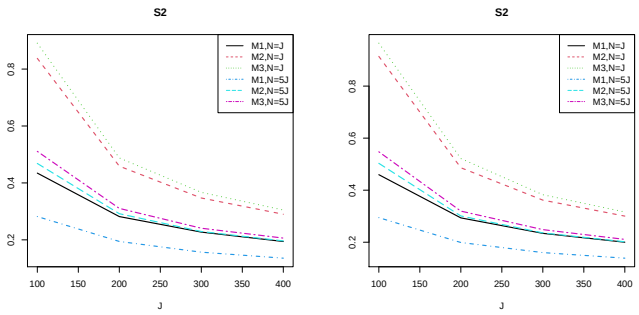

Fig. G.1: The loss $\max_{3\le K\le 5}\{(NJ)^{-1/2}\|\hat{M}^{(K)} - M^*\|_F\}$ for the recovery of the low-rank matrix M^*, where each point is the mean loss calculated by averaging over 100 independent replications. Panels (a) and (b) show the results under the two different factor strength settings, S1 and S2, respectively.

	$N = J$						$N = 5J$					
	S1			S2			S1			S2		
Under-selection	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
$J = 100$	0	0	0	58	40	33	0	0	0	100	100	100
$J = 200$	0	0	0	3	47	49	0	0	0	12	100	100
$J = 300$	0	0	0	0	2	3	0	0	0	0	87	83
$J = 400$	0	0	0	0	0	0	0	0	0	0	2	2
Over-selection												
$J = 100$	0	19	13	0	19	9	0	0	0	0	0	0
$J = 200$	0	0	0	0	0	0	0	0	0	0	0	0
$J = 300$	0	0	0	0	0	0	0	0	0	0	0	0
$J = 400$	0	0	0	0	0	0	0	0	0	0	0	0

Table G.2: The number of times that the true number of factors is under- or over-selected selected among 100 independent replications under each of the 48 simulation settings.

	$N = J$						$N = 5J$					
	S1			S2			S1			S2		
Average time	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
$J = 100$	1	1	1	1	1	1	12	6	7	11	6	7
$J = 200$	10	5	5	9	5	5	79	40	43	73	35	39
$J = 300$	28	14	16	25	13	14	249	122	125	222	107	110
$J = 400$	60	30	32	53	28	29	536	267	278	475	229	242

Table G.3: The average computation time (in seconds) for running one independent replication for each of the 48 simulation settings.

G.3. *A Scree Plot Example*

Scree plots are a widely used tool for selecting the number of factors in factor analysis (Cattell, 1966). A scree plot displays the eigenvalues of the covariance matrix of data in a downward curve, ordering the eigenvalues from largest to smallest. The number of factors is then determined by finding the "elbow" of the graph. The "elbow" is the eigenvalue where the eigenvalues seem to level off and the number of factors is determined by the number of eigenvalues that are greater than the elbow. This approach typically works

well for data following a linear factor model. This is because, under a linear factor model, the covariance matrix of data is approximately a low-rank matrix plus a diagonal matrix, where the low-rank part drives the "elbow" phenomenon. When data are generated from a nonlinear factor model, such as the logistic or Poisson factor models considered in the current work, the factor structure of data cannot be fully characterized by the covariance matrix. In particular, the covariance matrix cannot be approximated by a low-rank matrix plus a diagonal matrix. As a result, the elbow of the scree plot may no longer correspond to the number of factors. We provide a simulated example to illustrate this point. Figure G.2 shows the scree plot for data generated from a Poisson factor model under a setting when $N = J = 200$, $K^* = 3$, and there are no missing entries. Based on the scree plot, one may tend to choose seven or eight factors, which is larger than the true number of factors.

