Supplement to

"A Matrix–free Likelihood Method for Exploratory Factor Analysis of High-dimensional Gaussian Data"

Fan Dai, Somak Dutta and Ranjan Maitra

S1 Supplementary materials for Methodology

S1.1 The EM algorithm for factor analysis on Gaussian data

The complete data log-likelihood function is

$$\begin{aligned}\ell_C(\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\Psi}) &= c - \frac{n}{2}\log\det\boldsymbol{\Psi} - \frac{1}{2}\sum_{i=1}^{n}\{(\mathbf{Y}_i - \boldsymbol{\mu} - \boldsymbol{\Lambda}\mathbf{Z}_i)^\top \boldsymbol{\Psi}^{-1}(\mathbf{Y}_i - \boldsymbol{\mu} - \boldsymbol{\Lambda}\mathbf{Z}_i)\} \\ &= c - \frac{n}{2}\log\det\boldsymbol{\Psi} - \frac{1}{2}\mathrm{Tr}\,\{\boldsymbol{\Psi}^{-1}\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\mu})(\mathbf{Y}_i - \boldsymbol{\mu})^\top - 2\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\sum_{i=1}^{n}\mathbf{Z}_i(\mathbf{Y}_i - \boldsymbol{\mu})^\top\} \\ &\quad - \frac{1}{2}\mathrm{Tr}\,\{\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\sum_{i=1}^{n}\mathbf{Z}_i\mathbf{Z}_i^\top\},\end{aligned} \tag{S1}$$

where c is a constant that does not depend on the parameters.

S1.1.1 E-Step computations

Since the ML estimate of $\boldsymbol{\mu}$ is $\bar{\mathbf{Y}}$, at the current estimates $\boldsymbol{\Lambda}_t$ and $\boldsymbol{\Psi}_t$, the expected complete log-likelihood or so called Q function is given by

$$\begin{aligned}\mathrm{Q}(\boldsymbol{\Lambda}_{t+1}, \boldsymbol{\Psi}_{t+1}|\bar{\mathbf{Y}}, \boldsymbol{\Lambda}_t, \boldsymbol{\Psi}_t) =& \mathrm{E}[\ell_C(\boldsymbol{\Lambda}, \boldsymbol{\Psi}|\mathbf{Y}, \boldsymbol{\Lambda}_t, \boldsymbol{\Psi}_t] \\ =& -\frac{n}{2}\log\det\boldsymbol{\Psi} - \frac{n}{2}\mathrm{Tr}\,\boldsymbol{\Psi}^{-1}\mathbf{S} - \mathrm{Tr}\,\{\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\sum_{i=1}^{n}\mathrm{E}[\mathbf{Z}_i|\mathbf{Y}, \boldsymbol{\Lambda}_t, \boldsymbol{\Psi}_t](\mathbf{Y}_i - \bar{\mathbf{Y}})^\top\} \\ & + \frac{1}{2}\mathrm{Tr}\,\{\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[\mathbf{Z}_i\mathbf{Z}_i^\top|\mathbf{Y}, \boldsymbol{\Lambda}_t, \boldsymbol{\Psi}_t]\}.\end{aligned} \tag{S2}$$

Since $\mathbf{Z}_i|\mathbf{Y}, \boldsymbol{\Lambda}_t, \boldsymbol{\Psi}_t \sim \mathcal{N}_q(\boldsymbol{\Lambda}_t^\top\boldsymbol{\Sigma}_t^{-1}(\mathbf{Y}_i - \bar{\mathbf{Y}}), (\mathbf{I}_q + \boldsymbol{\Lambda}_t^\top\boldsymbol{\Psi}_t^{-1}\boldsymbol{\Lambda}_t)^{-1})$. Then,

$$\mathrm{E}[\mathbf{Z}_i|\mathbf{Y}, \boldsymbol{\Lambda}_t, \boldsymbol{\Psi}_t] = \boldsymbol{\Lambda}_t^\top\boldsymbol{\Sigma}_t^{-1}(\mathbf{Y}_i - \bar{\mathbf{Y}}) \tag{S3}$$

