Expandable Factor Analysis

Sanvesh Srivastava
Postdoctoral Researcher
Duke University & SAMSI

joint work with

David B. Dunson (Duke University)
Barbara E. Engelhardt (Princeton University)

February 12, 2015
Indiana University
Big Data

parameters (p) \[ p \rightarrow \infty \]

samples (n) \[ n \rightarrow \infty \]  

“big p”

life sciences lab

parameters (p)

n \rightarrow \infty

“big n”

“big data”

cancer genome atlas

Google
Multivariate Data & Factor Analysis

Approaches for Factor Analysis

Expandable Factor Analysis (MAP Estimation)

Expandable Factor Analysis ("Model" Averaging)

Applications
Examples

psychometrics
Examples

psychometrics

battery of tests, evaluations
Examples

Psychometrics

- battery of tests, evaluations
- observations: test scores, evaluation scores

Social Sciences

- questionnaires, surveys
- observations: responses measuring evolution of a society
- factors: group memberships, sub-networks, political affiliations

Genomics & Neuroscience

- biological or neurological measurements
- observations: genetic activity, brain signals
- factors: pathway, brain sub-networks
Examples

- **psychometrics**
  - battery of tests, evaluations
  - observations: test scores, evaluation scores
  - factors: mental abilities, personality traits, skills

- **social sciences**
  - questionnaires, surveys
  - observations: responses measuring evolution of a society
  - factors: group memberships, sub-networks, political affiliations

- **genomics & neuroscience**
  - biological or neurological measurements
  - observations: genetic activity, brain signals
  - factors: pathway, brain sub-networks
Examples

psychometrics
- battery of tests, evaluations
- observations: test scores, evaluation scores
- factors: mental abilities, personality traits, skills

social sciences
Examples

Psychometrics
- battery of tests, evaluations
- observations: test scores, evaluation scores
- factors: mental abilities, personality traits, skills

Social Sciences
- questionnaires, surveys
Examples

psychometrics

- battery of tests, evaluations
- observations: test scores, evaluation scores
- factors: mental abilities, personality traits, skills

social sciences

- questionnaires, surveys
- observations: responses measuring evolution of a society

Multivariate Data & Factor Analysis 3
Examples

psychometrics
- battery of tests, evaluations
- observations: test scores, evaluation scores
- factors: mental abilities, personality traits, skills

social sciences
- questionnaires, surveys
- observations: responses measuring evolution of a society
- factors: group memberships, sub-networks, political affiliations
Examples

- **psychometrics**
  - battery of tests, evaluations
  - **observations**: test scores, evaluation scores
  - **factors**: mental abilities, personality traits, skills

- **social sciences**
  - questionnaires, surveys
  - **observations**: responses measuring evolution of a society
  - **factors**: group memberships, sub-networks, political affiliations

- **genomics & neuroscience**
  - biological or neurological measurements
  - **observations**: genetic activity, brain signals
  - **factors**: pathway, brain sub-networks
Examples

psychometrics
- battery of tests, evaluations
- observations: test scores, evaluation scores
- factors: mental abilities, personality traits, skills

social sciences
- questionnaires, surveys
- observations: responses measuring evolution of a society
- factors: group memberships, sub-networks, political affiliations

genomics & neuroscience
- biological or neurological measurements
Examples

- **psychometrics**
  - battery of tests, evaluations
  - **observations**: test scores, evaluation scores
  - **factors**: mental abilities, personality traits, skills

- **social sciences**
  - questionnaires, surveys
  - **observations**: responses measuring evolution of a society
  - **factors**: group memberships, sub-networks, political affiliations

- **genomics & neuroscience**
  - biological or neurological measurements
  - **observations**: genetic activity, brain signals
Examples

psychometrics
- battery of tests, evaluations
- observations: test scores, evaluation scores
- factors: mental abilities, personality traits, skills

social sciences
- questionnaires, surveys
- observations: responses measuring evolution of a society
- factors: group memberships, sub-networks, political affiliations

genomics & neuroscience
- biological or neurological measurements
- observations: genetic activity, brain signals
- factors: pathway, brain sub-networks
Behavioral Neuroscience
Data Analysis

**hypothesis:** factors “explain” observations
Data Analysis

- **hypothesis:** factors “explain” observations

- **model:** observations depend *linearly* on the factors
Data Analysis

- **hypothesis:** factors “explain” observations
- **model:** observations depend linearly on the factors
- **goal:** \( \text{observations}(n) = \text{matrix} \times \text{factors}(n) + \text{residuals}(n) \)
Brief History of Factor Analysis

Spearman invented factor analysis for educational psychology applications in 1904.
Brief History of Factor Analysis

- Spearman invented factor analysis for educational psychology applications in 1904
- Anderson & Rubin added the statistical flavor in 1956
Brief History of Factor Analysis

- Spearman invented factor analysis for educational psychology applications in 1904
- Anderson & Rubin added the statistical flavor in 1956
- Rubin & Thayer (1982) used Expectation-Maximization for maximum likelihood parameter estimation
Brief History of Factor Analysis

† Spearman invented factor analysis for educational psychology applications in 1904

† Anderson & Rubin added the statistical flavor in 1956

† Rubin & Thayer (1982) used Expectation-Maximization for maximum likelihood parameter estimation

† extensively used in multivariate statistics [Muthén (2002); Bartholomew et al. (2011)]
Brief History of Factor Analysis

- Spearman invented factor analysis for educational psychology applications in 1904
- Anderson & Rubin added the statistical flavor in 1956
- Rubin & Thayer (1982) used Expectation-Maximization for maximum likelihood parameter estimation
- extensively used in multivariate statistics [Muthén (2002); Bartholomew et al. (2011)]
- dimension reduction
Brief History of Factor Analysis

❖ Spearman invented factor analysis for educational psychology applications in 1904

❖ Anderson & Rubin added the statistical flavor in 1956

❖ Rubin & Thayer (1982) used Expectation-Maximization for maximum likelihood parameter estimation

❖ extensively used in multivariate statistics [Muthén (2002); Bartholomew et al. (2011)]

❖ dimension reduction

❖ covariance matrices
Modern Relevance of Factor Analysis

- **pre-2000 period** [Muthén (2002)]
  - typical applications in educational psychology, epidemiology
  - dimensions $\ll$ samples ($P \ll N$)
  - factors are *known* apriori
  - parameter estimation using maximum likelihood

- **post-2000 period** [Bai, West; Dunson; Ghahramani, Jordan; Agarwal, Anandkumar]
  - typical applications in genomics and neuroscience
  - dimensions $\gg$ samples ($P \gg N$)
  - factors are unknown apriori
  - regularized parameter estimation using “Bet on Sparsity” principle
Modern Relevance of Factor Analysis

**pre-2000 period** [Muthén (2002)]
- typical applications in educational psychology, epidemiology
- dimensions $\ll$ samples ($P \ll N$)
- factors are known apriori
- parameter estimation using maximum likelihood

**post-2000 period** [Bai, West; Dunson; Ghahramani, Jordan; Agarwal, Anandkumar, Kakade]
- typical applications in genomics and neuroscience
- dimensions $\gg$ samples ($P \gg N$)
- factors are unknown apriori
- regularized parameter estimation using “Bet on Sparsity” principle
Multivariate Data & Factor Analysis

Approaches for Factor Analysis

Expandable Factor Analysis (MAP Estimation)

Expandable Factor Analysis (“Model” Averaging)

Applications
Generative Model

observations: \( y_1, \ldots, y_N \in \mathbb{R}^P \)
Generative Model

- observations: \( y_1, \ldots, y_N \in \mathbb{R}^P \)
- factors: \( z_1, \ldots, z_N \in \mathbb{R}^K \)
Generative Model

- **observations:** $y_1, \ldots, y_N \in \mathbb{R}^P$
- **factors:** $z_1, \ldots, z_N \in \mathbb{R}^K$
- **errors:** $e_1, \ldots, e_N \in \mathbb{R}^P$ and $\mathbb{E}[e_n] = 0$
Generative Model

- **observations:** \( y_1, \ldots, y_N \in \mathbb{R}^P \)
- **factors:** \( z_1, \ldots, z_N \in \mathbb{R}^K \)
- **errors:** \( e_1, \ldots, e_N \in \mathbb{R}^P \) and \( \mathbb{E}[e_n] = 0 \)
- **model:** \( y_9 = \Lambda z_9 + e_9 \)
Generative Model

- **observations**: \( y_1, \ldots, y_N \in \mathbb{R}^P \)
- **factors**: \( z_1, \ldots, z_N \in \mathbb{R}^K \)
- **errors**: \( e_1, \ldots, e_N \in \mathbb{R}^P \) and \( \mathbb{E}[e_n] = 0 \)
- **model**: \( y_9 = \Lambda z_9 + e_9 \)
- \( (y_9)_5 = z_9^T \lambda_5 + (e_9)_5 \)
Generative Model

- **observations**: $y_1, \ldots, y_N \in \mathbb{R}^P$
- **factors**: $z_1, \ldots, z_N \in \mathbb{R}^K$
- **errors**: $e_1, \ldots, e_N \in \mathbb{R}^P$ and $\mathbb{E}[e_n] = 0$
- **model**: $y_9 = \Lambda z_9 + e_9$
- **(y_9)_5 = z_9^T \lambda_5 + (e_9)_5**
- **parameters**: $\Lambda \in \mathbb{R}^{P \times K}$ and $\mathbb{V}[e_n] = \text{diag}(\sigma^2_{11}, \ldots, \sigma^2_{PP}) \equiv \Sigma$

Approaches for Factor Analysis
Generative Model

- **observations:** \( y_1, \ldots, y_N \in \mathbb{R}^P \)
- **factors:** \( z_1, \ldots, z_N \in \mathbb{R}^K \)
- **errors:** \( e_1, \ldots, e_N \in \mathbb{R}^P \) and \( \mathbb{E}[e_n] = 0 \)
- **model:** \( y_9 = \Lambda z_9 + e_9 \)
- \( (y_9)_5 = z_9^T \lambda_5 + (e_9)_5 \)
- **parameters:** \( \Lambda \in \mathbb{R}^{P \times K} \) and \( \mathbb{V}[e_n] = \text{diag}(\sigma^2_{11}, \ldots, \sigma^2_{PP}) \equiv \Sigma \)
- **assumptions:** \( \mathbb{E}[z_n] = 0, \mathbb{V}[z_n] = I_K \), and normality

