Feature Selection for Latent Factor Models

Rittwika Kansabanik, Adrian Barbu
Department of Statistics, Florida State University

1. Extended Literature Review

Class-specific feature selection and SNR-based methods have been explored in recent years. Cluster-based pattern discrimination [7] relies on classifiers to compute the similarity between unknown patterns and clusters, which is crucial for feature selection, while our approach operates independently of classifiers. [9] proposes a one-vs-all binary classification technique with traditional feature selectors for class-specific feature selection. SMBA-CSFS[8] employs a sparse model with a two-step strategy. It builds classifier ensembles and incurs higher computational costs. Authors in [6] leverage fuzzy entropy and mutual information to evaluate feature relevance and redundancy. These methods lack scalability and solid theoretical foundations. Our approach stands out using low-rank generative models and SNR, ensuring strong theoretical support for feature recovery and improved scalability using independent class models. [1] integrates Independent Components Analysis and SNR for feature selection via a linear transformation of all features for classification. In contrast, our method calculates SNRs on original features and uses only those selected for classification.

2. Theorems and Proofs

2.1. ELF Parameter Estimation

Theorem 1 *The ELF objective is:*

$$(\hat{\mathbf{W}}_{ELF}, \hat{\mathbf{\Gamma}}_{ELF}) = \underset{(\mathbf{\Gamma},\mathbf{W}),\mathbf{\Gamma}^T\mathbf{\Gamma}=\mathbf{I}_r}{\text{argmin}} \|(\mathbf{X} - \mathbf{\Gamma}\mathbf{W}^T)\mathbf{\Psi}^{-\frac{1}{2}}\|_F^2, \tag{1}$$

without the constraint $\mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}_r$*, (1) is minimized w.r.t* $\mathbf{\Gamma}$ *and* $\mathbf{W}$ *by*

$$\hat{\mathbf{\Gamma}} = \mathbf{X}\mathbf{\Psi}^{-1}\mathbf{W}(\mathbf{W}^T\mathbf{\Psi}^{-1}\mathbf{W})^{-1} \textit{ and } \hat{\mathbf{W}} = \mathbf{X}^T\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}. \tag{2}$$

Proof. Let $l(\mathbf{\Gamma}) = \|(\mathbf{X} - \mathbf{\Gamma}\mathbf{W}^T)\sqrt{\mathbf{\Psi}^{-1}}\|_F^2$. Then