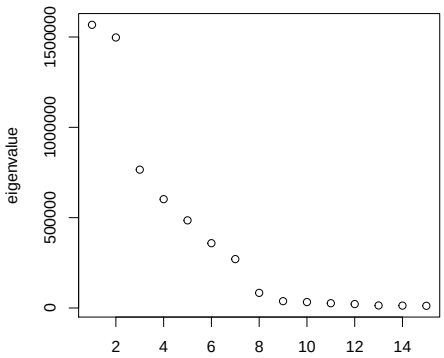

Fig. G.2: The scree plot for data that are generated from a Poisson factor model with three factors

G.4. *Comparison with Bai et al. (2018)*

We now compare the proposed JIC with the method proposed in Bai et al. (2018) via a simulation study. In this study, data are generated from a linear factor model, with $N = 2J$. More specifically, y_{ij} follows a normal distribution with mean $d_j + A_j^T F_i$ and variance 1, so that the data follow a spike covariance structure model assumed in Bai et al. (2018), with the eigenvalues of the covariance matrix of $(y_{i1}, ..., y_{iJ})^T$ satisfying $\lambda_{K^*} > \lambda_{K^*+1} = \cdots = \lambda_J = 1$. Again, we let $K^* = 3$ and the true model parameters be generated under the setting S1 the simulation study in Section 4.1, where the three factors are of the same strength. More precisely, the true parameters d_j^*, a_{j1}^*, ..., a_{j3}^* are generated by sampling independently from the uniform distribution over the interval $[-2, 2]$ and the true factor values are generated f_{i1}^*, ..., f_{i3}^* are generated by sampling independently from the uniform distribution over the interval $[-2, 2]$. We consider $J = 10, 20, ..., 50$ and $N = 2J$. Note that under this linear factor model with no missing data and assuming that the variance of y_{ij} is known to be 1, then the proposed JIC is the same as the PC_{p3} criterion proposed in Bai & Ng (2002).

We use the proposed JIC to select K from the candidate set $\{1, 2, 3, 4, 5\}$ and the constraint constant C in (3) is set to be 5. In addition, we also use the AIC and BIC proposed in Bai et al. (2018) to select K from the candidate set $\{1, 2, 3, 4, 5\}$. The results are given in Table G.4. As we can see, all three information criteria become more accurate when N and J simultaneously grow. Specifically, the proposed JIC and the BIC in Bai et al. (2018) perform similarly. When $J \geq 20$ and $N = 2J$, both methods correctly identify the true number of factors all the time. When $J = 10$ and $N = 20$, both the JIC and the BIC are correct 92% of the times, though the two methods are slightly different in the numbers of over- and under-selections. Finally, consistent with the observations in Bai et al. (2018), the AIC is less accurate and tends to over-select.

	Under-selection			Over-selection		
	JIC	BIC	AIC	JIC	BIC	AIC
$J = 10$	2	4	2	6	4	31
$J = 20$	0	0	0	0	0	18
$J = 30$	0	0	0	0	0	7
$J = 40$	0	0	0	0	0	3
$J = 50$	0	0	0	0	0	2

Table G.4: The number of times that the true number of factors is under- or over-selected selected among 100 independent replications under each of the 5 simulation settings, under a linear factor model with a spike covariance structure.

H. ADDITIONAL RESULTS FOR REAL DATA ANALYSIS

In what follows, we provide additional results for the real data analysis. In Tables H.1 and H.2, we show the loading matrix and the sample covariance matrix for the estimated factor scores, after applying the oblimin rotation. Note that the items have been reordered, with items 1-32, 33-55, and 56-79 designed to measure the psychoticism, extraversion, and neuroticism traits, respectively. The content of the items can be found in Eysenck et al. (1985). Note that our data have been pre-processed so that the negatively worded items are reversely scored. As we can see, items 1-32, 33-55, and 56-79 tend to have high loadings on F2, F1, and F3, respectively. According to Table H.2, the correlations between the three estimated factors are relatively small, suggesting that the three factors tend to be uncorrelated.