Approaches for Factor Analysis
Two Interpretations of Factor Analysis

geometry

- $\lambda_1, \ldots, \lambda_K \in \mathbb{R}^P$ span $K$-dimensional space
- $z_1, \ldots, z_N$ are the coordinates of $y_1, \ldots, y_N$ in factor space
- $[z_n, 0_{P-K \times 1}]$ is the $P$-dimensional embedding of $y_n$
- pathway coordinate axis; subnetwork coordinate axis
Two Interpretations of Factor Analysis

geometry

\[ \lambda_1, \ldots, \lambda_K \in \mathbb{R}^P \] span \( K \)-dimensional space

\[ z_1, \ldots, z_N \] are the coordinates of \( y_1, \ldots, y_N \) in factor space

\[ [z_n, \mathbf{0}_{P-K \times 1}] \] is the \( P \)-dimensional embedding of \( y_n \)

pathway coordinate axis; subnetwork coordinate axis

covariance matrix

\[ \forall [y_n] = \Lambda \Lambda^T + \Sigma \equiv \Omega \]

low-rank matrix factorization
Current Approaches

Non-Bayesian: $\Lambda_{\rho,\gamma} = \arg\min_{\Lambda} \text{loss}(\Lambda) + \sum_{p,k} \mathcal{P}_{\rho,\gamma}(|\lambda_{pk}|)$
Current Approaches

Non-Bayesian: $\Lambda_{\rho, \gamma} = \arg\min_{\Lambda} \text{loss}(\Lambda) + \sum_{p,k} \mathcal{P}_{\rho, \gamma}(|\lambda_{pk}|)$
Non-Bayesian: $\Lambda_{\rho,\gamma} = \arg\min_{\Lambda} \text{loss}(\Lambda) + \sum_{p,k} \mathcal{P}_{\rho,\gamma}(|\lambda_{pk}|)$

(Breiman’s black-box: loss + regularizer $\rho, \gamma$)

<table>
<thead>
<tr>
<th>method</th>
<th>$\Lambda$</th>
<th>$K$ unknown?</th>
<th>$P \geq 10^4$ feasible?</th>
<th>thresholding?</th>
<th>$\Lambda = \Lambda^*$?</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMD (W)</td>
<td>✓</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
<td>X</td>
</tr>
<tr>
<td>FANC (HY)</td>
<td>✓</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
<td>X</td>
</tr>
</tbody>
</table>

tuning parameters: $\rho$ and $\gamma$

[Witten et al. (2009); Hirose & Yamamoto (2014)]
Current Approaches

- **Non-Bayesian:** $\Lambda_{\rho,\gamma} = \arg\min_{\Lambda} \text{loss}(\Lambda) + \sum_{p,k} \mathcal{P}_{\rho,\gamma}(|\lambda_{pk}|)$

  (Breiman’s black-box: loss + regularizer $\rho,\gamma$)

<table>
<thead>
<tr>
<th>method</th>
<th>$\Lambda$</th>
<th>$K$ unknown?</th>
<th>$P \geq 10^4$ feasible?</th>
<th>thresholding?</th>
<th>$\Lambda = \Lambda^*$?</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMD (W)</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>FANC (HY)</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

**tuning parameters:** $\rho$ and $\gamma$

[Witten et al. (2009); Hirose & Yamamoto (2014)]

- **Bayesian:** $-\log\text{-posterior}_{\rho,\gamma}(\Lambda) \propto -\log\text{-likelihood}(\Lambda) - \log\text{-prior}_{\rho,\gamma}(|\Lambda|)$
Current Approaches

Non-Bayesian: $\Lambda_{\rho,\gamma} = \text{argmin } \Lambda \text{ loss}(\Lambda) + \sum_{p,k} \mathcal{D}_{\rho,\gamma}(|\lambda_{pk}|)$

(Breiman’s black-box: loss + regularizer $\rho,\gamma$)

<table>
<thead>
<tr>
<th>method</th>
<th>$\Lambda$</th>
<th>$K$ unknown?</th>
<th>$P \geq 10^4$ feasible?</th>
<th>thresholding?</th>
<th>$\Lambda = \Lambda^*$?</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMD (W)</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>FANC (HY)</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

tuning parameters: $\rho$ and $\gamma$

[Witten et al. (2009); Hirose & Yamamoto (2014)]

Bayesian: $-\text{log-posterior}_{\rho,\gamma}(\Lambda) \propto -\text{log-likelihood}(\Lambda) - \text{log-prior}_{\rho,\gamma}(|\Lambda|)$
Current Approaches

Non-Bayesian: $\Lambda_{\rho,\gamma} = \arg\min_{\Lambda} \text{loss}(\Lambda) + \sum_{p,k} P_{\rho,\gamma}(|\lambda_{pk}|)$
(Breiman’s black-box: loss + regularizer $\rho,\gamma$)

<table>
<thead>
<tr>
<th>method</th>
<th>$\Lambda$</th>
<th>$K$ unknown?</th>
<th>$P \geq 10^4$ feasible?</th>
<th>thresholding?</th>
<th>$\Lambda = \Lambda^*$?</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMD (W)</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>FANC (HY)</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

tuning parameters: $\rho$ and $\gamma$

Bayesian: $-\log\text{-posterior}_{\rho,\gamma}(\Lambda) \propto -\log\text{-likelihood}(\Lambda) - \log\text{-prior}_{\rho,\gamma}(|\Lambda|)$
([\Lambda 0_{P \times \infty}] \in \mathbb{R}^{P \times \infty})

<table>
<thead>
<tr>
<th>method</th>
<th>$\Lambda$</th>
<th>$K$ unknown?</th>
<th>$P \geq 10^4$ feasible?</th>
<th>thresholding?</th>
<th>$\Lambda = \Lambda^*$?</th>
</tr>
</thead>
<tbody>
<tr>
<td>BFRM (C)</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>IBP (KG)</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>MGP (BD)</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
</tbody>
</table>

[Carvalho et al. (2008); Knowles & Ghahramani (2011); Bhattacharya & Dunson (2011)]
Outline

Multivariate Data & Factor Analysis

Approaches for Factor Analysis

Expandable Factor Analysis (MAP Estimation)

Expandable Factor Analysis (“Model” Averaging)

Applications
\[ y_p = Z \lambda_p + e_p \text{ for } p = 1, \ldots, P \]

\[ y_{np} = z_n^T \lambda_p + e_{np} \]

Factor Analysis as Regression

\[ P \quad y_n = \Lambda \]

- \( y_p = Z \lambda_p + e_p \) for \( p = 1, \ldots, P \)
- \( y_{np} = z_n^T \lambda_p + e_{np} \)
- \( P \) and \( \Lambda \)
- Expandable Factor Analysis (MAP Estimation)
Factor Analysis as Regression

\[ y_p = Z \lambda_p + e_p \text{ for } p = 1, \ldots, P \]

\[ y_{np} = z_n^T \lambda_p + e_{np} \]

\[ xFA \text{ assumes } \lambda_p \in \mathbb{R}^{\infty \times 1} \text{ and } Z \in \mathbb{R}^{N \times \infty} \]
Factor Analysis as Regression

\[ \mathbf{y}_p = \mathbf{Z} \lambda_p + \mathbf{e}_p \] for \( p = 1, \ldots, P \)

\[ \mathbf{y}_{np} = \mathbf{z}_n^T \lambda_p + \mathbf{e}_{np} \]

- \( \lambda_p \)s ∈ \( \mathbb{R}^{\infty \times 1} \) and \( \mathbf{Z} \) ∈ \( \mathbb{R}^{N \times \infty} \)
- \( \mathbf{xFA} \) identifies four properties of prior/penalty

---

**Expandable Factor Analysis (MAP Estimation)**

11
Factor Analysis as Regression

\[ y_p = Z \lambda_p + e_p \text{ for } p = 1, \ldots, P \]
\[ y_{np} = z_n^T \lambda_p + e_{np} \]

\[ \text{xFA assumes } \lambda_p \text{ s} \in \mathbb{R}^{\infty \times 1} \text{ and } Z \in \mathbb{R}^{N \times \infty} \]

\[ \text{xFA identifies four properties of prior/penalty} \]
Factor Analysis as Regression

\[ y_p = Z \lambda_p + e_p \text{ for } p = 1, \ldots, P \]
\[ y_n = z_n^T \lambda_p + e_n \]

xFA assumes \( \lambda_p \in \mathbb{R}^{\infty \times 1} \) and \( Z \in \mathbb{R}^{N \times \infty} \)

xFA identifies four properties of prior/penalty

(a) **near unbiasedness** of the estimate \( \hat{\lambda}_{ij} \)
(b) **sparsity** of the estimate \( \hat{\lambda}_{ij} \)
(c) **continuity** of the estimate \( \hat{\lambda}_{ij} \) as a function of data
(d) **ordered factors** with loadings increasingly shrunk towards 0 for \( k = 1, \ldots, \infty \)

<table>
<thead>
<tr>
<th>method</th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMD (W)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>FANC (HY)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>BFRM (C)</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>IBP (KG)</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>MGP (BD)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>xFA</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

[Fan & Li (2001); Carvalho et al. (2008); Witten et al. (2009); Knowles & Ghahramani (2011); Bhattacharya & Dunson (2011); Hirose & Yamamoto (2014)]
Properties (a) – (c): Generalized Double Pareto Prior

\[
\text{GDP}(\alpha, \eta) = p(\lambda | \alpha, \eta) = \frac{\alpha}{2\eta} \left(1 + \frac{|\lambda|}{\eta}\right)^{-(\alpha+1)}
\]

[Armagan et al. (2013)]

\[
\mathcal{P}(\alpha, \eta) = -\log p(\lambda) \propto (\alpha + 1) \log \left(1 + \frac{|\lambda|}{\eta}\right)
\]

[Candes et al. (2008)]

\[
\text{minimize} \ \text{loss}(\lambda_{pk}) + \mathcal{P}_{\alpha,\eta}(|\lambda_{pk}|)
\]