$$\begin{aligned}
\arg\min_{\mathbf{\Gamma}} l(\mathbf{\Gamma}) &= \arg\min_{\mathbf{\Gamma}} \|\mathbf{X}\sqrt{\mathbf{\Psi}^{-1}} - \mathbf{\Gamma}\mathbf{W}^T\sqrt{\mathbf{\Psi}^{-1}}\|_F^2 \\
&= \arg\min_{\mathbf{\Gamma}} \text{Tr}((\mathbf{X}\sqrt{\mathbf{\Psi}^{-1}} - \mathbf{\Gamma}\mathbf{W}^T\sqrt{\mathbf{\Psi}^{-1}})^T(\mathbf{X}\sqrt{\mathbf{\Psi}^{-1}} - \mathbf{\Gamma}\mathbf{W}^T\sqrt{\mathbf{\Psi}^{-1}})) \\
&= \arg\min_{\mathbf{\Gamma}} \text{Tr}(\sqrt{\mathbf{\Psi}^{-1}}\mathbf{X}^T\mathbf{X}\sqrt{\mathbf{\Psi}^{-1}} - 2\sqrt{\mathbf{\Psi}^{-1}}\mathbf{X}^T\mathbf{\Gamma}\mathbf{W}^T\sqrt{\mathbf{\Psi}^{-1}} + \sqrt{\mathbf{\Psi}^{-1}}\mathbf{W}\mathbf{\Gamma}^T\mathbf{\Gamma}\mathbf{W}^T\sqrt{\mathbf{\Psi}^{-1}}) \\
&= \arg\min_{\mathbf{\Gamma}} \text{Tr}(\mathbf{\Psi}^{-1}\mathbf{X}^T\mathbf{X} - 2\mathbf{\Psi}^{-1}\mathbf{X}^T\mathbf{\Gamma}\mathbf{W}^T + \mathbf{\Psi}^{-1}\mathbf{W}\mathbf{\Gamma}^T\mathbf{\Gamma}\mathbf{W}^T) \\
&= \arg\min_{\mathbf{\Gamma}} \text{Tr}(-2\mathbf{\Psi}^{-1}\mathbf{X}^T\mathbf{\Gamma}\mathbf{W}^T + \mathbf{\Psi}^{-1}\mathbf{W}\mathbf{\Gamma}'\mathbf{\Gamma}\mathbf{W}^T) \\
&= \arg\min_{\mathbf{\Gamma}} \text{Tr}(-2\mathbf{W}^T\mathbf{\Psi}^{-1}\mathbf{X}^T\mathbf{\Gamma} + \mathbf{\Gamma}^T\mathbf{\Psi}^{-1}\mathbf{W}\mathbf{\Gamma}^T) \\
\frac{\partial l(\mathbf{\Gamma})}{\partial \mathbf{\Gamma}} &= \frac{\partial}{\partial \mathbf{\Gamma}} \text{Tr}(-2\mathbf{W}^T\mathbf{\Psi}^{-1}\mathbf{X}^T\mathbf{\Gamma} + \mathbf{\Gamma}^T\mathbf{\Psi}^{-1}\mathbf{W}\mathbf{\Gamma}^T) = -2\mathbf{X}\mathbf{\Psi}^{-1}\mathbf{W} + 2\mathbf{\Gamma}\mathbf{W}^T\mathbf{\Psi}^{-1}\mathbf{W} \\
\frac{\partial l(\mathbf{\Gamma})}{\partial \mathbf{\Gamma}} &= 0 \implies \mathbf{\Gamma} = \mathbf{X}\mathbf{\Psi}^{-1}\mathbf{W}(\mathbf{W}^T\mathbf{\Psi}^{-1}\mathbf{W})^{-1}.
\end{aligned}$$

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\text{argmin}} \sum_{j=1}^{d} \frac{\|\mathbf{X}_{\cdot j} - \mathbf{\Gamma}\mathbf{W}_{j\cdot}^T\|^2}{\sigma_j^2}$$

which is minimized individually for each $\mathbf{W}_j$ as $\mathbf{W}_{j\cdot}^T = (\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{X}_{\cdot j}$, which gives the result. □

Proposition 1 *If $\mathbf{UDV}^T = \mathbf{\Gamma}$ is the SVD of $\mathbf{\Gamma}$, then $\mathbf{\Gamma}_1 = \mathbf{U}$, and $\mathbf{W}_1 = \mathbf{WVD}$ satisfy $\mathbf{\Gamma}_1\mathbf{W}_1^T = \mathbf{\Gamma}\mathbf{W}^T$ along with $\mathbf{\Gamma}_1^T\mathbf{\Gamma}_1 = \mathbf{I}_r$.*

Proof. It is easy to verify that $\mathbf{\Gamma}_1^T\mathbf{\Gamma}_1 = \mathbf{I}_r$.□

2.2. True Feature Recovery Guarantees

We have considered a model that aims to find a relationship between the observed $\mathbf{x} \in \mathbb{R}^d$ and a hidden set of variables (latent variables) $\boldsymbol{\gamma} \in \mathbb{R}^r$ with $r << d$ and assumes the latent factors and noise variables are independent of each other. It is as follows:

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \quad \text{with } E(\boldsymbol{\epsilon}) = \mathbf{0} \text{ and } \operatorname{var}(\boldsymbol{\epsilon}) = \mathbf{\Psi}. \tag{3}$$

We have also considered the following assumptions:

(A1) The observations, $(\mathbf{x}_i, i = 1, 2, \cdots, n)$ are independently generated from the LFA model (3) with $\boldsymbol{\mu} = 0$ and $\boldsymbol{\epsilon} \overset{i.i.d}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{\Psi})$.