We further provide results for the two- and four-factor models, whose JIC values are also relatively small. These results may provide us further insights about the latent structure of this personality inventory. Tables H.3 and H.5 provide the the loading matrices for the two models, respectively, after applying the oblimin rotation. Moreover, Tables H.4 and H.6 show the sample covariance matrices for the estimated factor scores, from the two models, respectively. According to Table H.3, the items that are designed to measure the extraversion trait tend to have high loadings for the first factor and items designed to measure neuroticism tend to have high loadings for the second factor, while most items designed to measure psychoticism have small loadings for both factors. These results suggest that the psychoticism factor may not be captured by the two-factor model.

From the loading structure given in Table H.5, the extracted factors F4, F2, and F3 tend to correspond to the psychoticism, extraversion, and neuroticism traits, respectively. In addition, most items have small loadings on F1, except for items 14. "Do you stop to think things over before doing anything?", 28. "Do you generally 'look before you leap'?", 45. "Have people said that you sometimes act too rashly?", and 48. "Do you often make decisions on the spur of the moment?", where items 14 and 28 are negatively worded and thus reversely scored. It seems a minor factor about impulsive decision.

We compare the proposed method with the classical Akaike information criterion (AIC) and Bayesian information criterion (BIC) calculated based on the marginal likelihood function, where the latent factors are treated as random variables. More specifically, the latent factors are assumed to follow a multivariate normal distribution, in the calculation of the marginal likelihood. The marginal maximum likelihood estimator is computed using the R package "mirt" (Chalmers, 2012), where the computation for the marginal maximum likelihood estimator is carried out using an Expectation-Maximization (EM) algorithm. The EM algorithm is very time-consuming when only involving a moderate number of factors (Reckase, 2009). The AIC and BIC values for the one- through five-factor models are given in Table H.7 below. In calculating the AIC and BIC values, the number of parameters for a K-factor model is $J(K+1) - K(K-1)/2$, recalling that J is the number of items. The three-factor model fits best according to the BIC value, which is consistent with the selection based on the proposed JIC. On the other hand, AIC selects the four-factor model. Note that under the classical asymptotic regime and the true model is one of the candidate models, the BIC guarantees consistency for model selection, while the AIC tends to over-select (Shao, 1997).

Item	F1	F2	F3	Item	F1	F2	F3
1	0.31	2.33	0.42	**41**	1.84	-0.23	-0.09
2	0.34	1.37	-0.11	**42**	2.98	0.36	-0.07
3	0.53	1.18	0.49	**43**	0.91	-0.07	-0.05
4	0.27	1.47	0.79	**44**	2.59	-0.98	0.15
5	0.89	1.37	0.03	**45**	1.19	0.96	0.65
6	0.44	1.11	0.23	**46**	0.49	-0.02	-0.10
7	-0.25	1.99	0.04	**47**	0.79	0.36	-0.33
8	0.35	0.83	-0.23	**48**	0.93	0.59	0.18
9	-0.58	1.16	0.50	**49**	0.43	-0.02	0.11
10	-0.04	1.59	0.71	**50**	2.59	-0.01	-0.12
11	0.22	0.85	-0.10	**51**	1.92	-0.12	0.00
12	0.03	1.78	0.36	**52**	3.78	-0.02	0.10
13	0.03	0.45	0.50	**53**	3.79	0.54	-0.16
14	0.92	0.95	0.27	**54**	1.81	-0.18	-0.01
15	-0.15	1.04	-0.97	**55**	2.73	0.08	-0.08
16	0.55	1.13	-0.53	**56**	0.34	0.73	2.29
17	0.08	0.63	-0.01	**57**	0.13	0.41	1.58
18	-0.06	0.93	-0.35	**58**	0.37	-1.10	2.13
19	0.13	0.58	-0.31	**59**	0.01	0.67	1.64
20	0.08	1.78	-0.22	**60**	-0.01	-0.18	1.78
21	-0.50	2.37	-0.63	**61**	0.01	0.45	2.10
22	-0.49	2.17	-0.64	**62**	0.39	-0.02	1.68
23	-0.54	1.55	0.02	**63**	-0.47	0.11	2.10
24	0.23	1.15	-0.47	**64**	-0.35	-0.54	2.84
25	0.18	0.77	-0.06	**65**	-0.09	-0.18	1.38
26	-0.35	1.15	0.10	**66**	-0.23	0.49	1.91
27	0.44	1.85	0.13	**67**	0.13	-0.07	0.88
28	0.95	1.02	0.38	**68**	0.04	0.34	0.59
29	-0.16	0.50	0.45	**69**	0.16	0.40	1.25
30	0.16	1.31	-0.23	**70**	0.04	0.61	1.36
31	-0.08	1.25	-0.05	**71**	0.71	-0.16	1.17
32	-0.18	0.58	-0.25	**72**	-0.11	0.73	0.77
33	0.33	-0.17	-0.24	**73**	-0.28	-0.58	2.25
34	2.75	-0.19	0.48	**74**	-0.23	0.30	2.02
35	3.61	-0.31	-0.08	**75**	-0.17	0.70	1.55
36	2.05	0.08	-0.10	**76**	-0.26	-0.42	1.81
37	2.08	-0.34	-0.41	**77**	0.85	0.38	1.48
38	1.59	0.03	0.03	**78**	0.31	0.08	1.32
39	1.97	-0.77	-0.44	**79**	0.45	0.53	1.00
40	1.00	0.18	-0.58				