Expandable Factor Analysis (MAP Estimation)
Property (d): Multiscale GDP Prior

multiscale generalized double Pareto (mGDP) prior

\[
p_{m\text{GDP}}(\lambda_p) = \prod_{k=1}^{\infty} p_{\text{GDP}}(\lambda_{pk}|\alpha_k, \eta_k) = \prod_{k=1}^{\infty} \frac{\alpha_k}{2\eta_k} \left(1 + \frac{|\lambda_{pk}|}{\eta_k}\right)^{-(\alpha_k+1)}
\]

\[
p_{m\text{GDP}}(\Lambda) = \prod_{p=1}^{P} p_{m\text{GDP}}(\lambda_p) \equiv \text{mGDP}(\alpha_{1:\infty}, \eta_{1:\infty})
\]
Property (d): Multiscale GDP Prior

multiscale generalized double Pareto (mGDP) prior

\[ p_{\text{mGDP}}(\lambda_p) = \prod_{k=1}^{\infty} p_{\text{GDP}}(\lambda_{pk} | \alpha_k, \eta_k) = \prod_{k=1}^{\infty} \frac{\alpha_k}{2\eta_k} \left(1 + \frac{|\lambda_{pk}|}{\eta_k}\right)^{-\left(\alpha_k + 1\right)} \]

\[ p_{\text{mGDP}}(\Lambda) = \prod_{p=1}^{P} p_{\text{mGDP}}(\lambda_p) \equiv \text{mGDP}(\alpha_{1:\infty}, \eta_{1:\infty}) \]

\[ \mathcal{C}_{\text{load}} = \left\{ \Lambda \mid \max_{1 \leq p \leq P} \sum_{k=1}^{\infty} \lambda_{pk}^2 < \infty \right\} \text{ [Bhattacharya & Dunson (2011)]} \]
Property (d): Multiscale GDP Prior

The multiscale generalized double Pareto (mGDP) prior is given by:

\[ p_{\text{mGDP}}(\lambda_p) = \prod_{k=1}^{\infty} p_{\text{GDP}}(\lambda_p | \alpha_k, \eta_k) = \prod_{k=1}^{\infty} \frac{\alpha_k}{2\eta_k} \left(1 + \frac{\lambda_p}{\eta_k}\right)^{-(\alpha_k + 1)} \]

and

\[ p_{\text{mGDP}}(\Lambda) = \prod_{p=1}^{P} p_{\text{mGDP}}(\lambda_p) \equiv \text{mGDP}(\alpha_{1:\infty}, \eta_{1:\infty}) \]

Let \( C_{\text{load}} = \left\{ \Lambda \mid \max_{1 \leq p \leq P} \sum_{k=1}^{\infty} \lambda_p^2 \eta_k^2 < \infty \right\} \) [Bhattacharya & Dunson (2011)]

Lemma

If \( \alpha_k > 2 \) and \( \frac{\eta_k}{\alpha_k} = \Theta \left( \frac{1}{k^m} \right) \) for \( k = 1, \ldots, \infty \) and \( m > 0.5 \), then

\[ \mathbb{P}_{\text{load}}\{C_{\text{load}}\} = 1 \]
Property (d): Multiscale GDP Prior

- multiscale generalized double Pareto (mGDP) prior

\[ p_{mGDP}(\lambda_p) = \prod_{k=1}^{\infty} p_{GDP}(\lambda_{pk}|\alpha_k, \eta_k) = \prod_{k=1}^{\infty} \frac{\alpha_k}{2\eta_k} \left(1 + \frac{\lambda_{pk}}{\eta_k}\right)^{-\left(\alpha_k + 1\right)} \]

\[ p_{mGDP}(\Lambda) = \prod_{p=1}^{P} p_{mGDP}(\lambda_p) \equiv mGDP(\alpha_{1:\infty}, \eta_{1:\infty}) \]

\[ \mathcal{C}_{load} = \left\{ \Lambda \mid \max_{1 \leq p \leq P} \sum_{k=1}^{\infty} \lambda_{pk}^2 < \infty \right\} \] [Bhattacharya & Dunson (2011)]

**Lemma**
If \( \alpha_k > 2 \) and \( \frac{\eta_k}{\alpha_k} = \Theta\left(\frac{1}{k^m}\right) \) \( k = 1, \ldots, \infty \) and \( m > 0.5 \), then

\[ \mathbb{P}_{load}\{\mathcal{C}_{load}\} = 1 \]

**limitations:** (i) \( \infty \) hyperparameters (ii) \( \infty \) columns for \( \Lambda \)
(i) relax $\infty$ hyperparameters limitation

Lemma

$\mathbb{P}_{load}\{C_{load}\} = 1$ holds when $\delta > 2$, $\rho > 0$,

$$\alpha_k = \delta^k \text{ and } \eta_k = \rho, \quad k = 1, \ldots, \infty \quad [mGDP(\alpha_{1:\infty}, \eta_{1:\infty})]$$
Structured Sparsity by Prior Finessing

(i) relax $\infty$ hyperparameters limitation

Lemma
$\mathbb{P}_{load}\{C_{load}\} = 1$ holds when $\delta > 2$, $\rho > 0$,

$$\alpha_k = \delta^k \text{ and } \eta_k = \rho, \quad k = 1, \ldots, \infty \quad [mGDP(\alpha_{1:\infty}, \eta_{1:\infty})]$$

(ii) relax $\infty$ $\Lambda$ columns limitation

Consequence
Given $\epsilon > 0$, there exists $K_0 = \Theta \left( \log^{-1} \delta \log \frac{P}{\epsilon^2} \right)$ such that

$$\mathbb{P}\{d_\infty(\Omega, \Omega^K) < \epsilon\} > 1 - \epsilon \quad \forall K \geq K_0,$$

where $\Omega^K = \Lambda^K \Lambda^{K^T} + \Sigma$ and $d_\infty(A, B) = \max_{1 \leq i, j \leq P} |a_{ij} - b_{ij}|$

Proof

Expandable Factor Analysis (MAP Estimation) 14
mGDP penalty

\[ \text{mGDP penalty} = - \log p_{mGDP}(\lambda_{pk}) \propto (\alpha_k + 1) \log \left( 1 + \frac{|\lambda_{pk}|}{\eta_k} \right) \]
mGDP penalty

\[ m\text{GDP penalty} = -\log p_{m\text{GDP}}(\lambda_{pk}) \propto (\alpha_k + 1) \log \left(1 + \frac{|\lambda_{pk}|}{\eta_k}\right) \]

local linear approximation (LLA) of the mGDP penalty \( \propto \frac{\delta^{k+1}}{\rho + |\lambda^{(0)}_{pk}|} |\lambda_{pk}| \) [Zou & Li (2008)]

![Graph showing LLA at 0.1, 2, and 4 compared to the truth for different values of k and delta.](image)

(a) mGDP Penalty when \( \rho = 0.01 \)
Estimation of $\Lambda$ using regularized regression

minimize: $-\log$-likelihood $- \log$-mGDP-prior(30)

at the $(t+1)$-th iteration,

$$
\lambda_p^{(t+1)} = \arg \min_{\lambda_p} \frac{N}{2} \frac{\| w_p^{(t)} - X^{(t)} \lambda_p \|^2}{\sigma_{pp}^{2(t)}} + \sum_{k=1}^{K} (\delta^k + 1) \log \left( 1 + \frac{|\lambda_{pk}|}{\rho} \right)
$$

log-likelihood loss

mGDP penalty
**Estimation of $\Lambda$ using regularized regression**

**minimize:** $- \log\text{-likelihood} - \log\text{-mGDP-prior}(30)$

at the $(t + 1)$-th iteration,

$$
\lambda_p^{(t+1)} = \arg\min_{\lambda_p} \frac{N}{2} \| w_p^{(t)} - X^{(t)} \lambda_p \|^2 \sum_{k=1}^{K} (\delta^k + 1) \log \left( 1 + \frac{\| \lambda_{pk} \|}{\rho} \right) + \sum_{k=1}^{K} \frac{\delta^k + 1}{\rho + |\lambda_{pk}|} \| \lambda_{pk} \|
$$

using local linear approximation (LLA) of the mGDP penalty

$$
\lambda_p^{lla^{(t+1)}} = \arg\min_{\lambda_p} \frac{N}{2} \| w_p^{(t)} - X^{(t)} \lambda_p \|^2 \sum_{k=1}^{K} \frac{\delta^k + 1}{\rho + |\lambda_{pk}|} |\lambda_{pk}|
$$

$$
= \left[ (\lambda_{p1}^{(t)} - c_{p1}^{(t)}) + , \ldots , (\lambda_{pk}^{(t)} - c_{pk}^{(t)}) + , \ldots , (\lambda_{pK}^{(t)} - c_{pK}^{(t)}) + \right]^T
$$
Estimation of $\Lambda$ and $\Sigma$ using EM-type algorithm

\[ \text{soft-thresholding operator} \]
Estimation of $\Lambda$ and $\Sigma$ using EM-type algorithm

保密 soft-thresholding operator
Estimation of $\Lambda$ and $\Sigma$ using EM-type algorithm

Soft-thresholding operator

$$(\lambda^{(t)}_{pk} - c^{(t)}_{pk})^+ = \begin{cases} 
\lambda^{(t)}_{pk} - c^{(t)}_{pk} & \text{if } \lambda^{(t)}_{pk} > c^{(t)}_{pk} \\
\lambda^{(t)}_{pk} + c^{(t)}_{pk} & \text{if } \lambda^{(t)}_{pk} < -c^{(t)}_{pk} \\
0 & \text{otherwise}
\end{cases}$$

$$c^{(t)}_{pk} = \frac{\sigma^{2(t)}_{pp} (\delta^k + 1)}{N(\rho + |\lambda^{(t)}_{pk}|)}$$
Estimation of $\Lambda$ and $\Sigma$ using EM-type algorithm

**soft-thresholding operator**

$$(\lambda_{pk}^{(t)} - c_{pk}^{(t)})_+ = \begin{cases} 
\lambda_{pk}^{(t)} - c_{pk}^{(t)} & \text{if } \lambda_{pk}^{(t)} > c_{pk}^{(t)} \\
\lambda_{pk}^{(t)} + c_{pk}^{(t)} & \text{if } \lambda_{pk}^{(t)} < -c_{pk}^{(t)} \\
0 & \text{otherwise}
\end{cases}$$

$$c_{pk}^{(t)} = \frac{\sigma_{pp}^{2(t)}(\delta^k + 1)}{N(\rho + |\lambda_{pk}^{(t)}|)}$$