(A2) Denote $\mathbf{\Gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \cdots, \boldsymbol{\gamma}_n)^T$. $\mathbf{\Gamma}_{ij}$ are i.i.d random variables with $E(\mathbf{\Gamma}_{ij}) = 0, Var(\mathbf{\Gamma}_{ij}) = 1, E(\mathbf{\Gamma}_{ij}^4) < \infty$, for all $(i, j) \in \{1, 2, \cdots, n\} \times \{1, 2, \cdots, r\}$.

(A3) There are m true features with indices $S = \{s_1, s_2, \cdots, s_m\}$ and $(d - m)$ noisy features in our data, which satisfy, for some positive constant $\gamma > 0$:

$$\min\{\boldsymbol{SNR}_i^*, i \in S\} \geq \max\{\boldsymbol{SNR}_i^*, i \notin S\} + \gamma. \tag{4}$$

To prove the following proposition, we first introduce the spiked covariance model[5]:

Definition 1 (Spiked Covariance model [due to [5]]:) *Under this model, the data matrix $\boldsymbol{X}$, can be viewed as $\boldsymbol{X}^T = \boldsymbol{E}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{Z}$, where $\boldsymbol{E} = [\boldsymbol{e_1}, \boldsymbol{e_2}, \cdots, \boldsymbol{e_d}]$ is a $d \times d$ orthogonal matrix, $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_d)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ and $\boldsymbol{Z}$ is a $d \times n$ matrix constructed with iid random variables $\boldsymbol{Z}_{ij}$ with $E(\boldsymbol{Z}_{ij}) = 0$, $E(\boldsymbol{Z}_{ij}^2) = 1$ and $E(\boldsymbol{Z}_{ij}^4) \leq \infty$. The population covariance matrix is $\boldsymbol{\Sigma} = \boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E^T}$. Here, λ_k's are assumed to follow a specific structure, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > \lambda_{r+1} = \cdots = \lambda_d = 1$.*

For the following section, we assume that $\lim_{n\to\infty} \frac{d}{n} = \delta$. Also there are k population eigenvalues such that $\lambda_i > 1 + \sqrt{\delta}$, for $i \leq k$. The following result is due to [3].

Theorem 2 (due to [3]) *For $\delta \in (0, 1)$, the following holds:*

$$s_i \overset{a.s.}{\to} \begin{cases} \rho(\lambda_i), & \textit{if } i \leq k \\ (1+\sqrt{\delta})^2 & \textit{otherwise}, \end{cases}$$

where $\rho(x) = x(1 + \frac{\delta}{x-1})$.

Although consistency could not be proved for $\delta > 0$, [5] proved consistency for $\delta = 0$.

Lemma 1 (due to [5]) *If $\lim_{n\to\infty} \frac{d}{n} = \delta = 0$, then,*

$$s_i \overset{a.s.}{\to} \begin{cases} \lambda_i & \textit{if } i \leq m \\ 1 & \textit{otherwise.} \end{cases}$$

Proposition 2 *Under the assumptions (**A1**, **A2**), if $\frac{d}{n} \to \delta \in (0, 1)$ as $n \to \infty$, and $\mathbf{\Psi} = \sigma^{*2}\mathbf{I}_d$ then $\sigma_{ML}^2 \overset{a.s.}{\to} \sigma^{*2}(1 + \sqrt{\delta})^2$. If $\delta = 0$ then $\sigma_{ML}^2 \overset{a.s.}{\to} \sigma^{*2}$.*

Proof. The covariance matrix under the PPCA model, $\mathbf{\Sigma}^* = \mathbf{W}^*\mathbf{W}^{*T} + \sigma^{*2}\mathbf{I}_d = \mathbf{E}^*\mathbf{L}^*\mathbf{E}^{*T} + \sigma^{*2}\mathbf{I}_d$ can be viewed as a spiked covariance model.