and

$$\begin{aligned}\mathrm{E}[\mathbf{Z}_i\mathbf{Z}_i^\top|\mathbf{Y},\boldsymbol{\Lambda}_t,\boldsymbol{\Psi}_t] =&\mathrm{Var}[\mathbf{Z}_i|\mathbf{Y},\boldsymbol{\Lambda}_t,\boldsymbol{\Psi}_t]+\mathrm{E}[\mathbf{Z}_i|\mathbf{Y},\boldsymbol{\Lambda}_t,\boldsymbol{\Psi}_t]\mathrm{E}[\mathbf{Z}_i^\top|\mathbf{Y}_i,\bar{\mathbf{Y}},\boldsymbol{\Lambda}_t,\boldsymbol{\Psi}_t]\\ =&(\mathbf{I}_q+\boldsymbol{\Lambda}_t^\top\boldsymbol{\Psi}_t^{-1}\boldsymbol{\Lambda}_t)^{-1}+\boldsymbol{\Lambda}_t^\top\boldsymbol{\Sigma}_t^{-1}(\mathbf{Y}_i-\bar{\mathbf{Y}})(\mathbf{Y}_i-\bar{\mathbf{Y}})^\top\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\Lambda}_t.\end{aligned} \tag{S4}$$

S1.1.2 M-Step computations

The parameters $\boldsymbol{\Lambda}_{t+1}$ and $\boldsymbol{\Psi}_{t+1}$ are obtained by maximizing $\mathrm{Q}(\boldsymbol{\Lambda}_{t+1},\boldsymbol{\Psi}_{t+1}|\bar{\mathbf{Y}},\boldsymbol{\Lambda}_t,\boldsymbol{\Psi}_t)$ following equation S2. Specifically, given $\mathbf{Y}$, $\boldsymbol{\Lambda}_t$ and $\boldsymbol{\Psi}_t$, the maximizer $\boldsymbol{\Lambda}_{t+1}$ is given by

$$\begin{aligned}\hat{\boldsymbol{\Lambda}}_{t+1} =&(\frac{1}{n}\sum_{i=1}^n(\mathbf{Y}_i-\bar{\mathbf{Y}})\mathrm{E}[\mathbf{Z}_i^\top|\mathbf{Y},\boldsymbol{\Lambda}_t,\boldsymbol{\Psi}_t])(\frac{1}{n}\sum_{i=1}^n\mathrm{E}[\mathbf{Z}_i\mathbf{Z}_i^\top|\mathbf{Y},\boldsymbol{\Lambda}_t,\boldsymbol{\Psi}_t])^{-1}\\ =&\mathbf{S}\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\Lambda}_t((\mathbf{I}_q+\boldsymbol{\Lambda}_t^\top\boldsymbol{\Psi}_t^{-1}\boldsymbol{\Lambda}_t)^{-1}+\boldsymbol{\Lambda}_t^\top\boldsymbol{\Sigma}_t^{-1}\mathbf{S}\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\Lambda}_t)^{-1}\end{aligned} \tag{S5}$$

where $\boldsymbol{\Sigma}_t=\boldsymbol{\Lambda}_t\boldsymbol{\Lambda}_t^\top+\boldsymbol{\Psi}_t$. By Woodbury matrix identity (Henderson and Searle, 1981), $\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\Lambda}_t=\boldsymbol{\Psi}_t^{-1}\boldsymbol{\Lambda}_t(\mathbf{I}_q+\boldsymbol{\Lambda}_t^\top\boldsymbol{\Psi}_t^{-1}\boldsymbol{\Lambda}_t)^{-1}$, so S5 can be simplified as

$$\begin{aligned}\hat{\boldsymbol{\Lambda}}_{t+1} =&\mathbf{S}\boldsymbol{\Psi}_t^{-1}\boldsymbol{\Lambda}_t(\mathbf{I}_q+\boldsymbol{\Lambda}_t^\top\boldsymbol{\Psi}_t^{-1}\boldsymbol{\Lambda}_t)^{-1}((\mathbf{I}_q+\boldsymbol{\Lambda}_t^\top\boldsymbol{\Psi}_t^{-1}\boldsymbol{\Lambda}_t)^{-1}+\boldsymbol{\Lambda}_t^\top\boldsymbol{\Sigma}_t^{-1}\mathbf{S}\boldsymbol{\Psi}_t^{-1}\boldsymbol{\Lambda}_t(\mathbf{I}_q+\boldsymbol{\Lambda}_t^\top\boldsymbol{\Psi}_t^{-1}\boldsymbol{\Lambda}_t)^{-1})^{-1}\\ =&\mathbf{S}\boldsymbol{\Psi}_t^{-1}\boldsymbol{\Lambda}_t(\mathbf{I}_q+\boldsymbol{\Lambda}_t^\top\boldsymbol{\Sigma}_t^{-1}\mathbf{S}\boldsymbol{\Psi}_t^{-1}\boldsymbol{\Lambda}_t)^{-1}.\end{aligned} \tag{S6}$$