Table H.1: Estimated loading matrix for the three-factor model after applying the oblimin rotation.

Finally, we provide the estimation results for the three-factor model from the marginal-likelihood approach, in comparison with those from the joint-likelihood approach. The results are given in Tables H.8 and H.9. Similar to the analysis above, the results are under the oblimin rotation. As we can see, although the estimates are slightly different from those given by the joint likelihood, the loading structure is similar and suggests that the three factors correspond to the extraversion, psychoticism, and neuroticism traits, respectively.

	F1	F2	F3
F1	1.00	-0.03	-0.19
F2	-0.03	1.00	-0.02
F3	-0.19	-0.02	1.00

Table H.2: The sample covariance matrix for the estimated factor scores, under the three-factor model after applying the oblimin rotation. Note that the model parameters have been rescaled, so that the sample variance for each factor is one.

Item	F1	F2	Item	F1	F2
1	0.59	0.80	**41**	1.69	-0.13
2	0.46	0.17	**42**	2.54	-0.01
3	0.72	0.78	**43**	0.85	-0.08
4	0.48	1.01	**44**	2.18	-0.14
5	1.04	0.38	**45**	1.24	0.81
6	0.55	0.39	**46**	0.46	-0.08
7	0.16	0.40	**47**	0.86	-0.22
8	0.45	0.01	**48**	1.04	0.31
9	-0.36	0.68	**49**	0.40	0.10
10	0.22	0.89	**50**	2.38	-0.10
11	0.38	0.11	**51**	1.88	-0.02
12	0.45	0.68	**52**	3.51	0.10
13	0.19	0.64	**53**	3.79	0.03
14	0.98	0.48	**54**	1.74	-0.06
15	0.11	-0.53	**55**	2.85	-0.05
16	0.73	-0.20	**56**	0.43	2.38
17	0.25	0.18	**57**	0.15	1.62
18	0.19	-0.10	**58**	-0.09	1.33
19	0.27	-0.11	**59**	0.12	1.69
20	0.32	0.15	**60**	-0.14	1.54
21	0.11	0.04	**61**	0.08	2.10
22	0.12	0.02	**62**	0.24	1.46
23	-0.14	0.39	**63**	-0.55	1.93
24	0.43	-0.13	**64**	-0.54	2.07
25	0.35	0.12	**65**	-0.16	1.17
26	-0.11	0.32	**66**	-0.18	1.99
27	0.71	0.55	**67**	0.06	0.79
28	1.06	0.63	**68**	0.10	0.63
29	-0.02	0.54	**69**	0.22	1.34
30	0.35	0.09	**70**	0.09	1.44
31	0.14	0.21	**71**	0.53	0.99
32	-0.03	-0.09	**72**	0.02	0.88
33	0.24	-0.28	**73**	-0.48	1.65
34	2.56	0.39	**74**	-0.24	1.97
35	3.29	-0.14	**75**	-0.07	1.59
36	2.12	-0.07	**76**	-0.40	1.48
37	1.98	-0.50	**77**	0.86	1.52
38	1.60	0.06	**78**	0.22	1.26
39	1.62	-0.59	**79**	0.51	1.13
40	1.05	-0.48			

