Notes on Statistical and Machine Learning

Factor Analysis

Chapter: *26* **Prepared by:** *Chenxi Zhou*

This note is prepared based on

- *Chapter 15, Latent Variable Models for Blind Source Separation* in Izenman (2009), and
- *Chapter 14, Unsupervised Learning* in Hastie, Tibshirani, and Friedman (2009).

I. Introduction

1. **Overview:** In *factor analysis model*, a set of observed continuous variables is explained by a linear combination of a much smaller set of continuous latent variables, called *factors*.

2. **Factor Analysis Model:** Let $X = (X_1, X_2, \cdots, X_p)^\top \in \mathbb{R}^p$ be a real-valued random vector that we can observe. The factor analysis model takes on the form of

$$X = \mathbf{A}S + \varepsilon, \tag{1}$$

where

(a) $S = (S_1, S_2, \cdots, S_m)^\top \in \mathbb{R}^m$ and $S_1, S_2, \cdots, S_m$ are m unobservable random variables called *latent variables* or *common factors* and $m \le p$,

(b) $\mathbf{A} \in \mathbb{R}^{p\times m}$ is a mixing matrix of full rank containing unknown coefficients called the *factor loadings*, and

(c) $\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_p)^\top \in \mathbb{R}^p$ and $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_p$ are unobservable random variables that are called *specific* (or *unique*) factors because ε_j *only* appears in the equation involving X_j.

Let the (j, i)-th entry of $\mathbf{A}$ be $a_{j,i}$ for all $j = 1, 2, \cdots, p$ and $i = 1, 2, \cdots, m$. We can rewrite (1) by the following system of linear equations:

$$X_j = a_{j,1}S_1 + a_{j,2}S_2 + \cdots + a_{j,m}S_m + \varepsilon_j, \qquad \text{for all } j = 1, \cdots, p. \tag{2}$$

3. **Assumptions:** We assume the following:

(a) Each of the p observed random variables $X_1, X_2, \cdots, X_p$ has been standardized to have zero mean and unit variance;

(b) The relationships between the observed variables, $X_1, X_2, \cdots, X_p$, are explained *only* by the underlying common factors and *not* by the errors;

(c) The common factors have mean zero and unit variance, and are uncorrelated, that is,

$$\mathbb{E}[S] = \mathbf{0}_m \qquad \text{and} \qquad \mathrm{Var}[S] = \mathbf{I}_m;$$

(d) The random error term ε has zero mean and a diagonal covariance matrix, $\mathrm{Var}[\varepsilon] = \mathbf{\Psi}$, with positive diagonal entries;

(e) S and ε are independent so that $\mathbb{E}[S\varepsilon^\top] = \mathbf{0}_{m\times p}$.

4. **Goal:** The goal of the factor analysis is to estimate $\mathbf{A}$ and recover S.

5. **Moments of X:** From (1) and the assumptions, we obtain

$$\mathbb{E}[X] = \mathbf{0}_p, \tag{3}$$

and

$$\mathbf{\Sigma}_{XX} := \mathrm{Var}[X] = \mathbf{A}\mathbf{A}^\top + \mathbf{\Psi}. \tag{4}$$

Since X_j has unit variance for all $j = 1, 2, \cdots, p$, the j-th diagonal element of $\mathbf{\Sigma}_{XX}$ is

$$1 = h_j^2 + \psi_{j,j}, \tag{5}$$

where $h_j^2 = \sum_i a_{j,i}^2$ is called the *communality* and $\psi_{j,j}$ is called the *uniqueness* given by the j-th diagonal element of $\mathbf{\Psi}$.

6. **Orthogonal and Oblique Factors:** The common factors, $\{S_j\}$, are called *orthogonal* if they are pairwise uncorrelated, and are called *oblique* if they are correlated.

III. Principal Components Factor Analysis

1. **Goal:** Without making any distributional assumption for the sources S in (1), we determine $\mathbf{A}$ using a least-squares approach.