*Adaptive* Threshold

Constant Threshold

Unbiased
Estimation of $\Lambda$ and $\Sigma$ using EM-type algorithm

soft-thresholding operator

$$
(\lambda^{(t)}_{pk} - c^{(t)}_{pk})^+ =
\begin{cases}
\lambda^{(t)}_{pk} - c^{(t)}_{pk} & \text{if } \lambda^{(t)}_{pk} > c^{(t)}_{pk} \\
\lambda^{(t)}_{pk} + c^{(t)}_{pk} & \text{if } \lambda^{(t)}_{pk} < -c^{(t)}_{pk} \\
0 & \text{otherwise}
\end{cases}
$$

$$c^{(t)}_{pk} = \frac{\sigma^2_{pp}(\delta^k + 1)}{N(\rho + |\lambda^{(t)}_{pk}|)}$$
Estimation of $\Lambda$ and $\Sigma$ using EM-type algorithm

The soft-thresholding operator is defined as:

$$(\lambda_{pk}^{(t)} - c_{pk}^{(t)})^+ = \begin{cases} 
\lambda_{pk}^{(t)} - c_{pk}^{(t)} & \text{if } \lambda_{pk}^{(t)} > c_{pk}^{(t)} \\
\lambda_{pk}^{(t)} + c_{pk}^{(t)} & \text{if } \lambda_{pk}^{(t)} < -c_{pk}^{(t)} \\
0 & \text{otherwise}
\end{cases}$$

$$c_{pk}^{(t)} = \frac{\sigma_{pp}^{2(t)}(\delta^k + 1)}{N(\rho + |\lambda_{pk}^{(t)}|)}$$

Below is a table comparing different methods:

<table>
<thead>
<tr>
<th>method</th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(d)</th>
<th>scalable?</th>
<th>factor selection?</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMD</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓-?</td>
</tr>
<tr>
<td>FANC</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓-?</td>
</tr>
<tr>
<td>BFRM</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>IBP</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>MGP</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>xFA</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Expandable Factor Analysis (MAP Estimation)
Theoretical Guarantees

Theorem

The objective function of xFA does not decrease at every iteration. Under standard regularity conditions, $\Lambda^{lla(t)}$ and $\Sigma^{lla(t)}$ converge to their fixed points.

(A1) $y_n = \Lambda^{*}z_n + \epsilon_n$

(A2) If $N \rightarrow \infty$, then $\alpha_{kN} \rightarrow \infty$, $\alpha_{kN}/p_N \rightarrow 0$, and $p_N \eta_{kN} \rightarrow c_k > 0$ (modifications: $\alpha_{kN} = \delta_k \log N$ and $\eta_{kN} = \rho/p_N$)

Theorem (A1) – (A2) hold and $\Lambda^{lla(t)}$ and $\Sigma^{lla(t)}$ are MAP estimates of $\Lambda^{*}$ and $\Sigma^{*}$.

Then,

(i) $\Lambda^{lla(t)}$ consistently estimates $\Lambda^{*}$ and its nonzero elements are asymptotically normal

(ii) $\Sigma^{lla(t)}$ is asymptotically normal and consistently estimates $\Sigma^{*}$

Proof

Expandable Factor Analysis (MAP Estimation) 18
Theoretical Guarantees

Theorem
The objective function of xFA does not decrease at every iteration. Under standard regularity conditions, $\Lambda^{lla(t)}$ & $\Sigma^{lla(t)}$ converge to their fixed points.

(A1) $y_n = \Lambda^* z_n + e_n$

(A2) If $N \to \infty$, then $\alpha_{kN} \to \infty$, $\alpha_{kN} / \sqrt{N} \to 0$, and $\sqrt{N} \eta_{kN} \to c_k > 0$
(modifications: $\alpha_{kN} = \delta^k \log N$ and $\eta_{kN} = \rho / \sqrt{N}$)

Theorem
(A1) – (A2) hold and $\Lambda_{N}^{lla}$ and $\Sigma_{N}^{lla}$ are MAP estimates of $\Lambda^*$ and $\Sigma^*$. Then,

(i) $\Lambda_{N}^{lla}$ consistently estimates $\Lambda^*$ and its nonzero elements are asymptotically normal

(ii) $\Sigma_{N}^{lla}$ is asymptotically normal and consistently estimates $\Sigma^*$
Outline

Multivariate Data & Factor Analysis

Approaches for Factor Analysis

Expandable Factor Analysis (MAP Estimation)

Expandable Factor Analysis (“Model” Averaging)

Applications
Tuning Parameters $\rho$ and $\delta$

- no guarantee that there exists $(\rho_0, \delta_0)$ with “the optimal” $\Lambda$

- xFA’s estimate $\overline{\Lambda} = \sum_{g=1}^{G} \pi_g \Lambda_g$

- $\log \pi_g$ and BIC are related, so FANC should be unstable to choice of tuning parameter selection
Laplace’s Method for $\log \pi_g$

$\pi_g$ depends on $p(Y | M_g)$ and $p(M_g)$

$$p(Y | M_g) = \int_{\Lambda, \Sigma} p(Y | \Lambda, \Sigma, M_g) p(\Lambda) p(\Sigma) d\Lambda d\Sigma$$

Proof
Laplace’s Method for $\log \pi_g$

- $\pi_g$ depends on $p(Y|M_g)$ and $p(M_g)$

$$p(Y|M_g) = \int_{\Lambda, \Sigma} p(Y|\Lambda, \Sigma, M_g) p(\Lambda) p(\Sigma) d\Lambda d\Sigma$$

- $p(Y|M_g) \approx \int_{\Lambda} p(Y|\Lambda, \Sigma^g, M_g) p(\Lambda) d\Lambda$ [Tierney et al. (1989); Rue et al. (2009)]
Laplace’s Method for $\log \pi_g$

- $\pi_g$ depends on $p(Y|\mathcal{M}_g)$ and $p(\mathcal{M}_g)$

$$p(Y|\mathcal{M}_g) = \int_{\Lambda, \Sigma} p(Y|\Lambda, \Sigma, \mathcal{M}_g) p(\Lambda) p(\Sigma) \, d\Lambda \, d\Sigma$$

- $p(Y|\mathcal{M}_g) \approx \int_{\Lambda} p(Y|\Lambda, \Sigma^g, \mathcal{M}_g) p(\Lambda) \, d\Lambda$ [Tierney et al. (1989); Rue et al. (2009)]

- $\Lambda^{\text{ll}a}$ “averages” $\Lambda_{\rho, \delta}$; more stable than selection of optimal $\Lambda$ based on BIC
Laplace’s Method for $\log \pi_g$

$\pi_g$ depends on $p(Y|M_g)$ and $p(M_g)$

$$p(Y|M_g) = \int_{\Lambda, \Sigma} p(Y|\Lambda, \Sigma, M_g) p(\Lambda) p(\Sigma) d\Lambda d\Sigma$$

$p(Y|M_g) \approx \int_{\Lambda} p(Y|\Lambda, \Sigma_g, M_g) p(\Lambda) d\Lambda$ [Tierney et al. (1989); Rue et al. (2009)]

$L_{ll}a$ “averages” $\Lambda_{\rho,\delta}$s; more stable than selection of optimal $\Lambda$ based on BIC

idea can be used for constructing approximate posteriors

Proof
Model Selection Consistency of xFA

**Assumption:** $\rho - \delta$ grid is fine enough so that grid index $g^*$ corresponds to the true factor analysis model $\mathcal{M}_{g^*}$

**Theorem**

*If the eigenvalues of $\Psi^{g^*}$ are bounded away from zero and the xFA model is over-fitted, then $\lim_{N \uparrow \infty} \mathbb{P}\{\pi_{g^*} = 1\} = 1$ and $\lim_{N \uparrow \infty} \mathbb{P}\{\overline{\Lambda}_{ll} = \Lambda^*\} = 1$*

Proof
Outline

Multivariate Data & Factor Analysis

Approaches for Factor Analysis

Expandable Factor Analysis (MAP Estimation)

Expandable Factor Analysis (“Model” Averaging)

Applications
Simulated Data Analysis

compared xFA with FANC (HY) and SPC (W) using

cumulative number of non-zero loadings
\[ \text{CNNL}_k = \sum_{l=1}^{k} \sum_{p=1}^{P} 1_{\lambda_{pl} \neq 0} \]

cumulative proportion of explained variance
\[ \text{CPEV}_k = \frac{\text{trace}(\lambda_{1:k} \lambda_{1:k}^T)}{\text{trace}(\sum_{n=1}^{N} y_n y_n^T) / N} \]

[Shen & Huang (2008); Witten et al. (2009); Hirose & Yamomoto (2014)]

simulated lower-triangular $\Lambda$ for comparisons

set $K = 20$
Simulations

<table>
<thead>
<tr>
<th>signal-to-noise ratio</th>
<th>loadings matrix</th>
<th>priors</th>
<th>methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td>sparse</td>
<td>MGP &amp; Uniform</td>
<td>FANC, SPC, &amp; xFA</td>
</tr>
<tr>
<td>low</td>
<td>dense</td>
<td>MGP &amp; Uniform</td>
<td>FANC, SPC, &amp; xFA</td>
</tr>
</tbody>
</table>

- high signals $\sim$ Uniform(-3, 3)
- low signals $\sim$ Uniform(-0.5, 0.5)
- error variance $\sim$ Uniform(1, 3)
- $P = 500; N = P \log P; K = 5$
- 50 replications
Factor Selection

Dense Loadings Matrix  
Low Signal Strength

Sparse Loadings Matrix  
High Signal Strength

selected rank of loadings matrix = largest column index with non-zero entries
maximum possible rank = 20
true rank = 5
xFAn performs better than its competitors in sparse loadings matrices with high signal-to-noise ratio.

xFA’s performance in estimating dense loadings matrices depends on the type of signal.
analyzed two publicly available microarray data sets

- 608 samples and 17,229 genes (hereafter *HapMap3* data)
- 480 samples and 8,718 genes (hereafter *LCL* data)

used 5 fold cross-validation

only median results are reported
\( \log \pi_g: \log \text{Posterior Model Weights} \)