Let l_i^* be the i^{th} diagonal element of $\mathbf{L}^*$. As $\mathbf{W}^* \in \mathbb{R}^{d\times r}$ is a tall skinny matrix, the first r diagonal elements of $\mathbf{L}$ are greater than 0, and the rest of the diagonal elements are equal to 0.

Therefore, the SVD on $\mathbf{\Sigma}^*$ provides $\mathbf{\Sigma}^* = \mathbf{E}^*\mathbf{\Lambda}^*\mathbf{E}^{*T}$, where λ_i^*, the i^{th} element of $\mathbf{\Lambda}^*$ is:

$$\lambda_i^* = \begin{cases} l_i^* + \sigma^{*2} & \text{if } i \leq r \\ \sigma^{*2} & \text{otherwise.} \end{cases} \tag{5}$$

The ML estimates of σ^{*2} from [10] are given below:

$$\sigma^2_{ML} = \frac{1}{d-r}\sum_{j=r+1}^{d} s_j \tag{6}$$

where s_i is the i^{th} largest eigen value of $\hat{\mathbf{\Sigma}}$.

From (5), we have $\lambda^*_{r+1} = \lambda^*_{r+2} = \cdots = \lambda^*_d = \sigma^{*2}$. Let, $\mathbf{\Sigma}_0^* = \mathbf{\Sigma}^*/\sigma^2$. Therefore the estimated $\hat{\mathbf{\Sigma}}_0 = \hat{\mathbf{\Sigma}}/\sigma^2$. Let λ_i^{*0} and s_i^0 be the i^{th} largest eigenvalues of $\mathbf{\Sigma}_0^*$ and $\hat{\mathbf{\Sigma}}_0$ respectively.

It is easy to verify that the set of principal eigenvectors for $\mathbf{\Sigma}_0^*$ and $\mathbf{\Sigma}^*$ are the same and $\lambda_i^{*0} = \frac{\lambda_i^*}{\sigma^{*2}}$. A similar logic holds for $\hat{\mathbf{\Sigma}}_0$ and $\hat{\mathbf{\Sigma}}$ as well, i.e. $s_i^0 = \frac{s_i}{\sigma^{*2}}$.

Therefore, $\Lambda_0^* = \text{diag}(\lambda_1^{*0}, \cdots, \lambda_d^{*0})$ and $\lambda_1^{*0} \geq \lambda_2^{*0} \geq \cdots \geq \lambda_r^{*0} > \lambda_{r+1}^{*0} = \lambda_{r+2}^{*0} = \cdots = \lambda_d^{*0} = 1$

It can be easily verified that $\tilde{\mathbf{X}}_{n\times d} = \frac{1}{\sigma^*}\mathbf{X}$ follows a spiked covariance model. As, $\tilde{\mathbf{X}}^T = \mathbf{\Sigma}_0^{*\frac{1}{2}}\mathbf{Z}_{d\times n}$. From the assumptions (**A1**,**A2**), $\mathbf{Z}$ can be viewed as a matrix constructed with iid random variables $\boldsymbol{Z}_{ij}$ with $E(\boldsymbol{Z}_{ij}) = 0$, $E(\boldsymbol{Z}_{ij}^2) = 1$ and $E(\boldsymbol{Z}_{ij}^4) \leq \infty$.

Therefore, from the Theorem 2 we get, for $\delta \in (0, 1)$,

$$s_i^0 \overset{a.s.}{\to} (1+\sqrt{\delta})^2, \forall i > r.$$

Now $s_i^0 = s_i/\sigma^{*2}$, therefore $s_i \overset{a.s.}{\to} \sigma^{*2}(1+\sqrt{\delta})^2, \forall i > r$.

Let us denote $\mathbf{s}_{d-r} = \{s_{r+1}, s_{r+2}, \cdots, s_d\}$. It can be easily seen that σ^2_{ML} is a continuous transformation of $\mathbf{s}_{d-r}$, which can be defined as: $\sigma^2_{ML} = h(\mathbf{s}_{d-r}) = (\mathbf{s}^T_{d-r}\mathbf{1}_{d-r})/(d-r)$.