Next, given $\mathbf{Y}$, $\boldsymbol{\Lambda}_t$, $\boldsymbol{\Psi}_t$ and $\hat{\boldsymbol{\Lambda}}_{t+1}$, the ML estimate of $\boldsymbol{\Psi}_{t+1}$ is given by

$$\begin{aligned}\hat{\boldsymbol{\Psi}}_{t+1} = \mathrm{diag}\Big(&\mathbf{S}-\frac{2}{n}\sum_{i=1}^n(\mathbf{Y}_i-\bar{\mathbf{Y}})\mathrm{E}[\mathbf{Z}_i^\top|\mathbf{Y},\boldsymbol{\Lambda}_t,\boldsymbol{\Psi}_t]\hat{\boldsymbol{\Lambda}}_{t+1}^\top\\ &+\hat{\boldsymbol{\Lambda}}_{t+1}\frac{1}{n}\sum_{i=1}^n\mathrm{E}[\mathbf{Z}_i\mathbf{Z}_i^\top|\mathbf{Y},\boldsymbol{\Lambda}_t,\boldsymbol{\Psi}_t])^{-1}\hat{\boldsymbol{\Lambda}}_{t+1}^\top\Big).\end{aligned} \tag{S7}$$

Substitute with S5, we get

$$\begin{aligned}\hat{\boldsymbol{\Psi}}_{t+1} =&\mathrm{diag}\Big(\mathbf{S}-\frac{2}{n}\sum_{i=1}^n(\mathbf{Y}_i-\bar{\mathbf{Y}})\mathrm{E}[\mathbf{Z}_i^\top|\mathbf{Y},\boldsymbol{\Lambda}_t,\boldsymbol{\Psi}_t]\hat{\boldsymbol{\Lambda}}_{t+1}^\top\\ &+\frac{1}{n}\sum_{i=1}^n(\mathbf{Y}_i-\bar{\mathbf{Y}})\mathrm{E}[\mathbf{Z}_i^\top|\mathbf{Y},\boldsymbol{\Lambda}_t,\boldsymbol{\Psi}_t]\hat{\boldsymbol{\Lambda}}_{t+1}^\top\Big)\\ =&\mathrm{diag}\Big(\mathbf{S}-2\mathbf{S}\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\Lambda}_t\hat{\boldsymbol{\Lambda}}_{t+1}^\top\Big).\end{aligned} \tag{S8}$$

S1.2 Proof of Lemma 1

From equation 1, the ML estimates of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ are obtained by solving the score equations

$$\begin{cases} \mathbf{\Lambda}(\mathbf{I}_q + \mathbf{\Lambda}^\top \mathbf{\Psi}^{-1}\mathbf{\Lambda}) = \mathbf{S}\mathbf{\Psi}^{-1}\mathbf{\Lambda} \\ \mathbf{\Psi} = \mathrm{diag}(\mathbf{S} - \mathbf{\Lambda}\mathbf{\Lambda}^\top) \end{cases} \tag{S9}$$

From $\mathbf{\Lambda}(\mathbf{I}_q + \mathbf{\Lambda}^\top \mathbf{\Psi}^{-1}\mathbf{\Lambda}) = \mathbf{S}\mathbf{\Psi}^{-1}\mathbf{\Lambda}$, we have

$$\mathbf{\Psi}^{-1/2}\mathbf{\Lambda}(\mathbf{I}_q + (\mathbf{\Psi}^{-1/2}\mathbf{\Lambda})^\top \mathbf{\Psi}^{-1/2}\mathbf{\Lambda}) = \mathbf{\Psi}^{-1/2}\mathbf{S}\mathbf{\Psi}^{-1/2}\mathbf{\Psi}^{-1/2}\mathbf{\Lambda}. \tag{S10}$$