**HapMap3 data**

\( \log \pi_g \) is multimodal (\( \Rightarrow \) BIC surface is multimodal)

\( \Rightarrow \) FANC and SPC tuning parameter selection is unstable

**factor selection:** FANC = 20; SPC = 1; \( x_{FA} = 7 \)

**LCL data**

\( \log \pi_g \) is multimodal (\( \Rightarrow \) BIC surface is multimodal)

\( \Rightarrow \) FANC and SPC tuning parameter selection is unstable

**factor selection:** FANC = 20; SPC = 1; \( x_{FA} = 7 \)
structured mGDP penalty ensures
\[ j \] exponential decay of non-zero loadings across factors
\[ j \] smoothly varies between the fraction of non-zero loadings in FANC
\[ j \] and SPC
\[ j \] model averaging requires no selection of tuning parameters
\[ j \] FANC’s loadings matrix is overly dense
\[ j \] SPC’s loadings matrix is extremely sparse
SPC’s CPEV & CNNL is smallest for HapMap3 and LCL data

xFA’s CPEV was largest for HapMap3; FANC’s CPEV was largest for LCL

xFA explained greater fraction of variance with a small fraction of non-zero loadings in HapMap3
Discussion

**extension:** linear Structural Equation Models (SEM)

**measurement part:**

$$\text{obs}(n) = \Lambda_{\text{meas}} \text{factors}(n) + \Gamma_{\text{meas}} \text{covariates vector}(n) + \text{err}_{\text{meas}}(n)$$

**structural part:**

$$\text{factors}(n) = \Phi \text{factors}(n) + \Theta \text{covariates vector}(n) + \text{err}_{\text{stru}}(n)$$

**example:** dependence among multiple sources of high-dimensional genomic data (Bayesian Canonical Correlation Analysis)

**non-parametric extensions:** scalable extensions for nonlinear SEM using Gaussian processes

**sampling:** Laplace approximation as a warm start for Gibbs sampling
Computationally Feasible Parameter Estimation: Variational Inference

$p \to \infty$

“big $p$”


Srivastava, S. and V. Baladandayuthapani (in preparation). Regularized and supervised Bayesian canonical correlation analysis for integrating multi-platform genomic data
Computationally Feasible Parameter Estimation: Laplace-type Approximations

$p \rightarrow \infty$

“big p”


Computationally Feasible Sampling: Big Bayes

- **big data features**
  - complex dependencies
  - don’t fit on a single machine

- **computational framework assumptions**
  - sampling is feasible in the parameter space

- **advantages**
  - reuse of existing sampler code with minimal modifications (wide applicability)
  - computation involves no communication (easy asynchronous and distributed implementations)

\[ n \rightarrow \infty \quad \text{“big n”} \]
Computationally Feasible Sampling: Big Bayes

Big Data → Data Subsets → Subset Posteriors → ?

“big n” → n → ∞
Computationally Feasible Sampling: Big Bayes

**M-Posterior:** Median of posteriors


  [http://cran.r-project.org/web/packages/Mposterior/index.html](http://cran.r-project.org/web/packages/Mposterior/index.html)

**WASP:** Wasserstein barycenter of posteriors


Computationally Feasible Parameter Estimation: Big Bayes

Distributed Expectation-Maximization


Big Data

parameters (p) \( p \to \infty \)

samples (n)

“big p”

parameters (p) \( p \to \infty \)

samples (n)

“big n”

life sciences lab

Google

cancer genome atlas
M-Posterior & WASP

(a) M-Posterior

(b) WASP
mGDP Support

We prove that \( P_{\text{load}}(\mathcal{E}_{\text{load}}) = 1 \) using the definition of \( \mathcal{E}_{\text{load}} \). Specifically,

\[
\begin{align*}
P_{\text{load}}(\mathcal{E}_{\text{load}}) &= P_{\text{load}}(\Lambda | \max_{1 \leq p \leq P} \sum_{k=1}^{\infty} \lambda^2 p_k < \infty) = 1 - \lim_{t \to \infty} P_{\text{load}}(\Lambda | \max_{1 \leq p \leq P} \sum_{k=1}^{\infty} \lambda^2 p_k \geq t) \\
&\geq 1 - \lim_{t \to \infty} \sum_{p=1}^{P} P_{\text{load}}(\Lambda | \sum_{k=1}^{\infty} \lambda^2 p_k \geq t) \geq 1 - \lim_{t \to \infty} \frac{\sum_{k=1}^{\infty} E[\lambda^2_{1k}]}{t} = 1 - \lim_{t \to \infty} \frac{\sum_{k=1}^{\infty} V[\lambda_{1k}]}{t}.
\end{align*}
\]

(1)

Armagan et al. (2013) show that \( \lambda_{1k} \sim \text{GDP}(\alpha, \eta) \) has \( V[\lambda_{1k}] = 2\eta^2 (\alpha - 1)^{-1}(\alpha - 2)^{-1} \) for \( \alpha > 2 \), so

\[
\sum_{k=1}^{\infty} V[\lambda_{1k}] = 2 \sum_{k=1}^{\infty} \eta^2_k \frac{1}{\alpha_k - 1} \frac{1}{\alpha_k - 2} \leq 2 \sum_{k=1}^{\infty} \eta^2_k \frac{1}{\alpha_k} \frac{1}{\alpha_k - 2} \leq 2 \sum_{k=1}^{\infty} \frac{\eta^2_k}{\alpha^2_k} \left(1 - \frac{2}{\alpha_k}\right)^{-2} \leq 2 \sum_{k=1}^{\infty} \frac{\eta^2_k}{\alpha^2_k} \left(1 + \frac{4}{\alpha_k} + o\left(\frac{1}{\alpha_k}\right)\right) < 2(1 + 2 + O(1)) \sum_{k=1}^{\infty} \frac{\eta^2_k}{\alpha^2_k} < \infty
\]

(2)

if \( \alpha_k > 2 \) and \( \eta_k / \alpha_k = O\left(1/k^m\right) \) for \( m > 0.5 \); therefore, \( P \sum_{k=1}^{\infty} V[\lambda_{1k}] \) in (1) is bounded and \( P_{\text{load}}(\mathcal{E}_{\text{load}}) = 1 \).
Truncation Proof

Expandable factor analysis finds $K_0 = K(P, \delta, \rho, \epsilon)$ that upper bounds $\mathbb{P}\{d_\infty(\Omega, \Omega^{K_0}) \geq \epsilon\}$ by $\epsilon$.

$$\mathbb{P}\{\Omega^{K_0} | d_\infty(\Omega, \Omega^{K_0}) < \epsilon\} \geq 1 - \sum_{k=K_0+1}^{\infty} \frac{E[\sum_{i=1}^{P} \sum_{j=1}^{P} |\lambda_{ik}||\lambda_{jk}|]}{\epsilon}.$$  

Using Hölder’s inequality and noticing that $\lambda_{ik}$’s are sampled independently from GDP($\alpha_k, \eta_k$),

$$\mathbb{P}\{\Omega^{K_0} | d_\infty(\Omega, \Omega^{K_0}) < \epsilon\} \geq 1 - \sum_{k=K_0+1}^{\infty} \frac{E[\left(\sum_{i=1}^{P} |\lambda_{ik}|\right)^2]}{\epsilon}$$

$$= 1 - \frac{p^2}{\epsilon} \sum_{k=K_0+1}^{\infty} \left(V[|\lambda_{1k}|] + E^2[|\lambda_{1k}|]\right).$$ (3)

Using (3), $K_0$ that satisfies $\mathbb{P}\{\Omega^{K_0} | d_\infty(\Omega, \Omega^{K_0}) < \epsilon\} \geq 1 - \epsilon$ depends as follows on the sufficient conditions of Lemma:

(I) $\frac{p^2}{\epsilon} \Theta\left(\frac{1}{\delta^2 K_0}\right) \leq \epsilon \Rightarrow K_0 = \Theta\left(\log^{-1} \delta \log \frac{p}{\epsilon}\right)$;

(II) $\frac{p^2}{\epsilon} \Theta\left(\rho^2 K_0\right) \leq \epsilon \Rightarrow K_0 = \Theta\left(\log^{-1} \frac{1}{\rho} \log \frac{p}{\epsilon}\right)$;

(III) $\frac{p^2}{\epsilon} \Theta\left(\frac{\rho^2 K_0}{\delta^2 K_0}\right) \leq \epsilon \Rightarrow K_0 = \Theta\left(\log^{-1} \frac{\delta}{\rho} \log \frac{p}{\epsilon}\right)$.  

Appendix
Nonconcave Optimization

\[ f(x) \]

Global Optimum

Local Optimum

Lasso

\( x \) \( (P - \text{dimensional}) \)
Log-posterior based on the Expectation-Maximization algorithm and local linear approximation

Denoting $Z$ as missing data and $Y$ as observed data, expandable factor analysis extends the maximum likelihood estimation to maximum a posteriori estimation of $\Lambda$ and $\Sigma$ using multiscale generalized double Pareto prior on $\Lambda$ and Jeffreys’ prior on the diagonal elements of $\Sigma$. This involves maximization of $E\{ \log p(Z, \Lambda, \Sigma \mid Y, \Lambda^{(t)}, \Sigma^{(t)}, \alpha_{1:K}, \eta_{1:K}) \}$

$$= - \sum_{p=1}^{P} \log p_{mis}(\lambda_{p}, \sigma_{pp}^{2} \mid S_{yy}, \Psi^{(t)}, \tilde{\Lambda}^{(t)}) - \frac{N+2}{2} \sum_{p=1}^{P} \log \sigma_{pp}^{2}, \quad (4)$$

where the analytic forms of conditional expectations, $\Psi^{(t)}$, and $\tilde{\Lambda}^{(t)}$ are as follows:

$$S_{yy} = \frac{1}{N} \sum_{n=1}^{N} y_{n}y_{n}^{T}; \quad S_{zz} = \frac{1}{N} \sum_{n=1}^{N} z_{n}z_{n}^{T}; \quad S_{yz} = \frac{1}{N} \sum_{n=1}^{N} y_{n}z_{n}^{T}; \quad (5)$$

$$\Delta = I - \Lambda^{T} (\Lambda \Lambda^{T} + \Sigma)^{-1} \Lambda; \quad \Gamma = (\Lambda \Lambda^{T} + \Sigma)^{-1} \Lambda; \quad \Psi = \Delta + \Gamma^{T} S_{yy} \Gamma; \quad \tilde{\Lambda} = S_{yy} \Gamma;$$

$$\mathbb{E}[S_{zz} \mid Y, \Lambda^{(t)}, \Sigma^{(t)}] = \Delta^{(t)} + \Gamma^{(t)}^{T} S_{yy} \Gamma^{(t)} = \Psi^{(t)}; \quad \mathbb{E}[S_{yz} \mid Y, \Lambda^{(t)}, \Sigma^{(t)}] = \tilde{\Lambda}^{(t)}.$$

We observe that (4) splits into $P$ separate terms corresponding to each dimension of $Y$; therefore, estimating $\Lambda$ and $\Sigma$ by maximizing (4) at the $(t+1)$th iteration is equivalent to separately minimizing $P$ objectives of the form

$$\log p_{mis}(\lambda_{p}, \sigma_{pp}^{2} \mid S_{yy}, \Psi^{(t)}, \tilde{\Lambda}^{(t)}) + \frac{N+2}{2} \log \sigma_{pp}^{2}, \quad (6)$$

with respect to $\lambda_{p}$ and $\sigma_{pp}^{2}$ for $p = 1, \ldots, P$; each of which is bi-convex.