Then $h(\mathbf{s}_{d-r}) \overset{a.s.}{\to} h((\sigma^{*2}(1+\sqrt{\delta})^2)\mathbf{1}_{d-r}) \implies \sigma^2_{ML} \overset{a.s.}{\to} \sigma^{*2}(1+\sqrt{\delta})^2$.

Now if $\delta \to 0$, it is evident from the Lemma 1 that $\sigma^2_{ML} \overset{a.s.}{\to} \sigma^{*2}$. $\square$

Theorem 3 *Let d be fixed, and let $n \to \infty$:*

(C1) *Under assumptions (**A1**,**A2**), if $\mathbf{\Psi}^* = \sigma^{*2}\mathbf{I}_d$, then*

$$\boldsymbol{S\hat{N}R}_i^{PPCA} \overset{p}{\to} \boldsymbol{SNR}_i^*,$$

for all $i \in \{1, 2, \ldots, d\}$.

(C2) *Under the assumption (**A1**), if $\mathbf{\Psi}^* = \text{diag}(\sigma_1^{*2}, \sigma_2^{*2}, \ldots, \sigma_d^{*2})$ and $\boldsymbol{\gamma}_i \overset{i.i.d}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_r)$ then*

$$\boldsymbol{S\hat{N}R}_i^{LFA} \overset{p}{\to} \boldsymbol{SNR}_i^*,$$

for all $i \in \{1, 2, \ldots, d\}$.

*Furthermore, under the assumption (**A3**), the probability that the m features with the highest SNRs are the true features converges to 1 as $n \to \infty$ for both (**C1**) and (**C2**).*

Here, $\boldsymbol{S\hat{N}R}^{PPCA}$ and $\boldsymbol{S\hat{N}R}^{LFA}$ denote the estimated SNRs from PPCA and LFA, respectively.

Proof. We have defined SNRs as:

$$\boldsymbol{SNR}_i = \frac{\sum_{j=1}^r \mathbf{W}_{ij}^2}{\sigma_i^2} = \frac{(\mathbf{W}\mathbf{W}^T)_{ii}}{\sigma_i^2}, i \in \{1, 2, \cdots, d\}. \tag{7}$$

Under the assumptions (**A1**,**A2**), for a particular dimension (i), the n observations corresponding to dimension i, denoted as $\{\mathbf{X}_{ji}, j = 1, 2, \cdots, n\}$ are i.i.d single valued random variables with

$$E(\mathbf{X}_{ji}) = 0 \tag{8}$$

$$Var(\mathbf{X}_{ji}) = \sum_{k=1}^{d} \lambda_k^* \mathbf{E}_{ik}^{*2} \tag{9}$$

Where $Cov(\mathbf{X}) = \mathbf{\Sigma}^* = \mathbf{E}^*\mathbf{\Lambda}^*\mathbf{E}^{*T}$. Therefore, when $\delta = 0$ (that is, d is fixed and $n \to \infty$), by the law of large numbers, the corresponding sample variance, which is also the element i^{th} of the principal diagonal of $\hat{\mathbf{\Sigma}}$ (denoted as $\hat{\mathbf{\Sigma}}_{ii}$) converges to $\mathbf{\Sigma}^*_{ii}$ in probability. i.e.

$$\hat{\mathbf{\Sigma}}_{ii} \xrightarrow{p} \mathbf{\Sigma}^*_{ii} \tag{10}$$

When $\mathbf{\Psi}^* = \sigma^{*2}\mathbf{I}_d$,

$$\mathbf{\Sigma}^*_{ii} = (\mathbf{W}^*\mathbf{W}^{*T})_{ii} + \sigma^{*2} \tag{11}$$

$$\hat{\mathbf{\Sigma}}_{ii} = (\hat{\mathbf{W}}\hat{\mathbf{W}}^T)_{ii} + \hat{\sigma}^2 \tag{12}$$