Suppose that $\mathbf{\Psi}^{-1/2}\mathbf{S}\mathbf{\Psi}^{-1/2} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$ and that the diagonal elements in $\mathbf{D}$ are in decreasing order with $\theta_1 \geq \theta_2 \geq, \cdots, \geq \theta_p$. Let $\mathbf{D} = \begin{bmatrix} \mathbf{D}_q & 0 \\ 0 & \mathbf{D}_m \end{bmatrix}$ with $m = p - q$ and $\mathbf{D}_q$ containing the largest q eigenvalues $\theta_1 \geq \theta_2 \geq, \cdots, \geq \theta_q$. The corresponding q eigenvectors forms columns of matrix $\mathbf{V}_q$ so that $\mathbf{V} = [\mathbf{V}_q, \mathbf{V}_m]$. Then, if $\mathbf{D}_q > \mathbf{I}_q$, S10 shows that

$$\mathbf{\Lambda} = \mathbf{\Psi}^{1/2}\mathbf{V}_q(\mathbf{D}_q - \mathbf{I}_q)^{1/2}. \tag{S11}$$

The square roots of $\theta_1, \cdots, \theta_q$ are the q largest singular values of $n^{-1/2}(\mathbf{Y} - \mathbf{1}\bar{\mathbf{Y}}^\top)\mathbf{\Psi}^{-1/2}$ and columns in $\mathbf{V}_q$ are then the corresponding q right-singular vectors. Hence, conditional on $\mathbf{\Psi}$, $\mathbf{\Lambda}$ is maximized at $\hat{\mathbf{\Lambda}} = \mathbf{\Psi}^{1/2}\mathbf{V}_q\mathbf{\Delta}$, where $\mathbf{\Delta}$ is a diagonal matrix with elements $\max(\theta_i - 1, 0)^{1/2}, i = 1, \cdots, q$.

From the construction of $\mathbf{V}_q$ and $\mathbf{V}_m$, we have $\mathbf{V}_q^\top\mathbf{V}_q = \mathbf{I}_q$, $\mathbf{V}_m^\top\mathbf{V}_m = \mathbf{I}_m$, $\mathbf{V}_q\mathbf{V}_q^\top + \mathbf{V}_m\mathbf{V}_m^\top = \mathbf{I}_p$, $\mathbf{V}_q^\top\mathbf{V}_m = \mathbf{0}$ and hence, $(\mathbf{V}_q\mathbf{D}_q\mathbf{V}_q^\top + \mathbf{V}_m\mathbf{V}_m^\top)(\mathbf{V}_q\mathbf{D}_q^{-1}\mathbf{V}_q^\top + \mathbf{V}_m\mathbf{V}_m^\top) = \mathbf{I}_p$.

Let $\mathbf{A} = \mathbf{V}_q\mathbf{\Delta}^2\mathbf{V}_q^\top$. Then $\mathbf{A}\mathbf{A} = \mathbf{V}_q\mathbf{\Delta}^4\mathbf{V}_q^\top$ and

$$\begin{aligned} |\mathbf{A} + \mathbf{I}_p| &= |(\mathbf{A} + \mathbf{I}_p)\mathbf{A}|/|\mathbf{A}| \\ &= |\mathbf{V}_q(\mathbf{\Delta}^4 + \mathbf{\Delta}^2)\mathbf{V}_q^\top|/|\mathbf{V}_q\mathbf{\Delta}^2\mathbf{V}_q^\top| \\ &= |\mathbf{\Delta}^2 + \mathbf{I}_q| = \prod_{j=1}^{q}\theta_j \end{aligned} \tag{S12}$$

and

$$\begin{aligned} (\mathbf{A} + \mathbf{I}_p)^{-1} &= (\mathbf{V}_q\mathbf{\Delta}^2\mathbf{V}_q^\top + \mathbf{V}_q\mathbf{V}_q^\top + \mathbf{V}_m\mathbf{V}_m^\top)^{-1} \\ &= (\mathbf{V}_q(\mathbf{\Delta}^2 + \mathbf{I}_q)\mathbf{V}_q^\top + \mathbf{V}_m\mathbf{V}_m^\top)^{-1} \\ &= (\mathbf{V}_q\mathbf{D}_q\mathbf{V}_q^\top + \mathbf{V}_m\mathbf{V}_m^\top)^{-1} \\ &= \mathbf{V}_q\mathbf{D}_q^{-1}\mathbf{V}_q^\top + \mathbf{V}_m\mathbf{V}_m^\top. \end{aligned} \tag{S13}$$