Following the standard Expectation-Maximization approach, for $p = 1, \ldots, P$, we first fix $\sigma_{pp}^{2}$ at $\sigma_{pp}^{2(t)}$ and minimize

$$\log p_{mis}(\lambda_{p}, \sigma_{pp}^{2(t)} \mid S_{yy}, \Psi^{(t)}, \tilde{\Lambda}^{(t)})$$

in (6). We then fix $\lambda_{p}$ at $\lambda_{p}^{(t)}$ in (6) and minimize with respect to $\sigma_{pp}^{2}$, and repeat these steps until convergence to the local optimum.
Block coordinate descent algorithm for estimation of $\Lambda$

We suppress $(t)$ in $w_p$ and $X$ to ease notation. The cycle of coordinate descent algorithm for updating $\lambda_p^{l(a)(t)}$ to $\lambda_p^{l(a)(t+1)}$ starts by initializing $\tilde{\Lambda}^{(1)} = \Lambda^{l(a)(t)}$ and $(i+1)$th cycle updates $\tilde{\Lambda}_p^{(i)}$ using the objective as

$$\tilde{\lambda}_p^{(i+1)} = \arg\min_{\lambda_p^k} \lambda_p^k \frac{\lambda_p^k X^T X + 2 \lambda_p^k (\tilde{\Lambda}_p^{(i)} X_k - X_k w_p)}{2} + \frac{(\alpha_k + 1) \sigma_p^{(t)}}{(\eta_k + |\lambda_p^{(i)})^k| N} |\lambda_p^k|.$$

The solution of this convex program is obtained as a simple extension of glmnet updates, so that

$$\tilde{\lambda}_p^{(i+1)} = \frac{\text{sign}(l_p^{(i)} X^T X)}{|l_p^{(i)} X^T X|} \left( |l_p^{(i)} X^T X| - \frac{(\alpha_k + 1) \sigma_p^{(t)}}{(\eta_k + |\lambda_p^{(i)})^k| N} |\lambda_p^k| \right).$$

(7)

where $l_p^{(i)} = X^T X_p w - \tilde{\Lambda}_p^{(i)} X_p X_k$. Noticing that $\Psi = X^T X$ yields the update. We also exploit the form of the model and use it to update the $k$th column of $\tilde{\Lambda}^{(i)}$. This modification results in $K$ block updates for $\tilde{\Lambda}^{(i)}$ in a single cycle of expandable factor analysis coordinate descent algorithm. These update cycles are repeated multiple times until the change in $\tilde{\Lambda}$ is negligible, and then we set $\Lambda^{l(a)(t+1)} = \tilde{\Lambda}(\infty)$.  

Soft Thresholding
Asymptotic normality and consistency of $\Lambda_{ll}^a$

Denoting $\Lambda_{ll}^a$ as the expandable factor analysis estimate of $\lambda_p$ from local linear approximations using $\lambda_p^{(0)}$ as the $\sqrt{N}$-consistent sequence of estimators then for $p = 1, \ldots, P$

$$\lambda_{ll}^a = \arg\min_{\lambda_p} \frac{N}{2} \| w_p^{(0)} - X^{(0)} \lambda_p \|^2 + \frac{\sum_{k=1}^{K} \alpha_k + 1}{\eta_k + |\lambda_p^{(0)}|} |\lambda_p^k|,$$

where $w_p^{(0)} = \Psi^{(0)^{-1/2}} \tilde{\lambda}_p^{(0)}$, pseudo design matrix $X^{(0)} = \Psi^{(0)^{1/2}}$, and $\Sigma^{(0)}$ is obtained from (??) using $\lambda^{(0)}$. We denote the true value of $\lambda_p$ as $\lambda_p^*$ and define $u_p = (u_{p1}, \ldots, u_{pk}, \ldots, u_{pK})^T$ and

$$V(u_p) = \frac{N}{2} \| w_p^{(0)} - X^{(0)} (\lambda_p^* + \frac{u_p}{\sqrt{N}}) \|^2 + \frac{\sum_{k=1}^{K} \alpha_k + 1}{\eta_k + |\lambda_p^{(0)}|} |\lambda_p^k + \frac{u_p^k}{\sqrt{N}}|,$$

where vectors are added component-wise. By substituting $u_p^k = 0$ for $k = 1, \ldots, K$ in (9),

$$V(0) = \frac{N}{2} \| w_p^{(0)} - X^{(0)} \lambda_p^* \|^2 + \frac{\sum_{k=1}^{K} \alpha_k + 1}{\eta_k + |\lambda_p^{(0)}|} |\lambda_p^*|,$$

and

$$V(u_p) - V(0) = \frac{1}{2} u_p^T \Psi^{(0)} u_p - \frac{\sqrt{N}(\tilde{\lambda}_p^{(0)} - \Psi^{(0)^{1/2}} \lambda_p)}{\sigma_p^{2(0)}}$$

$$+ \frac{\sum_{k=1}^{K} \alpha_k + 1}{\eta_k + |\lambda_p^{(0)}|} \left( |\lambda_p^* - \lambda_p| - |\lambda_p^*| \right)$$

$$= T_1 + T_2 + \sum_{k=1}^{K} T_{3k}.$$
Asymptotic normality and consistency of \( \Lambda^{lla} \)

Let \( \hat{u}_p = \arg\min_{u_p} (V(u_p) - V(0)) \), then \( \Lambda^{lla}_p = \lambda^*_p + \hat{u}_p / \sqrt{N} \). When \( N \to \infty \), \( S_{yy} \longrightarrow \Omega^* = \Lambda^* \Lambda^*^T + \Sigma^* \) element-wise in probability. If we represent \( \Omega(0) = \Lambda(0) \Lambda(0)^T + \Sigma, \Gamma(0) = \Omega(0)^{-1} \Lambda(0) \), and \( \Psi(0) = I_K - \Lambda(0)^T \Omega(0)^{-1} \Lambda(0) \), then \( \Psi(0) \)

\[
= \Delta(0) + \Lambda(0)^T S_{yy} \Lambda(0)
\]

\[
= I_k + \Lambda^*^T \left( \Omega(0)^{-1} S_{yy} \Omega(0)^{-1} - \Omega(0)^{-1} \right) \Lambda^* + \frac{u^T}{\sqrt{N}} \left( \Omega(0)^{-1} S_{yy} \Omega(0)^{-1} - \Omega(0)^{-1} \right) \Lambda^* + \Lambda^*^T \left( \Omega(0)^{-1} S_{yy} \Omega(0)^{-1} - \Omega(0)^{-1} \right) \frac{u}{\sqrt{N}}.
\]

Furthermore, \( \Omega(0)^{-1} = \Omega^*^{-1} + o_P(1) \) by continuous mapping theorem and \( S_{yy} = \Omega^* + o_P(1) \); therefore,

\[
\Psi(0) = I_k + o_P(1) \left( \Lambda^*^T \Lambda^* + \frac{u^T}{\sqrt{N}} \Lambda^* + \Lambda^*^T \frac{u}{\sqrt{N}} + \frac{u^T u}{N} \right),
\]

which in turn implies that \( \Psi(0) = I_k + o_P(1) \) because \( \Lambda^* \) is fixed and \( u_{pk} = o_P(1) \) for \( p = 1, \ldots, P \) and \( k = 1, \ldots, K \). Then arguments similar to Theorem 2 of Zou (2006) and Theorem 5 of Zou & Li (2008) imply that

\[
T_1 = \frac{1}{2} \frac{u^T}{\sigma_{pp}^2(0)} \Psi(0)^{pp} u_p \longrightarrow \frac{1}{2} \frac{u_p^T u_p}{\sigma_{pp}^2(0)} ; \quad T_2 = \frac{\sqrt{N} u_p^T (\hat{\lambda}(0)^p - \Psi(0)^p \lambda^*_p)}{\sigma_{pp}^2(0)} \longrightarrow \frac{u_p^T l_p}{\sigma_{pp}^2(0)}
\]

respectively in probability and distribution, where \( l_p \sim \mathcal{N}(0, I_K) \).
Asymptotic normality and consistency of $\Lambda^{ll_a}$