From Proposition 2, we have $\hat{\sigma^2} \xrightarrow{a.s.} \sigma^{*2}$ for $\delta = 0$. Therefore combining this result with (10), from (11), we get:

$$(\hat{\mathbf{W}}\hat{\mathbf{W}}^T)_{ii} \xrightarrow{p} (\mathbf{W}^*\mathbf{W}^{*T})_{ii} \tag{13}$$

From the definition of SNR (7), we get that the estimated SNR is a continuous transformation of $((\hat{\mathbf{W}}\hat{\mathbf{W}}^T)_{ii}, \hat{\sigma^2}), i = 1, 2, \cdots, d$ and provided $\hat{\sigma}^2 > 0$. Therefore, under condition (**C1**), we get, $\hat{\boldsymbol{SNR}}_i^{PPCA} \xrightarrow{p} \boldsymbol{SNR}^*_i$.

Under condition (**C2**), when d is fixed and $n \to \infty$, the ML estimate of the noise covariance matrix, $\hat{\mathbf{\Psi}}$, we get from [4], is consistent[2], i.e.

$$\hat{\sigma}_i^2 \xrightarrow{p} \sigma_i^{*2}, i = 1, 2, \cdots, d \tag{14}$$

Also, the n observations corresponding to dimension i, denoted as $\{\mathbf{X}_{ji}, j = 1, 2, \cdots, n\}$ are i.i.d single valued random variables with $\mathbf{X}_{ji} \overset{i.i.d}{\sim} \mathcal{N}(0, \sum_{k=1}^d \lambda_k^* \mathbf{E}_{ik}^{*2})$ Therefore, by the law of large numbers, we get $\hat{\mathbf{\Sigma}}_{ii} \xrightarrow{p} \mathbf{\Sigma}^*_{ii}$ and using similar logic from the condition (**C1**), we get,

$$(\hat{\mathbf{W}}\hat{\mathbf{W}}^T)_{ii} \xrightarrow{p} (\mathbf{W}^*\mathbf{W}^{*T})_{ii}, i = 1, 2, \cdots, d \tag{15}$$

Under the condition (**C2**), the definition of SNR (7) suggests that the estimated SNR is a continuous transformation of $((\hat{\mathbf{W}}\hat{\mathbf{W}}^T)_{ii}, \hat{\sigma}_i^2)$, provided $\min_{i\in\{1,2,\cdots,d\}} \hat{\sigma}_i^2 > 0$. Therefore, combining the results of (14) and (15), we get: $\hat{\boldsymbol{SNR}}_i^{LFA} \xrightarrow{p} \boldsymbol{SNR}^*_i$

Now, we prove the last part of the theorem. Let $\hat{\boldsymbol{SNR}}_i$ be the estimate of true $\boldsymbol{SNR}^*_i$, for $i = 1, 2, \cdots, d$. Under condition (**C1** and **C2**), we get, $\hat{\boldsymbol{SNR}}_i \xrightarrow{p} \boldsymbol{SNR}^*_i$. Therefore, for any $\epsilon > 0$, there exists n_0 such that for any $n \geq n_0$

$$P(|\hat{\boldsymbol{SNR}}_i - \boldsymbol{SNR}^*_i| < \gamma/2) > 1 - \epsilon.$$

Therefore, with at least $(1 - \epsilon)$ probability,

$$\begin{aligned}
& \boldsymbol{SNR}^*_i - \gamma/2 < \hat{\boldsymbol{SNR}}_i < \boldsymbol{SNR}^*_i + \gamma/2, \forall i \\
& \implies \min_{i\in S} \hat{\boldsymbol{SNR}}_i > \min_{i\in S} \boldsymbol{SNR}^*_i - \gamma/2, \text{ and } \max_{j\notin S} \boldsymbol{SNR}^*_j + \gamma/2 > \max_{j\notin S} \hat{\boldsymbol{SNR}}_j, \\
\text{(by } \mathbf{A3}) & \implies \min_{i\in S} \hat{\boldsymbol{SNR}}_i > \min_{i\in S} \boldsymbol{SNR}^*_i - \gamma/2 > \max_{j\notin S} \boldsymbol{SNR}^*_j + (\gamma - \gamma/2) > \max_{j\notin S} \hat{\boldsymbol{SNR}}_j,
\end{aligned}$$

which proves that with probability $1-\epsilon$, the m features with the highest estimated SNRs are the true features S for $n \geq n_0$. □