Based on S12 and S13 and equation 2, the profile log-likelihood is given by

$$\begin{aligned}
\ell_p(\boldsymbol{\Psi}) &= c - \frac{n}{2}\log|\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}^\top + \boldsymbol{\Psi}| - \frac{n}{2}\mathrm{Tr}\,(\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}^\top + \boldsymbol{\Psi})^{-1}\mathbf{S} \\
&= c - \frac{n}{2}\Big\{\log|\boldsymbol{\Psi}^{1/2}(\mathbf{V}_q\boldsymbol{\Delta}^2\mathbf{V}_q^\top + \mathbf{I}_p)\boldsymbol{\Psi}^{1/2}| + \mathrm{Tr}\;\boldsymbol{\Psi}^{1/2}(\mathbf{V}_q\boldsymbol{\Delta}^2\mathbf{V}_q^\top + \mathbf{I}_p)\boldsymbol{\Psi}^{1/2})^{-1}\mathbf{S}\Big\} \\
&= c - \frac{n}{2}\Big\{\log\det\boldsymbol{\Psi} + \log|\mathbf{V}_q\boldsymbol{\Delta}^2\mathbf{V}_q^\top + \mathbf{I}_p| + \mathrm{Tr}\,(\mathbf{V}_q\mathbf{D}_q^{-1}\mathbf{V}_q^\top + \mathbf{V}_m\mathbf{V}_m^\top)\boldsymbol{\Psi}^{-1/2}\mathbf{S}\boldsymbol{\Psi}^{-1/2}\Big\} \\
&= c - \frac{n}{2}\Big\{\log\det\boldsymbol{\Psi} + \sum_{j=1}^{q}\log\theta_j + \mathrm{Tr}\,(\mathbf{D}_q^{-1}\mathbf{V}_q^\top\mathbf{V}\mathbf{D}\mathbf{V}^\top\mathbf{V}_q + \mathrm{Tr}\;\mathbf{V}_m^\top\mathbf{V}\mathbf{D}\mathbf{V}^\top\mathbf{V}_m\Big\} \\
&= c - \frac{n}{2}\Big\{\log\det\boldsymbol{\Psi} + \sum_{j=1}^{q}\log\theta_j + \mathrm{Tr}\;\mathbf{D}_q^{-1}\mathbf{D}_q + \mathrm{Tr}\;\mathbf{D}_m\Big\} \\
&= c - \frac{n}{2}\Big\{\log\det\boldsymbol{\Psi} + \sum_{j=1}^{q}\log\theta_j + q + \mathrm{Tr}\;\boldsymbol{\Psi}^{-1}\mathbf{S} - \sum_{j=1}^{q}\theta_j\Big\}.
\end{aligned} \tag{S14}$$

S2 Additional results for simulation studies

S2.1 Average CPU time

Table S1: Average CPU time (in seconds) of FAD and EM applied with 1-6 factors for randomly simulated datasets where true $q = 3$.

		1	2	3	4	5	6
$(n, p, q) = (10^2, 10^3, 3)$	FAD	0.101	0.092	0.096	0.116	0.122	0.128
	EM	2.494	2.519	3.076	2.012	2.075	2.162
$(n, p, q) = (15^2, 15^3, 3)$	FAD	0.639	0.514	0.486	0.841	0.966	1.025
	EM	24.798	22.885	27.906	16.822	16.630	15.722
$(n, p, q) = (20^2, 20^3, 3)$	FAD	2.933	2.658	2.580	7.135	7.863	8.590
	EM	57.052	57.527	81.463	49.689	48.508	49.504