2. **Model Transformation:** Premultiplying (1) by the Moore-Penrose generalized inverse of $\mathbf{A}$,

$$\mathbf{B} = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top, \tag{6}$$

where $(\mathbf{A}^\top\mathbf{A})^{-1}$ exists since $\mathbf{A} \in \mathbb{R}^{p\times m}$ has full rank and $p \geq m$, we obtain

$$\mathbf{B}X = \mathbf{B}\mathbf{A}S + \mathbf{B}\varepsilon = S + \mathbf{B}\varepsilon,$$

that is,

$$S = \mathbf{B}X - \mathbf{B}\varepsilon.$$

Substituting the preceding equation back into (1), we have

$$X = \mathbf{A}(\mathbf{B}X - \mathbf{B}\varepsilon) + \varepsilon,$$

or equvivalently,

$$X = \mathbf{C}X + E, \tag{7}$$

where $\mathbf{C} = \mathbf{AB}$ has rank m, $\mathbf{A}$ and $\mathbf{B}$ are full-rank matrices each of rank m, $E = (\mathbf{I}_p - \mathbf{C})\varepsilon$, and X and E both have mean zero.

3. **Principal Components Factor Analysis:** Assume $\mathbf{\Sigma}_{XX}$ is *known*. The model (7) is the *multivariate reduced-rank regression model* corresponding to principal component analysis. To obtain $\mathbf{A}$ and $\mathbf{B}$, we use the least-squares criterion and minimize

$$\mathbb{E}\Big[(X - \mathbf{AB}X)^\top (X - \mathbf{AB}X)\Big]. \tag{8}$$

The minimum is obtained by setting

$$\mathbf{A} = (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_m) = \mathbf{B}^\top,$$

where $\mathbf{v}_j$'s is the eigenvector associated with the j-th largest eigenvalue of $\mathbf{\Sigma}_{XX}$.

This approach is referred to as the *principal components* approach to the factor analysis.

Remark. The principal components approach essentially ignores the matrix $\mathbf{\Psi}$.

4. **Factor Scores:** The rows of $\mathbf{B}$ give the coefficients of the m principal components scores, $\mathbf{v}_j^\top X$, for all $j = 1, 2, \cdots, m$, and the eigenvalues of $\mathbf{\Sigma}_{XX}$ measure the variance (or power) of the m sources.

5. **Estimation:** Typically, $\mathbf{\Sigma}_{XX}$ is unknown. We estimate it from the standardized input data by $\widehat{\mathbf{\Sigma}}_{XX}$, the sample correlation matrix. Estimates of $\mathbf{A}$ and $\mathbf{B}$ are given by

$$\widehat{\mathbf{A}} = (\hat{\mathbf{v}}_1, \cdots, \hat{\mathbf{v}}_m) = \widehat{\mathbf{B}}^\top, \tag{9}$$

respectively, where $\hat{\mathbf{v}}_j$ is the eigenvector corresponding to the j-th largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{XX}$, for all $j = 1, 2, \cdots, m$.

6. **Choice of m:** We discuss how to determine the value of m, the number of common factors. Because the p eigenvalues of $\widehat{\mathbf{\Sigma}}_{XX}$ sum to p, i.e., the trace of $\widehat{\mathbf{\Sigma}}_{XX}$, a popular decision rule is that m should be taken to be the number of those sample eigenvalues that are greater than unity.

7. **Estimates of Factor Scores:** The m-vector of estimated factor scores corresponding to a standardized sample observation $\mathbf{x} = (x_1, x_2, \cdots, x_p)^\top$ is given by

$$\hat{\mathbf{f}} = \widehat{\mathbf{B}}\mathbf{x} = (\hat{\mathbf{v}}_1^\top \mathbf{x}, \hat{\mathbf{v}}_2^\top \mathbf{x}, \cdots, \hat{\mathbf{v}}_m^\top \mathbf{x})^\top, \tag{10}$$

Remark. If we have n observations, $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n \sim X$, we can plot the first two estimated factor scores,

$$(\hat{\mathbf{v}}_1^\top \mathbf{x}_i, \hat{\mathbf{v}}_2^\top \mathbf{x}_i), \qquad \text{for all } i = 1, 2, \cdots, n,$$

on a scatterplot, using which we can identify outliers.