Table H.3: Estimated loading matrix for the two-factor model after applying the oblimin rotation.

	F1	F2
F1	1.00	-0.22
F2	-0.22	1.00

Table H.4: The sample covariance matrix for the estimated factor scores, under the two-factor model after applying the oblimin rotation.

Item	F1	F2	F3	F4	Item	F1	F2	F3	F4
1	0.48	0.39	0.48	2.31	**41**	-0.19	1.97	-0.06	-0.11
2	0.50	0.25	-0.09	1.31	**42**	-0.29	3.96	0.03	0.72
3	0.31	0.56	0.54	1.43	**43**	-0.20	1.05	0.00	0.00
4	0.22	0.34	0.86	1.49	**44**	-0.48	2.76	0.22	-0.68
5	0.67	0.74	0.02	1.17	**45**	2.25	0.60	0.58	0.32
6	0.53	0.25	0.20	0.92	**46**	0.10	0.48	-0.10	-0.04
7	-0.19	0.12	0.17	2.68	**47**	0.61	0.57	-0.39	0.16
8	0.06	0.41	-0.20	0.96	**48**	5.43	0.32	-0.14	-0.69
9	0.00	-0.45	0.53	1.36	**49**	0.36	0.31	0.09	-0.25
10	0.07	0.08	0.81	1.89	**50**	-0.32	3.33	-0.05	0.22
11	0.37	0.08	-0.12	0.73	**51**	0.35	1.78	-0.04	-0.30
12	-0.12	0.30	0.46	2.36	**52**	0.49	3.55	0.08	-0.37
13	0.24	-0.01	0.50	0.38	**53**	0.24	3.96	-0.12	0.52
14	4.87	-0.27	-0.08	0.32	**54**	-0.26	1.97	0.04	-0.06
15	0.30	-0.26	-1.00	1.03	**55**	0.23	2.61	-0.08	-0.02
16	0.98	0.15	-0.67	0.73	**56**	0.84	0.07	2.20	0.33
17	0.43	-0.07	-0.03	0.53	**57**	0.57	-0.08	1.50	0.14
18	0.24	-0.05	-0.47	1.23	**58**	-0.15	0.47	2.14	-1.22
19	-0.06	0.24	-0.27	0.69	**59**	0.53	-0.13	1.65	0.43
20	0.07	0.22	-0.17	2.06	**60**	0.26	-0.08	1.74	-0.41
21	-0.47	0.07	-0.44	3.15	**61**	0.42	-0.09	2.05	0.24
22	-0.38	-0.03	-0.51	3.08	**62**	-0.01	0.44	1.73	-0.05
23	0.30	-0.52	0.08	1.55	**63**	-0.42	-0.26	2.48	0.35
24	0.23	0.21	-0.46	1.18	**64**	-0.65	-0.02	3.18	-0.36
25	0.12	0.15	-0.06	0.78	**65**	-0.27	0.10	1.52	-0.05
26	-0.07	-0.24	0.14	1.35	**66**	-0.05	-0.12	2.13	0.60
27	0.84	0.23	0.11	1.51	**67**	-0.17	0.25	0.94	0.03
28	5.39	-0.19	0.11	0.35	**68**	0.23	-0.02	0.60	0.26
29	0.11	-0.13	0.47	0.48	**69**	0.61	-0.10	1.19	0.17
30	0.13	0.25	-0.18	1.52	**70**	0.56	-0.16	1.32	0.41
31	-0.07	0.15	0.04	1.55	**71**	0.11	0.69	1.16	-0.25
32	-0.18	-0.01	-0.20	0.79	**72**	0.11	-0.09	0.83	0.72
33	-0.24	0.46	-0.20	-0.12	**73**	-0.05	-0.29	2.25	-0.63
34	0.46	2.55	0.45	-0.44	**74**	-0.29	-0.01	2.42	0.47
35	0.12	3.67	-0.05	-0.48	**75**	0.20	-0.17	1.57	0.63
36	-0.06	2.19	-0.05	0.13	**76**	0.32	-0.40	1.80	-0.73
37	-0.85	2.48	-0.25	0.00	**77**	0.94	0.46	1.41	-0.07
38	-0.07	1.66	0.06	0.04	**78**	0.57	0.08	1.29	-0.26
39	-0.27	2.08	-0.42	-0.71	**79**	0.53	0.29	0.99	0.36
40	0.71	0.71	-0.71	-0.13					