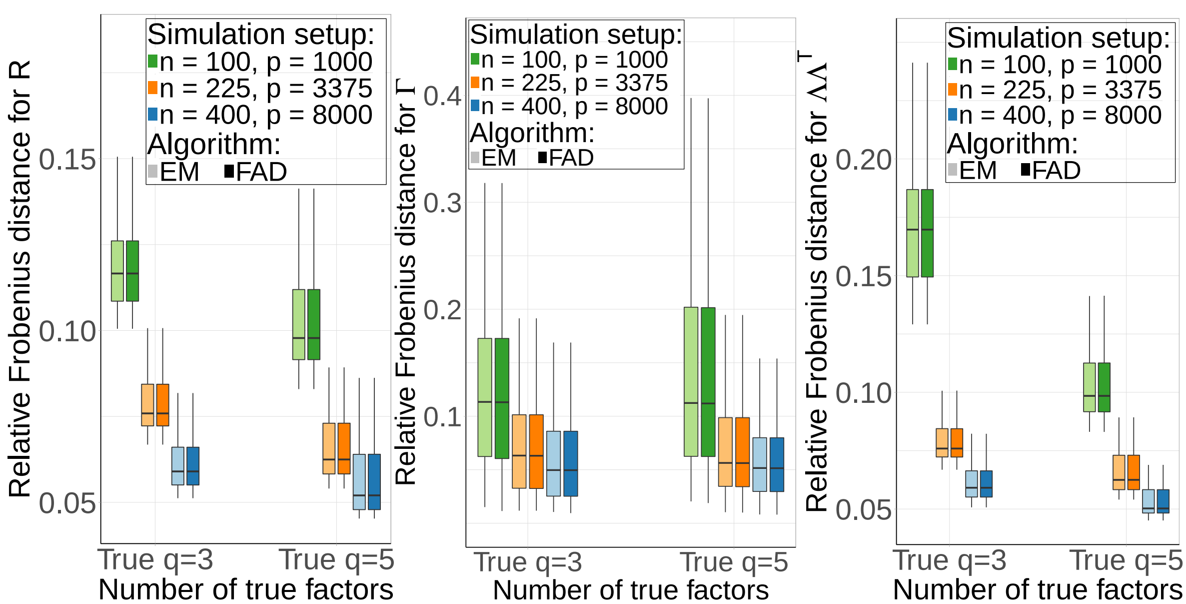
Table S2: Average CPU time (in seconds) of FAD and EM applied with 1-10 factors for randomly simulated datasets where true $q = 5$.

		1	2	3	4	5	6	7	8	9	10
$(n, p, q) = (10^2, 10^3, 5)$	FAD	0.102	0.095	0.096	0.096	0.094	0.119	0.124	0.128	0.134	0.137
	EM	2.290	2.539	2.808	2.800	2.985	2.301	2.327	2.097	2.166	2.196
$(n, p, q) = (15^2, 15^3, 5)$	FAD	0.667	0.513	0.501	0.507	0.497	0.828	0.919	1.039	1.108	1.143
	EM	22.545	21.066	22.300	22.197	26.300	16.544	15.796	14.767	14.292	14.789
$(n, p, q) = (20^2, 20^3, 5)$	FAD	2.937	2.687	2.553	2.583	2.590	7.114	8.167	9.238	10.426	11.157
	EM	47.200	47.333	49.867	49.956	71.308	47.469	47.119	43.828	45.304	44.141

Table S3: Average CPU time (in seconds) of FAD and EM applied for data-driven models.

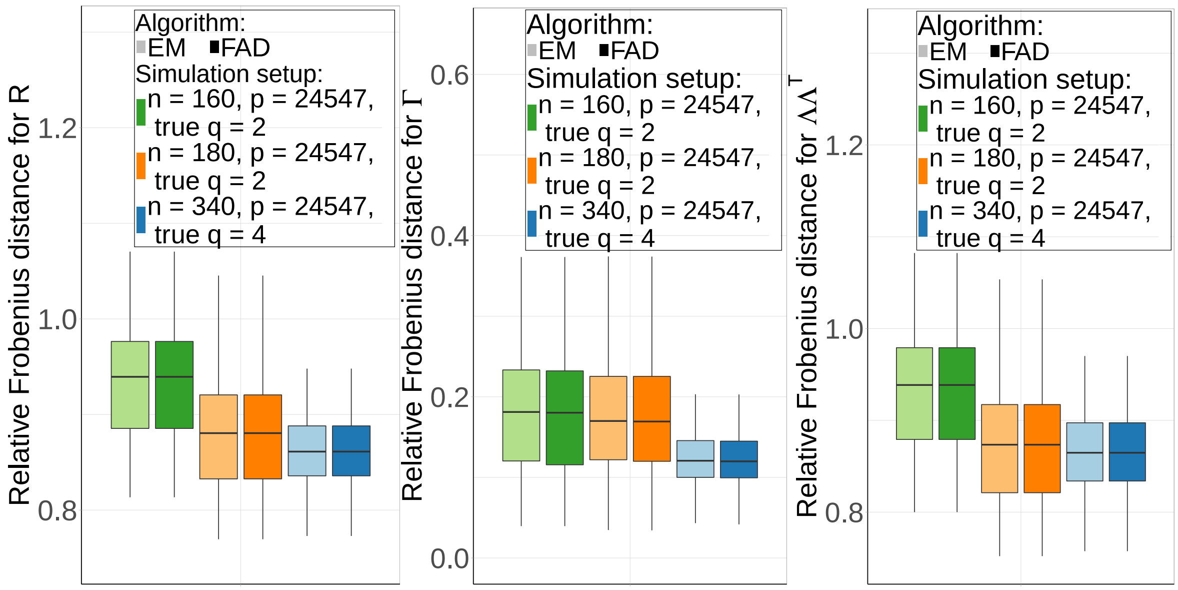
		1	2	3	4	5	6	7	8
$(n, p, q) = (160, 24547, 2)$	FAD	5.007	4.222	10.835	13.636	–	–	–	–
	EM	253.021	304.909	311.916	303.712	–	–	–	–
$(n, p, q) = (180, 24547, 2)$	FAD	4.927	4.104	10.411	12.058	–	–	–	–
	EM	287.824	345.504	331.919	314.723	–	–	–	–
$(n, p, q) = (340, 24547, 4)$	FAD	6.645	7.121	7.449	6.688	22.294	26.575	31.109	34.208
	EM	648.759	734.226	745.902	735.263	767.010	789.614	802.502	748.395