If $\lambda^{*}_{pk} \neq 0$ and $N \to \infty$, then $\lambda^{(0)}_{pk} \to \lambda^{*}_{pk}$ in probability due to $\sqrt{N}$-consistency of $\Lambda^{(0)}$, $\frac{\alpha_k+1}{\sqrt{N}} \to 0$ due to Assumption (A2); $(\eta_k + \lambda^{(0)}_{pk}) \to \lambda^{*}_{pk}$ in probability due to Assumption (A2), $\sqrt{N}$-consistency of $\Lambda^{(0)}$, and continuous mapping theorem; and $\sqrt{N}(|\lambda^{*}_{pk} + \frac{u_{pk}}{\sqrt{N}}| - |\lambda^{*}_{pk}|) = \text{sign}(\lambda^{*}_{pk})u_{pk}$. By Slutsky’s and continuous mapping theorems,

$$T_{3k} = \frac{\alpha_k+1}{\sqrt{N}} \frac{1}{(\eta_k + |\lambda^{(0)}_{pk}|)} \sqrt{N}(|\lambda^{*}_{pk} + \frac{u_{pk}}{\sqrt{N}}| - |\lambda^{*}_{pk}|) \to 0$$

in probability. If $\lambda^{*}_{pk} = 0$ and $N \to \infty$, then

$$T_{3k} = \frac{\alpha_k+1}{\sqrt{N}(\eta_k + |\lambda^{(0)}_{pk}|)} \sqrt{N}(|\lambda^{*}_{pk} + \frac{u_{pk}}{\sqrt{N}}| - |\lambda^{*}_{pk}|) \to \begin{cases} 0 & \text{if } u_{pk} = 0, \\ \infty & \text{otherwise} \end{cases}$$

in probability due to similar reasons as those for $\lambda^{*}_{pk} \neq 0$. Again, by Slutsky’s theorem, $V(u_{p}) - V(0) \to V^{*}(u_{p})$ in distribution (11), where

$$V^{*}(u_{p}) = \begin{cases} \frac{u^{T}_{p} p_{N} |\lambda^{*}_{pN}| u_{p}}{2\sigma_{2(0)}^{2} p_{p}} - \frac{u^{T}_{p} p_{N} l_{p} p_{N}}{\sigma_{2(0)}^{2} p_{p}} & \text{if } u_{pk} = 0 \text{ for all } k \notin \mathcal{A} p_{N}, \\ \infty & \text{otherwise.} \end{cases}$$

$V(u_{p}) - V(0)$ is convex, and the unique minimum of $V^{*}(u_{p})$ is $(l_{p} p_{N}, 0)$ for $p = 1, \ldots, P$. Following the epi-convergence results of Geyer (1994) and Knight & Fu (2008), $u_{p} p_{N} \to l_{p} p_{N}$ in distribution and $u_{p} c_{pN} \to 0$ in distribution for $p = 1, \ldots, P$. Because $l_{p} p_{N}$ is distributed as $N(0, I_{|\mathcal{A} p_{N}|})$, the asymptotic normality of non-zero loadings is proved.
Asymptotic normality and consistency of $\Lambda^{lla}$

For $p = 1, \ldots, P$ all $k \in A_{PN}$, $\Lambda^p \xrightarrow{p} \lambda^*$ in probability; therefore, for all $k \in A_{PN}$, $\mathbb{P}\{ k \mid k \in A^*_p \} \to 1$ in probability for $p = 1, \ldots, P$. To prove the consistency of $\Lambda^{lla}$, we finally show that for all $k' \in A_{PN}$, $\mathbb{P}\{ k' \mid k' \in A^*_p \} \to 0$ in probability for $p = 1, \ldots, P$. For any $p$, assume that $k' \in A^*_p$, then the necessary conditions of KKT optimality and the differential of $V(u_p)$ (9) with respect to $\lambda^p$ imply that $N X^{(0)}_{k'}^T (w_p^{(0)} - X^{(0)} \lambda^p) = (\alpha_{k'} + 1)/(\eta_{k'} + |\lambda^{(0)}_{pk'}|)$. Using (16), we know that $(\alpha_{k'} + 1)/(|k' + |\lambda^{(0)}_{pk'}|) 1/\sqrt{N} \to \infty$ in probability, and

$$\sqrt{N} X^{(0)}_{k'}^T (w_p^{(0)} - X^{(0)} \lambda^p) = X^{(0)}_{k'}^T X^{(0)} \sqrt{N}(\lambda^* - \lambda^p) + \sqrt{N}(X^{(0)}_{k'}^T w_p^{(0)} - X^{(0)}_{k'}^T X^{(0)} \lambda^*)$$  \hspace{1cm} (18)

The first and second terms on the right hand side of (18) are asymptotically normal respectively due to the asymptotic normality of non-zero $\lambda^{lla}_{pk'}$’s and (14). By Slutsky’s theorem, the left hand side of (18) is also asymptotically normal; therefore, for $p = 1, \ldots, P$,

$$\mathbb{P}\{ k' \in A^*_p \} \leq \mathbb{P}\left\{ k' \mid X^{(0)}_{k'}^T (w_p^{(0)} - X^{(0)} \lambda^p) = \frac{\alpha_{k'} + 1}{\eta_{k'} + |\lambda^{(0)}_{pk'}|} \right\} \to 0$$  \hspace{1cm} (19)

in probability, which further implies that $k' \in A_{PN} \implies \mathbb{P}\{ k' \in A^*_p \} \to 0$ in probability. This proves the consistency of $\lambda_{pk}$'s.
Asymptotic normality and consistency of $\Sigma^{lla}$

For the $\sqrt{N}$-consistent sequence of estimators $\Lambda_p^{(0)}$, Assumption (A2) and continuous mapping theorem imply that if $N \to \infty$, then $\hat{\Lambda}^{(0)} = (\Omega^* + o_P(1))(\Omega^*^{-1} \Lambda^* + o_P(1))$, $\lambda_p^{(0)} = \lambda^* + o_P(1)$, and

$$
\sigma_{pp}^{2lla} = \lambda_p^* T \lambda_p^* + o_P(1) - 2 \lambda_p^* T \lambda_p^* + o_P(1) + (\Omega^*)_{pp} + o_P(1) = -\lambda_p^* T \lambda_p^* + (\Omega^*)_{pp} + o_P(1) = \sigma_{pp}^{2}\lambda^* + o_P(1),
$$

which proves the consistency of $\sigma_{pp}^{2lla}$ and that of $\Sigma^{lla}$.

The asymptotic normality of $\Sigma^{lla}$ follows from Theorem 5.21 of van DerVaart (2000) due to the following two reasons. First, if $\sigma_{pp}^{2} > 0$, then differential of the objective is convex in $1/\sigma_{pp}^{2}$ and its derivatives are continuous and bounded. This implies that the differential is locally Lipschitz in $1/\sigma_{pp}^{2}$ with a square integrable Lipschitz constant. Second, $\Sigma^{lla} \to \Sigma^*$ in probability.
Laplace approximation at any grid point \((\rho_g, \delta_g)\) requires calculation of

\[
\frac{\partial^2 \log p(Y, \Lambda | \theta^g)}{\partial \text{vec}(\Lambda^T) \partial \text{vec}(\Lambda^T)^T} = \frac{\partial^2 \log p(Y | \Lambda, \theta^g)}{\partial \text{vec}(\Lambda^T) \partial \text{vec}(\Lambda^T)^T} + \frac{\partial^2 \log p(\Lambda | \rho_g, \delta_g)}{\partial \text{vec}(\Lambda^T) \partial \text{vec}(\Lambda^T)^T},
\]

for the non-zero entries of \(\Lambda\). The left hand side of (21) is obtained using the right hand side. While the first term term \(\log p(Y | \Lambda, \Lambda^g, \Sigma^g)\) in the right hand side is differentiable, the term \(\log p(\Lambda | \rho_g, \delta_g)\) is not. This problem is resolved by exploiting the Gaussian scale mixture representation of the multiscale generalized double Pareto prior \(p(\Lambda | \rho_g, \delta_g)\) at \(\theta^g\) so that

\[
\log p(\lambda_p | \rho_g, \delta_g) \propto - \sum_{k \in \mathcal{A} \cap p_N} \frac{\lambda^2_{pk}}{2} \frac{\alpha^g_k + 1}{|\lambda^g_{pk}|^2 (\eta^g_k + |\lambda^g_{pk}|)},
\]

where \(\lambda_p\) represents \(\lambda_{pk}\) with \(k \in \mathcal{A} \cap p_N\) to ease notation.
Differentiating (22) twice with respect to $\lambda_p$ yields

$$-\frac{\partial^2 \log p(\lambda_p \mid \rho_g, \delta_g)}{\partial \lambda_p \partial \lambda_p^T} = \text{diag} \left( \frac{\alpha_{k_1}^g + 1}{|\lambda_{pk_1}^g + |\lambda_{pk_1}^g|}, \ldots, \frac{\alpha_{kp}^g + 1}{|\lambda_{pkp}^g + |\lambda_{pkp}^g|} \right) \equiv D_p^g \tag{23}$$

for $k_1, \ldots, kp \in \mathcal{A}_N^p$ and $p = 1, \ldots, P$. The analytic form in (22) can be also obtained using local quadratic approximation, which has been used previously in penalized variable selection. For $p = 1, \ldots, P$, the $p$th block of the information matrix for $\text{vec}(\Lambda^T)$ evaluated at $\theta^g$

$$H_p^g = -\frac{\partial^2 \log p(Y, \Lambda \mid \theta^g)}{\partial \lambda_p \partial \lambda_p^T} = N\Psi_g \sigma_{pp}^2 + D_p^g; \tag{24}$$

therefore, the overall information matrix for $\text{vec}(\Lambda^T)$

$$H = -\frac{\partial^2 \log p(Y, \Lambda \mid \theta^g)}{\partial \text{vec}(\Lambda^T) \partial \text{vec}(\Lambda^T)^T} = \text{bdiag}(H_1, \ldots, H_p, \ldots, H_P), \tag{25}$$

where bdiag is the block diagonal operator. Using (25), the analytic form of $\log p(Y \mid \mathcal{M}_g)$ follows from the standard definition of Laplace approximation

$$\log \pi_g \propto p(Y \mid \mathcal{M}_g) = \log p(Y \mid \Lambda^g, \Sigma^g) + \log p(\Lambda^g) + \frac{\log 2\pi}{2} \sum_{p=1}^{P} |\lambda_{PN}^g| - \frac{1}{2} \sum_{p=1}^{P} \log \left| \frac{N\Psi_g}{\sigma_{pp}^2} + D_p^g \right|. \tag{26}$$
Model Selection Consistency of xFA

Assume that \( p(\mathcal{M}_g) = p_g \), where \( 0 < p_g < 1 \) and \( \sum_{g=1}^{G} p_g = 1 \), then

\[
\pi_g = \frac{p(Y | \mathcal{M}_g)p(\mathcal{M}_g)}{\sum_{g=1}^{G} p(Y | \mathcal{M}_g)p(\mathcal{M}_g)} \implies \pi_g = \frac{p(Y | \mathcal{M}_g)p_g}{\sum_{g=1}^{G} p(Y | \mathcal{M}_g)p_g}.
\]  

(27)

If index \( g^* \) corresponds to the true model, then

\[
\pi_{g^*} = p(\mathcal{M}_{g^*} | Y) = \left(1 + \sum_{g=1, g \neq g^*}^{G} \frac{p(Y | \mathcal{M}_g)p_g}{p(Y | \mathcal{M}_{g^*})p_{g^*}}\right)^{-1}.
\]  

(28)

To prove the model selection consistency of expandable factor analysis, we need to show that when \( N \to \infty \),

\[
\pi_{g^*} \to 1 \iff \frac{p(Y | \mathcal{M}_g)}{p(Y | \mathcal{M}_{g^*})} \to 0
\]

(29)

in probability for \( g \neq g^* \); therefore, it is enough to show that \( \log p(Y | \mathcal{M}_{g^*}) - \log p(Y | \mathcal{M}_g) \to \infty \) in probability as \( N \to \infty \).