Table H.5: Estimated loading matrix for the four-factor model after applying the oblimin rotation.

	F1	F2	F3	F4
F1	1.00	0.22	-0.07	0.07
F2	0.22	1.00	-0.20	-0.03
F3	-0.07	-0.20	1.00	0.01
F4	0.07	-0.03	0.01	1.00

Table H.6: The sample covariance matrix for the estimated factor scores, under the four-factor model after applying the oblimin rotation.

K	1	2	3	4	5
AIC	66304	63434	61942	**61732**	61750
BIC	67049	64547	**63418**	63566	63937

Table H.7: AIC and BIC values based on the marginal likelihood for the one- through five-factor models.

Item	F1	F2	F3
1	0.16	1.57	0.37
2	0.20	1.12	-0.15
3	0.47	1.10	0.57
4	0.14	1.11	0.73
5	0.84	1.08	0.04
6	0.27	0.91	0.15
7	-0.30	1.38	0.02
8	0.22	0.72	-0.24
9	-0.63	0.91	0.48
10	-0.20	1.21	0.59
11	0.17	0.70	-0.07
12	-0.12	1.39	0.30
13	0.00	0.43	0.47
14	0.72	0.77	0.23
15	-0.14	0.85	-0.81
16	0.50	0.94	-0.48
17	0.04	0.54	0.00
18	-0.26	1.13	-0.54
19	0.12	0.48	-0.29
20	0.00	1.22	-0.16
21	-0.39	1.41	-0.42
22	-0.45	1.37	-0.44
23	-0.55	1.15	0.10
24	0.19	0.90	-0.41
25	0.11	0.61	-0.06
26	-0.37	0.91	0.10
27	0.31	1.35	0.11
28	0.74	0.81	0.36
29	-0.19	0.40	0.41
30	0.08	1.04	-0.23
31	-0.14	0.92	-0.06
32	-0.14	0.49	-0.24
33	0.33	-0.14	-0.23
34	1.94	-0.11	0.36
35	2.58	-0.13	-0.10
36	1.59	0.11	-0.11
37	1.75	-0.22	-0.40
38	1.28	0.04	0.04
39	1.55	-0.55	-0.39
40	0.86	0.24	-0.51