S2.2 Estimation errors in parameters



(a) (b) (c)

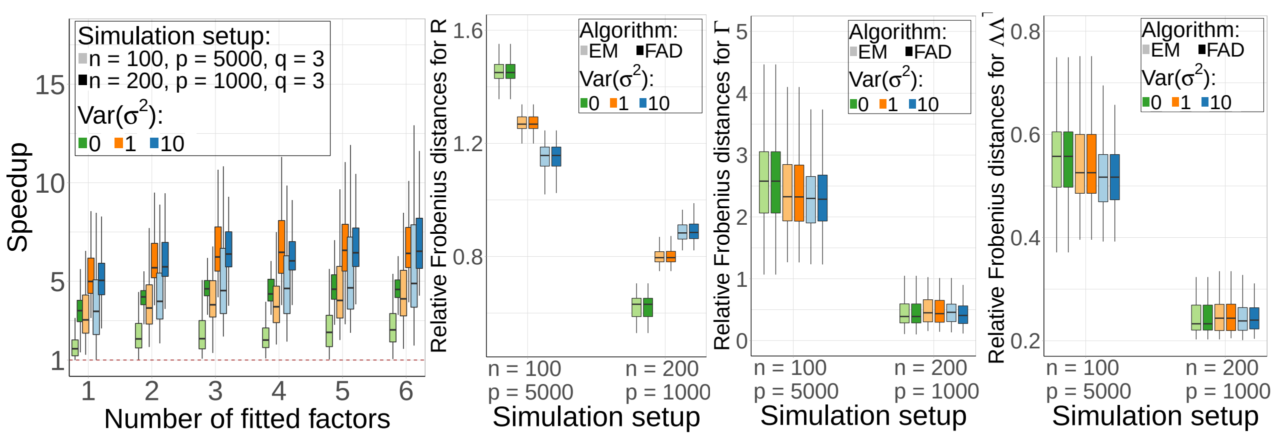
Figure S1: Relative Frobenius errors of FAD and EM for (a) correlation matrix $\mathbf{R}$, (b) signal matrix $\mathbf{\Gamma}$ and (c) $\mathbf{\Lambda}\mathbf{\Lambda}^\top$ on randomly simulated cases, with lighter ones for EM and darker ones for FAD.



(a) (b) (c)

Figure S2: Relative Frobenius errors of FAD and EM for (a) correlation matrix $\mathbf{R}$, (b) signal matrix $\mathbf{\Gamma}$ and (c) $\mathbf{\Lambda}\mathbf{\Lambda}^\top$ on data-driven cases, with lighter ones for EM and darker ones for FAD.

S2.3 Performance of FAD compared to EM for high-noise models



(a) (b) (c) (d)

Figure S3: Relative speed of FAD to EM on dataset with high noise (a). Relative Frobenius errors of FAD and EM for (b) correlation matrix $\mathbf{R}$, (c) signal matrix $\mathbf{\Gamma}$ and (d) $\mathbf{\Lambda}\mathbf{\Lambda}^\top$.

References

Henderson, H. V. and Searle, S. R. (1981), "On deriving the inverse of a sum of matrices," *SIAM Review*, **23**, 53–60.