8. **Factor Indeterminacy:** Let $\mathbf{T} \in \mathbb{R}^{m\times m}$ be an orthogonal matrix. Then, we have

$$\mathbf{C} = (\mathbf{A}\mathbf{T})(\mathbf{T}^\top \mathbf{B}).$$

   Thus, we can only determine $\mathbf{A}$ up to a rotation. This is called the *factor indeterminacy.*

   We choose $\mathbf{T}$ to have certain desirable properties:

   (a) We require certain elements of $\widehat{\mathbf{A}}\mathbf{T}$ to be zero; or

   (b) *Varimax Rotation:* We seek to find an orthogonal transformation $\mathbf{T}$ to maximize

$$\sum_{k=1}^{m} \left[\sum_{j=1}^{p} \tilde{a}_{j,k}^4 - \frac{1}{p} \left(\sum_{j=1}^{p} \tilde{a}_{j,k}^2 \right)^2 \right],$$

   where $\tilde{a}_{j,k}$ is the (j,k)-th entry of the matrix $\mathbf{A}\mathbf{T}$.

9. **Principal Factor Method:**

   (a) *Main Idea:* Principal factor method is a modification of the principal components method that takes the diagonal matrix $\mathbf{\Psi}$ into account. We replace the correlation matrix $\mathbf{\Sigma}_{XX}$ by $\mathbf{\Sigma}_{XX} - \mathbf{\Psi}$ so that we have the communalities $\{h_j^2\}_{j=1}^p$ on the diagonal.

   (b) *Estimation:* Since $\mathbf{\Psi}$ is unknown, we estimate $\{h_j^2\}_{j=1}^p$. We estimate h_j^2 by the squared multiple correlation between X_j and the remaining $p-1$ variables as

$$\hat{h}_j^2 = 1 - \frac{1}{r_{j,j}}, \qquad \text{for all } j = 1, 2, \cdots, p, \tag{11}$$

   where $r_{j,j}$ is the j-th diagonal element of the inverse of the sample correlation matrix.

   (c) *Caution:* The matrix $\widehat{\mathbf{\Sigma}}_{XX} - \widehat{\mathbf{\Psi}}$ is *not* necessarily be positive-definite, so that its eigenvalues can be both positive and negative. Because the sum of the positive eigenvalues exceeds the sum of the communalities, the number of factors, m, is usually taken to be at most the maximum number of positive eigenvalues whose sum is less than trace$(\widehat{\mathbf{\Sigma}}_{XX} - \widehat{\mathbf{\Psi}})$.

10. **Comparison between Factor Analysis and Principal Component Analysis (PCA):**

   (a) *Similarities:*

   - They both aim to reduce the dimensionality of a vector of random variables.
   - They both attempt to represent some aspect of the covariance matrix or the correlation matrix as well as possible.
   - The results of the principal factor method are equivalent to those of the PCA if all non-zero elements of $\mathbf{\Psi}$ are identical. More generally, the coefficients found from PCA and the loadings found from *orthogonal* factor analysis are often very similar.

(b) *Differences:*

- On the model:
  - Factor analysis has a definite model (1), but
  - PCA does *not*.
- On the covariance or correlation matrix:
  - PCA concentrates on the diagonal elements of the covariance or correlation matrix, but
  - Factor analysis focuses more on the off-diagonal elements, by noting the common factor term $\mathbf{A}S$ in (4) accounts completely for the off-diagonal elements.
- On the number of dimensionality m:
  - If any individual variables are almost independent of all others, there will be a principal component corresponding to each such variable;
  - A common factor in factor analysis must contribute to at least two of the variables, so it is not possible to have a "single variable" common factor.

  Remark. Any "single variable" factors appear as error terms and do *not* contribute to the dimensionality of the model.
- On the changes of the dimensionality m: changing m can have much more drastic effects on the factor analysis than it does on PCA. If we increase m from m_1 to m_2,
  - in PCA, additional $m_2 - m_1$ principal components are included, and the original m_1 principal components are *unaffected*;
  - in factor analysis, none of m_2 factors need bear any resemblance to the original m_1 factors.
- On the exact computation:
  - The principal components are exact linear functions of the data vector $\mathbf{x}$;
  - The factors are *not* exact linear functions of $\mathbf{x}$; instead, $\mathbf{x}$ is defined as a linear function of factors apart from an error term.

  Remark. The fact that the expected value of X is a linear function of S need *not* imply that the expected value of S is a linear function of X, unless multivariate normal assumptions are made.

IV. Maximum Likelihood Factor Analysis

1. **Assumptions:** The *maximum likelihood factor analysis* assumes a fully parametric model. We assume

   (a) the m sources in (1) are distributed as multivariate Gaussian, $S \sim \text{Normal}_m(\mathbf{0}_m, \mathbf{I}_m)$,

   (b) the error term ε is also distributed as multivariate Gaussian, $\varepsilon \sim \text{Normal}_p(\mathbf{0}_p, \mathbf{\Psi})$, where $\mathbf{\Psi}$ is diagonal,

   (c) S and ε are independent.