We have used Proposition 3 from [11], to prove the theorem related to the generalized score $r-$ (Theorem 4).

Proposition 3 (due to [11]) *If $\mathbf{\Sigma}$ admits a rank-r eigendecomposition of the form:*

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{D}\mathbf{L}^T + \lambda\mathbf{I}_m, \tag{16}$$

with $\mathbf{L} \in \mathbb{R}^{m\times r}$, diagonal $\mathbf{D} = \mathrm{diag}(\mathbf{d}) \in \mathbb{R}^{r\times r}$ with positive entries, and $\lambda > 0$, the Mahalanobis distance can be computed as:

$$MD(\mathbf{x}, \boldsymbol{\mu}, \mathbf{\Sigma}) = r(\mathbf{x}; \boldsymbol{\mu}, \mathbf{L}, \mathbf{D}, \lambda), \tag{17}$$

where

$$r(\mathbf{x}; \boldsymbol{\mu}, \mathbf{L}, \mathbf{D}, \lambda) = \|\mathbf{x} - \boldsymbol{\mu}\|_2^2/\lambda - \|\mathbf{u}(\mathbf{x})\|_2^2/\lambda, \tag{18}$$

with $\mathbf{u}(\mathbf{x}) = \mathrm{diag}(\frac{\sqrt{\mathbf{d}}}{\sqrt{\mathbf{d}+\lambda\mathbf{1}_r}})\mathbf{L}^T(\mathbf{x} - \boldsymbol{\mu})$, and $\mathbf{1}_r = (1, 1, \cdots, 1)^T$. The operation $\frac{\sqrt{\mathbf{d}}}{\sqrt{\mathbf{d}+\lambda\mathbf{1}_r}}$ is performed element-wise.

Theorem 4 *If*

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{D}\mathbf{L}^T + \mathbf{\Psi}, \tag{19}$$

with $\mathbf{L} \in \mathbb{R}^{m \times r}$*, the diagonal matrices* $\mathbf{D} \in \mathbb{R}^{r \times r}$ *and* $\mathbf{\Psi} \in \mathbb{R}^{m \times m}$ *with positive entries, the Mahalanobis distance can be computed as:*

$$MD(\mathbf{x}, \boldsymbol{\mu}, \mathbf{\Sigma}) = r(\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{x}; \mathbf{\Psi}^{-\frac{1}{2}}\boldsymbol{\mu}, \mathbf{L}', \mathbf{D}', 1) \tag{20}$$

where $r(\mathbf{x}; \boldsymbol{\mu}, \mathbf{L}, \mathbf{D}, \lambda)$ *is defined in Eq. (18), and* $\mathbf{L}'$ *and* $\mathbf{D}'$ *are obtained by SVD on* $\mathbf{\Sigma}' = \mathbf{\Psi}^{-\frac{1}{2}}\mathbf{\Sigma}\mathbf{\Psi}^{-\frac{1}{2}}$.