Item	F1	F2	F3
41	1.41	-0.13	-0.08
42	2.02	0.31	-0.11
43	0.78	-0.05	-0.09
44	2.24	-0.72	0.06
45	0.87	0.76	0.50
46	0.49	0.01	-0.07
47	0.72	0.34	-0.34
48	0.74	0.54	0.13
49	0.39	-0.02	0.10
50	1.85	0.06	-0.14
51	1.50	-0.06	-0.01
52	2.36	0.07	0.06
53	2.45	0.46	-0.15
54	1.44	-0.14	-0.01
55	1.92	0.09	-0.09
56	0.17	0.48	1.76
57	0.01	0.25	1.31
58	0.16	-0.86	1.64
59	-0.09	0.43	1.34
60	-0.10	-0.22	1.45
61	-0.08	0.23	1.64
62	0.24	-0.09	1.35
63	-0.46	-0.02	1.61
64	-0.34	-0.52	1.99
65	-0.08	-0.25	1.17
66	-0.30	0.30	1.54
67	0.09	-0.11	0.78
68	0.01	0.24	0.55
69	0.04	0.28	1.08
70	-0.11	0.42	1.14
71	0.57	-0.18	0.99
72	-0.19	0.56	0.66
73	-0.33	-0.54	1.71
74	-0.28	0.15	1.59
75	-0.25	0.46	1.29
76	-0.29	-0.41	1.46
77	0.62	0.26	1.24
78	0.07	0.13	1.10
79	0.19	0.25	0.93

Table H.8: Estimated loading matrix for the three-factor model based on the marginal likelihood. The results are obtained after applying the oblimin rotation.

	F1	F2	F3
F1	1.00	0.03	-0.16
F2	0.03	1.00	0.06
F3	-0.16	0.06	1.00

Table H.9: The estimated covariance matrix for the latent factors based on the marginal likelihood. The results are obtained after applying the oblimin rotation.

REFERENCES

BAI, J. & NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70**, 191–221.

BAI, Z., CHOI, K. P. & FUJIKOSHI, Y. (2018). Consistency of AIC and BIC in estimating the number of significant components in high-dimensional principal component analysis. *Annals of Statistics* **46**, 1050–1076.

BANDEIRA, A. S. & VAN HANDEL, R. (2016). Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *The Annals of Probability* **44**, 2479–2506.

BHASKAR, S. A. & JAVANMARD, A. (2015). 1-bit matrix completion under exact low-rank constraint. In *2015 49th Annual Conference on Information Sciences and Systems (CISS)*. pp. 1–6.

CAI, T. & ZHOU, W.-X. (2013). A max-norm constrained minimization approach to 1-bit matrix completion. *The Journal of Machine Learning Research* **14**, 3619–3647.

CATTELL, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research* **1**, 245–276.

CHALMERS, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software* **48**, 1–29.

CHEN, Y., LI, X. & ZHANG, S. (2019). Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. *psychometrika* **84**, 124–146.

CHEN, Y., LI, X. & ZHANG, S. (2020). Structured latent factor analysis for large-scale data: Identifiability, estimability, and their implications. *Journal of the American Statistical Association* **115**, 1756–1770.

DAVENPORT, M. A., PLAN, Y., VAN DEN BERG, E. & WOOTTERS, M. (2014). 1-bit matrix completion. *Information and Inference* **3**, 189–223.

EYSENCK, S. B., EYSENCK, H. J. & BARRETT, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences* **6**, 21–29.

GILBERT, E. N. (1952). A comparison of signalling alphabets. *The Bell system technical journal* **31**, 504–522.

HORN, R. A. (1995). Norm bounds for hadamard products and an arithmetic-geometric mean inequality for unitarily invariant norms. *Linear Algebra and Its Applications* **223**, 355–361.

HUBER, P., RONCHETTI, E. & VICTORIA-FESER, M.-P. (2004). Estimation of generalized linear latent variable models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66**, 893–908.

NI, R. & GU, Q. (2016). Optimal statistical and computational rates for one bit matrix completion. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. pp. 426–434.

RECKASE, M. (2009). *Multidimensional item response theory*. New York, NY: Springer.

SHAO, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica* **7**, 221–242.

TSYBAKOV, A. B. (2008). *Introduction to nonparametric estimation*. New York, NY: Springer.

ZHANG, H., CHEN, Y. & LI, X. (2020). A note on exploratory item factor analysis by singular value decomposition. *Psychometrika* **85**, 358–372.