These assumptions imply that X is also multivariate Gaussian,

$$X \sim \text{Normal}_p(\mathbf{0}_p, \mathbf{\Sigma}_{XX}), \tag{12}$$

where $\mathbf{\Sigma}_{XX} = \mathbf{A}\mathbf{A}^\top + \mathbf{\Psi}$.

2. **Data:** Let $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ be n independent observations from X.

3. **Estimation of $\mathbf{\Sigma}_{XX}$:** We estimate $\mathbf{\Sigma}_{XX}$ by the sample covariance matrix $\widehat{\mathbf{\Sigma}}_{XX}$, which has a Wishart distribution:

$$n\widehat{\mathbf{\Sigma}}_{XX} \sim W_p(n, \mathbf{\Sigma}_{XX}).$$

4. **ML Estimation of $\mathbf{A}$ and $\mathbf{\Psi}$:** We estimate $\mathbf{A}$ and $\mathbf{\Psi}$ using the method of maximum likelihood by maximizing the logarithm of the likelihood function

$$\ell(\mathbf{A}, \mathbf{\Psi}) := -\frac{n}{2}\log|\mathbf{A}\mathbf{A}^\top + \mathbf{\Psi}| - \frac{n}{2}\operatorname{trace}\big(\widehat{\mathbf{\Sigma}}_{XX}(\mathbf{A}\mathbf{A}^\top + \mathbf{\Psi})^{-1}\big), \tag{13}$$

where terms that do *not* involve $\mathbf{A}$ or $\mathbf{\Psi}$ have been ignored.

5. **EM Algorithm to Estimate $\mathbf{A}$ and $\mathbf{\Psi}$:** We use the EM algorithm to maximize ℓ. We treat the unobservable source variables as if they were missing data. If S were actually observed with values $\mathbf{s}_1, \cdots, \mathbf{s}_n$, the complete-data likelihood function is the joint density function of $\{\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_n\}$ and $\{\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \cdots, \boldsymbol{\varepsilon}_n\}$,

$$\begin{aligned} L(\mathbf{A}, \mathbf{\Psi}) &:= \prod_{i=1}^{n}\left[\frac{1}{(2\pi)^{p/2}|\mathbf{\Psi}|^{1/2}}\exp\left(-\frac{1}{2}\boldsymbol{\varepsilon}_i^\top\mathbf{\Psi}^{-1}\boldsymbol{\varepsilon}_i\right) \times \frac{1}{(2\pi)^{m/2}}\exp\left(-\frac{1}{2}\mathbf{s}_i^\top\mathbf{s}_i\right)\right] \\ &= \prod_{i=1}^{n}\Bigg[\frac{1}{(2\pi)^{p/2}|\mathbf{\Psi}|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}_i - \mathbf{A}\mathbf{s}_i)^\top\mathbf{\Psi}^{-1}(\mathbf{x}_i - \mathbf{A}\mathbf{s}_i)\right)\times \\ &\qquad\qquad \frac{1}{(2\pi)^{m/2}}\exp\left(-\frac{1}{2}\mathbf{s}_i^\top\mathbf{s}_i\right)\Bigg] \\ &\propto \left(\prod_{j=1}^{p}\psi_{j,j}\right)^{-\frac{n}{2}}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{(x_{i,j} - \mathbf{a}_j\mathbf{s}_i)^2}{\psi_{j,j}} - \frac{1}{2}\sum_{i=1}^{n}\mathbf{s}_i^\top\mathbf{s}_i\right), \end{aligned} \tag{14}$$

where $x_{i,j}$ is the j-th component of $\mathbf{x}_i$, $\mathbf{a}_j$ is the j-th row of $\mathbf{A}$, and $\psi_{j,j}$ is the j-th diagonal element of the diagonal matrix $\mathbf{\Psi}$, for all $j = 1, 2, \cdots, p$.