Proof. We consider the following transformation:

$$\begin{aligned}
\mathbf{x}' &= \mathbf{\Psi}^{-\frac{1}{2}}\mathbf{x}, \\
\boldsymbol{\mu}' &= \mathbf{\Psi}^{-\frac{1}{2}}\mu, \\
\mathbf{\Sigma}' &= (\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{W})(\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{W}^T) + \mathbf{I}_m.
\end{aligned}$$

$\mathbf{\Sigma}'$ looks similar to (16), with $\lambda = 1$. Therefore, using the Proposition 3, we get: $MD(\mathbf{x}', \boldsymbol{\mu}', \mathbf{\Sigma}') = r(\mathbf{x}'; \boldsymbol{\mu}', \mathbf{L}', \mathbf{D}', 1)$. Here, $\mathbf{\Sigma}' = \mathbf{\Psi}^{-\frac{1}{2}}\mathbf{\Sigma}\mathbf{\Psi}^{-\frac{1}{2}} = \mathbf{L}'\mathbf{D}'\mathbf{L}'^T$. Also,

$$\begin{aligned}
&MD(\mathbf{x}, \boldsymbol{\mu}, \mathbf{\Sigma}) \\
&= (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \\
&= (\mathbf{x} - \boldsymbol{\mu})^T (\mathbf{\Psi}^{-\frac{1}{2}})(\mathbf{\Psi}^{\frac{1}{2}})\mathbf{\Sigma}^{-1}(\mathbf{\Psi}^{\frac{1}{2}})(\mathbf{\Psi}^{-\frac{1}{2}})(\mathbf{x} - \boldsymbol{\mu}) \\
&= (\mathbf{x}' - \boldsymbol{\mu}')^T (\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{\Sigma}\mathbf{\Psi}^{-\frac{1}{2}})^{-1}(\mathbf{x}' - \boldsymbol{\mu}') \\
&= (\mathbf{x}' - \boldsymbol{\mu}')^T \mathbf{\Sigma}'^{-1}(\mathbf{x}' - \boldsymbol{\mu}') \\
&= MD(\mathbf{x}', \boldsymbol{\mu}', \mathbf{\Sigma}') \\
&= r(\mathbf{x}'; \boldsymbol{\mu}', \mathbf{L}', \mathbf{D}', 1) \\
&= r(\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{x}; \mathbf{\Psi}^{-\frac{1}{2}}\boldsymbol{\mu}, \mathbf{L}', \mathbf{D}', 1). \ \square
\end{aligned}$$

References

[1] R. Aziz, C. Verma, and N. Srivastava. A weighted-snr feature selection from independent component subspace for nb classification of microarray data. *Int J Adv Biotec Res*, 6:245–255, 2015. 1

[2] Jushan Bai and Kunpeng Li. Statistical analysis of factor models of high dimension. *Ann. Stat.*, pages 436–465, 2012. 4

[3] Jinho Baik and Jack W Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of multivariate analysis*, 97(6):1382–1408, 2006. 2

[4] Zoubin Ghahramani, Geoffrey E Hinton, et al. The em algorithm for mixtures of factor analyzers. Technical report, CRG-TR-96-1, University of Toronto, 1996. 4

[5] Seunggeun Lee, Fei Zou, and Fred A Wright. Convergence and prediction of principal component scores in high-dimensional settings. *Ann. Stat.*, 38(6):3605, 2010. 2

[6] Xi-Ao Ma, Hao Xu, Yi Liu, and Justin Zuopeng Zhang. Class-specific feature selection using fuzzy information-theoretic metrics. *Engineering Apps. of AI*, 136:109035, 2024. 1

[7] L. Nanni. Cluster-based pattern discrimination: A novel technique for feature selection. *Patt. Rec. Lett.*, 27(6), 2006. 1

[8] D. Nardone, A. Ciaramella, and A. Staiano. A sparse-modeling based approach for class specific feature selection. *PeerJ Computer Science*, 5:e237, 2019. 1

[9] B. Pineda-Bautista, J. Carrasco-Ochoa, and J. Martınez-Trinidad. General framework for class-specific feature selection. *Expert Sys. with Apps.*, 38(8):10018–10024, 2011. 1

[10] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *JRSS B*, 61(3):611–622, 1999. 3

[11] Boshi Wang and Adrian Barbu. Scalable learning with incremental probabilistic pca. In *IEEE Int. Conf. on Big Data*, pages 5615–5622, 2022. 4