The overfitted expandable factor analysis model that corresponds to grid index \( g \) has \( \mathcal{A}_{pN}^{g^*} \subset \mathcal{A}_{pN}^{g} \) and its number of factors \( K > K^* \); therefore,

\[
\log p(Y | \mathcal{M}_{g^*}) - \log p(Y | \mathcal{M}_g) = \log \frac{p(Y | \mathcal{M}_{g^*}, \Sigma^{g^*})}{p(Y | \mathcal{M}_g, \Sigma^g)} + \log \frac{p(\Lambda^{g^*})}{p(\Lambda^g)} + \frac{1}{2} \sum_{p=1}^{P} \log \left| \frac{N\Psi^g + D^g}{\sigma^g_{pp}} \right| - \log \frac{2\pi}{2} \sum_{p=1}^{P} \left( | \mathcal{A}_{pN}^{g^*} | - | \mathcal{A}_{pN}^{g} | \right)
\]

\[
\equiv T_1 + T_2 + T_3 + T_4.
\]  

(30)
Model Selection Consistency of xFA (bound $T_1$ and $T_4$)

The log likelihood of expandable factor analysis depends on the data only through $S_{yy}$, so it is bounded for any $\Lambda^g, \Sigma^g$ $g = 1, \ldots, G$. Also, $\Omega^g$ is estimated as a positive definite matrix, so $L_1 \leq |T_1| \leq L_2$, where $L_1$ and $L_2$ are fixed constants.

Since $\Lambda^g$ and $\Lambda^g\ast$ are estimated using $\sqrt{N}$-consistent estimators of $\Lambda^\ast$, so

$$
\sum_{p=1}^{P} \left( |\omega^g_{pN} \ast - \omega^g_{pN} | \right) = P(K^\ast - K) \sigma \left( \frac{1}{\sqrt{N}} \right) = o_P(1).
$$

(31)
The form of the multiscale generalized double Pareto prior implies that

\[ T_2 = - \sum_{p=1}^{P} \sum_{k \in \mathcal{A}_p^g \setminus \mathcal{A}_p^g^*} \log \frac{\alpha_{kN}^g}{2 \eta_{kN}^g} + \sum_{p=1}^{P} \sum_{k \in \mathcal{A}_p^g^*} \log \frac{\alpha_{kN}^g^* \eta_{kN}^g^*}{\alpha_{kN}^g \eta_{kN}^g} + \]

\[ \sum_{p=1}^{P} \left\{ \sum_{k \in \mathcal{A}_p^g} (\alpha_{kN}^g + 1) \log \left( 1 + \frac{|\lambda_{pN}^g|}{\eta_{kN}^g} \right) - \sum_{k \in \mathcal{A}_p^g^*} (\alpha_{kN}^g^* + 1) \log \left( 1 + \frac{|\lambda_{pN}^g^*|}{\eta_{kN}^g} \right) \right\} \] (32)

Because \( \mathcal{M}_g \) is overfitted in the number of factors, \( \alpha_{kN}^g^* > \alpha_{kN}^g \) for all \( k \), \( \eta_{kN}^g^* \leq \eta_{kN}^g \) for all \( k \), and \( \mathcal{A}_p^g^* \subset \mathcal{A}_p^g \) for all \( p \). Using this, (32) reduces to

\[ T_2 \geq - \sum_{p=1}^{P} \sum_{k \in \mathcal{A}_p^g \setminus \mathcal{A}_p^g^*} \left\{ \log \frac{\sqrt{N} \alpha_{kN}^g}{2 \sqrt{N} \eta_{1N}^g} + (\alpha_{kN}^g + 1) \log \left( 1 + \frac{\sqrt{N} |\lambda_{pN}^g|}{\sqrt{N} \eta_{1N}^g} \right) \right\} \]

\[ + \sum_{p=1}^{P} \sum_{k \in \mathcal{A}_p^g^*} (\alpha_{1N}^g + 1) \log \left( \frac{\sqrt{N} \eta_{kN}^g^* + \sqrt{N} |\lambda_{pN}^g|}{\sqrt{N} \eta_{kN}^g + \sqrt{N} |\lambda_{pN}^g^*|} \right). \] (33)
Model Selection Consistency of xFA (bound for $T_2$)

Assumptions imply that $\alpha g kN / \sqrt{N} \rightarrow 0$ and $\sqrt{N} \eta g kN \rightarrow c_k > 0$ for all $k$ and $g$ as $N \rightarrow \infty$; that $\lambda pk = \lambda^* pk + o_P(1)$, $\lambda^* pk = \lambda^* pk + o_P(1)$; and that $\left\{ \frac{\alpha g kN}{\eta g kN} \right\}_{k=1}^{\infty}$ is an increasing sequence for all $g$. Therefore, (33) reduces to

$$T_2 \geq \left\{ \log \frac{\sqrt{N} \alpha g K_N}{2c_1 + o_P(1)} + (\alpha g K_N + 1) \log \left( 1 + \frac{\Theta_P(1)}{c_1} \right) + \sum_{p=1}^{P} \left| \mathcal{A}^g p N \right| - \left| \mathcal{A}^g^* p N \right| \right\}$$

$$+ (\alpha g kN + 1) \left( \sum_{p=1}^{P} \left| \mathcal{A}^g^* p \right| + o_P(1) \right)$$

$$= -\left\{ \log \sqrt{N} + \log \alpha g K_N + \alpha g K_N \log \left( 1 + \frac{\Theta_P(1)}{c_1} \right) + \Theta_P(1) \right\} P(K - K^*) \Theta_P \left( \frac{1}{\sqrt{N}} \right)$$

$$+ (\alpha g kN + 1) \left( \sum_{p=1}^{P} \left| \mathcal{A}^g^* p \right| + o_P(1) \right)$$

$$= (\alpha g kN + 1) \text{const.} + o_P(1) \longrightarrow \infty$$

in probability as $N \longrightarrow \infty$, where const. > 1.
Model Selection Consistency of xFA (bound for $T_3$)

We use the assumption that $k$th eigen value of $\Psi^g \epsilon^*_g > 0$ for all $k$ and $\sigma^2_{pp} > 0$ for all $p$. Then,

$$2T_3 = \sum_{p=1}^{P} \log \left| \frac{N\Psi^g}{\sigma^2_{pp}} + D^g_p \right|$$

$$= \sum_{p=1}^{P} \left( |\omega^g_{pN}| - |\omega^g_{pN}^*| \right) \log N + \sum_{p=1}^{P} \sum_{k \in \mathcal{A}_{pN}^g \setminus \mathcal{A}^g_{pN}^*} \log \left( \frac{e^g_k}{\sigma^2_{pp}} + \frac{(\alpha^g_{kN} + 1)}{|\lambda^g_{pK}^*| (\sqrt{N} \eta_{kN}^g + \sqrt{N} |\lambda^g_{pK}|)} \right)$$

$$\sum_{p=1}^{P} \sum_{k \in \mathcal{A}_{pN}^g} \left\{ \log \left( \frac{e^g_k}{\sigma^2_{pp}} + \frac{(\alpha^g_{kN} + 1) / \sqrt{N}}{|\lambda^g_{pK}| (\sqrt{N} \eta_{kN}^g + \sqrt{N} |\lambda^g_{pK}|)} \right) - \log \left( \frac{e^g_k}{\sigma^2_{pp}} + \frac{(\alpha^g_{kN}^* + 1) / \sqrt{N}}{|\lambda^g_{pK}^*| (\sqrt{N} \eta_{kN}^g + \sqrt{N} |\lambda^g_{pK}^*|)} \right) \right\}. \quad (35)$$

There exists $d_{pk} > 0$ such that $|\lambda_{pk}| = d_{pk} / \sqrt{N}$ for $k \in \mathcal{A}_{pN}^g \setminus \mathcal{A}^g_{pN}$ and $|\mathcal{A}^g_{pN}| = |\mathcal{A}^g_{pN}^*| \geq 1$ for all $p$. Therefore,

$$2T_3 \geq P \left\{ \log N + \sum_{k \in \mathcal{A}_{pN}^g \setminus \mathcal{A}^g_{pN}^*} \log \left( \frac{e^g_k}{\sigma^2_{pp}} + \frac{\sqrt{N}(\alpha^g_{1N} + 1)}{d_{pk}(c_K + d_{pk} / \sqrt{N})} \right) + \sum_{p=1}^{P} \sum_{k \in \mathcal{A}_{pN}^g} \log \left( 1 + \frac{\sigma^2_{pp} (1)}{e^g_k |\lambda^*_{pK}| (c_K + \sqrt{N} |\lambda^*_{pK}|)} \right) \right\} + o_P(1)$$

$$= \Theta(\log N) + o_P \left( \log \alpha^g_{kN} \right) + o_P(1) \longrightarrow \infty \quad (36)$$

in probability as $N \longrightarrow \infty$. Therefore, $\log p(Y | \mathcal{M}_{g}^* ) - \log p(Y | \mathcal{M}_{g}) \longrightarrow \infty$ in probability as $N \longrightarrow \infty$. 

---

Model Selection Consistency Appendix 56