The log-likelihood function of the complete-data is

$$\ell_c(\mathbf{A}, \mathbf{\Psi}) = -\frac{n}{2}\sum_{j=1}^{p}\log\psi_{j,j} - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{(x_{i,j} - \mathbf{a}_j\mathbf{s}_i)^2}{\psi_{j,j}} - \frac{1}{2}\sum_{i=1}^{n}\mathbf{s}_i^\top\mathbf{s}_i. \tag{15}$$

Then, we have

(a) In the *E-step*, given the observed data $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ and the current values of $\mathbf{A}$ and $\mathbf{\Psi}$, we find the conditional expectation of ℓ_c.

Since the joint distribution of X and S, given $\mathbf{A}$ and $\mathbf{\Psi}$, is $(p+m)$-variate Gaussian, the conditional distribution of S given X is

$$S \mid X, \mathbf{A}, \mathbf{\Psi} \sim \mathrm{Normal}_m(\boldsymbol{\delta} X, \mathbf{D}),$$

where $\boldsymbol{\delta} := \mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top + \mathbf{\Psi})^{-1}$ and $\mathbf{D} = \mathbf{I}_m - \mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top + \mathbf{\Psi})^{-1}\mathbf{A}$.

To find the conditional expectation of ℓ_c, we need to compute the conditional expectations of the following statistics

$$\mathbf{C}_{XX} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top, \qquad \mathbf{C}_{XS} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{s}_i^\top, \qquad \mathbf{C}_{SS} = \frac{1}{n}\sum_{i=1}^n \mathbf{s}_i\mathbf{s}_i^\top.$$

Then, given data $\{\mathbf{x}_i\}_{i=1}^n$ and current values of $\mathbf{A}$, $\mathbf{\Psi}$, we have

$$\begin{aligned}
\mathbb{E}\big[\mathbf{C}_{XX} \mid \{\mathbf{x}_i\}_{i=1}^n, \mathbf{A}, \mathbf{\Psi}\big] &= \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top = \mathbf{C}_{XX}, \\
\mathbb{E}\big[\mathbf{C}_{XS} \mid \{\mathbf{x}_i\}_{i=1}^n, \mathbf{A}, \mathbf{\Psi}\big] &= \mathbf{C}_{XX}\boldsymbol{\delta}^\top =: \widetilde{\mathbf{C}}_{XS}, \\
\mathbb{E}\big[\mathbf{C}_{SS} \mid \{\mathbf{x}_i\}_{i=1}^n, \mathbf{A}, \mathbf{\Psi}\big] &= \boldsymbol{\delta}\mathbf{C}_{XX}\boldsymbol{\delta}^\top + \mathbf{D} =: \widetilde{\mathbf{C}}_{SS}.
\end{aligned}$$

(b) In the *M-step*, we maximize over $\mathbf{A}$ and $\mathbf{\Psi}$, and the resulting maximizers are

$$\widehat{\mathbf{A}} = \widetilde{\mathbf{C}}_{XS}\widetilde{\mathbf{C}}_{SS}^{-1}, \qquad \text{and} \qquad \widehat{\mathbf{\Psi}} = \mathrm{diag}(\mathbf{C}_{XX} - \widetilde{\mathbf{C}}_{XS}\widetilde{\mathbf{C}}_{SS}^{-1}\widetilde{\mathbf{C}}_{XS}^\top).$$

The complete EM algorithm is shown below:

Algorithm 1 EM Algorithm for Maximum Likelihood Factor Analysis

```
Let $\widehat{\mathbf{A}}^{(0)}$ and $\widehat{\boldsymbol{\Psi}}^{(0)}$ be initial guesses for the parameter matrices $\mathbf{A}$ and $\boldsymbol{\Psi}$, respectively;
Compute
    $\mathbf{C}_{XX} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top;$
For $k = 1, 2, \cdots$, iterate between the following two steps:
    (a) E-Step: Compute
        $\widetilde{\mathbf{C}}_{XS}^{(k-1)} = \mathbf{C}_{XX} \boldsymbol{\delta}^{(k-1)^\top},$
        $\widetilde{\mathbf{C}}_{SS}^{(k-1)} = \boldsymbol{\delta}^{(k-1)} \mathbf{C}_{XX} \boldsymbol{\delta}^{(k-1)^\top} + \mathbf{D}^{(k-1)},$
        where
        $\boldsymbol{\delta}^{(k-1)} = \widehat{\mathbf{A}}^{(k-1)^\top} \left( \widehat{\mathbf{A}}^{(k-1)} \widehat{\mathbf{A}}^{(k-1)^\top} + \widehat{\boldsymbol{\Psi}}^{(k-1)} \right)^{-1},$
        $\mathbf{D}^{(k-1)} = \mathbf{I}_m - \boldsymbol{\delta}^{(k-1)} \widehat{\mathbf{A}}^{(k-1)};$
    (b) M-Step: Update the parameter estimates,
        $\widehat{\mathbf{A}}^{(k)} = \widetilde{\mathbf{C}}_{XS}^{(k-1)} [\widetilde{\mathbf{C}}_{SS}^{(k-1)}]^{-1},$
        $\widehat{\boldsymbol{\Psi}}^{(k)} = \mathrm{diag}(\mathbf{C}_{XX} - \widetilde{\mathbf{C}}_{XS}^{(k-1)} [\widetilde{\mathbf{C}}_{SS}^{(k-1)}]^{-1} [\widetilde{\mathbf{C}}_{XS}^{(k-1)}]^\top);$
Stop when convergence has been attained.
```

6. **Critiques of the Maximum Likelihood Factor Analysis (MLFA):**

   (a) MLFA is based upon Gaussian assumptions but has been routinely applied to non-Gaussian or discrete data;

   (b) There may exist many local maxima;

   (c) The log-likelihood function is rotationally invariant in factor space. Therefore, the sources S and the mixing matrix $\mathbf{A}$ in can *only* be defined up to an arbitrary rotation.

V. Independent Factor Analysis

1. **Blind Source Separation Problem:** The *blind source separation* (BSS) problem involves decomposing an unknown mixture of non-Gaussian signals into its independent component signals.

2. **Overview:** Independent factor analysis (IFA) was proposed as an alternative to ICA and factor analysis to deal with the BSS problem. IFA adopts the maximum likelihood factor analysis model but employs arbitrary *non-Gaussian* densities for the factors.

3. **Model Specification:** The model is given by

$$X = \mathbf{A}S + \varepsilon.$$

We make the following assumptions:

(a) The random error term ε has a p-variate Gaussian distribution $\text{Normal}_p(\mathbf{0}_p, \mathbf{\Psi})$, where $\mathbf{\Psi}$ is *not* necessarily diagonal;

(b) Each unobserved source S_j is assumed to be independently distributed according to a non-Gaussian density $q_{S_j}(\,\cdot\,|\,\boldsymbol{\theta}_j)$ characterized by the parameter vector $\boldsymbol{\theta}_j$, for all $j = 1, 2, \cdots, m$.

In this set-up, the collection of parameters is given by $(\mathbf{A}, \mathbf{\Psi}, \boldsymbol{\theta})$, where $\boldsymbol{\theta} := (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_m)$.

4. **Specification of q_{S_j}:** We require each source density, q_{S_j}, is modeled parametrically by an arbitrary mixture of univariate Gaussian densities,

$$q_{S_j}(s_j\,|\,\boldsymbol{\theta}_j) = \sum_{k=1}^{I_j} w_{j,k} \cdot \varphi(s_j\,|\,\mu_{j,k}, \sigma_{j,k}^2),$$

where

- $\varphi(\,\cdot\,|\,\mu_{j,k}, \sigma_{j,k}^2)$ is the density function of a normal random variable with mean $\mu_{j,k}$ and variance $\sigma_{j,k}^2$,
- $w_{j,k} > 0$ is the mixing proportion associated with the k-th component of the j-th source density, for all $k = 1, 2, \cdots, I_j$, satisfying $\sum_{k=1}^{I_j} w_{j,k} = 1$ for all $j = 1, 2, \cdots, m$.

In particular, note that

$$\boldsymbol{\theta}_j = \Big\{(w_{j,k}, \mu_{j,k}, \sigma_{j,k}^2)\,\Big|\,k = 1, \cdots, I_j\Big\}.$$

These parameters can be estimated using the EM algorithm.

5. **Critiques:**

(a) The total number of parameters can grow to be very large;

(b) The IFA procedure by maximizing the log-likelihood function using the EM algorithm is an extremely computationally intensive procedure when there are many sources to be separated;

(c) EM is a slow algorithm that does *not* necessarily converge to a global maximum of the log-likelihood;

(d) It is hard to determine the number of Gaussian components in the mixture for each component and whether such a mixture of Gaussian formulation appears justified.

References

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning*. Vol. 1. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Izenman, Alan J (Mar. 2009). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. en. Springer Science & Business Media.