AUSTRIAN JOURNAL OF STATISTICS
Volume 38 (2009), Number 3, 135–150

A Corrected Criterion for Selecting the Optimum Number of Principal Components

Hannes Kazianka and Jürgen Pilz
Institute of Statistics, University of Klagenfurt

Abstract: Determining the optimum number of components to be retained is a key problem in principal component analysis (PCA). Besides the rule of thumb estimates there exist several sophisticated methods for automatically selecting the dimensionality of the data. Based on the probabilistic PCA model Minka (2001) proposed an approximate Bayesian model selection criterion. In this paper we correct this criterion and present a modified version. We compare the novel criterion with various other approaches in a simulation study. Furthermore, we use it for finding the optimum number of principal components in hyper-spectral skin cancer images.

Zusammenfassung: Ein zentrales Problem der Hauptkomponentenanalyse (PCA) ist es, die Anzahl an wichtigen Komponenten zu wählen. Neben den bekannten Faustregeln gibt es verschiedene fortgeschrittene Methoden zur automatischen Wahl der optimalen Dimensionalität der Daten. Beispielsweise wurde von Minka (2001), basierend auf dem probabilistischen PCA Modell, ein approximativ Bayes'sches Kriterium zur Modellwahl vorgeschlagen. In der vorliegenden Arbeit korrigieren wir dieses Kriterium und stellen eine modifizierte Version vor. Das neue Kriterium wird mit anderen Verfahren in einer Simulationsstudie verglichen. Weiters wird es zum Finden der optimalen Anzahl an Hauptkomponenten in hyper-spektralen Hautkrebsdaten verwendet.

Keywords: PCA, Probabilistic PCA, Model Selection.

1 Introduction

Principal component analysis (PCA) was introduced by Pearson (1901) and is a linear orthogonal data transformation. The data is transformed in such a way that in the new coordinate system the components are uncorrelated and sorted in a descending order according to their variance. PCA is a standard tool for reducing multidimensional data sets to lower dimensions for further statistical analysis by omitting higher-order components. Determining the optimum number of components to be retained is a major problem in PCA. Besides certain rule of thumb estimates there exist numerous sophisticated methods for automatically selecting the dimensionality of the data. For the Gaussian probabilistic PCA model of Tipping and Bishop (1999) an approximate Bayesian model selection criterion was proposed by Minka (2001). Although mentioned in books and used in various applications, we found some mistakes in its proof. In the following we correct this criterion and present a modified version.

The paper is organized as follows. Section 2 presents the probabilistic PCA model and mentions its basic properties. In Section 3 the model selection criterion proposed by

Minka (2001) is corrected. Section 4 compares the corrected criterion with several other approaches in a simulation study while Section 5 presents the results when it is applied to hyper-spectral skin cancer image data. Section 6 is devoted to conclusions.

2 Probabilistic PCA

A probabilistic model for PCA was introduced by Tipping and Bishop (1999) and is a special case of the factor analysis model (see Basilevsky, 1994). It assumes that the random vector $\boldsymbol{x}$ is a linear combination of basis vectors and an additive noise term,

$$\boldsymbol{x} = \sum_{i=1}^{k} \boldsymbol{h}_j w_j + \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \boldsymbol{H}\boldsymbol{w} + \boldsymbol{\mu} + \boldsymbol{\varepsilon}\,, \tag{1}$$

where $\boldsymbol{x}$ has length d, $\boldsymbol{w}$ has smaller length k, $\boldsymbol{\mu}$ is the mean of $\boldsymbol{x}$ and the noise model is isotropic Gaussian, $p(\boldsymbol{\varepsilon}) \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_d)$. Moreover, the density of $\boldsymbol{w}$ is assumed to be standard Gaussian, $p(\boldsymbol{w}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k)$. Therefore, the probability of observing $\boldsymbol{x}$ follows a normal distribution, $p(\boldsymbol{x} \mid \boldsymbol{w}, \boldsymbol{H}, \boldsymbol{\mu}, \sigma^2) \sim \mathcal{N}(\boldsymbol{H}\boldsymbol{w} + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_d)$, and the marginal distribution of $\boldsymbol{x}$ is $p(\boldsymbol{x} \mid \boldsymbol{H}, \boldsymbol{\mu}, \sigma^2) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{H}\boldsymbol{H}^T + \sigma^2 \boldsymbol{I}_d)$. With this, the likelihood of a data set, $\mathcal{D} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, is

$$p(\mathcal{D} \mid \boldsymbol{H}, \boldsymbol{\mu}, \sigma^2) = (2\pi)^{-\frac{nd}{2}} |\boldsymbol{H}\boldsymbol{H}^T + \sigma^2 \boldsymbol{I}_d|^{-\frac{n}{2}} \exp\left\{ -\frac{n}{2} \mathrm{tr}\left[(\boldsymbol{H}\boldsymbol{H}^T + \sigma^2 \boldsymbol{I}_d)^{-1} \boldsymbol{S} \right] \right\}, \tag{2}$$

where $\boldsymbol{S} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$. Consequently, the maximum-likelihood estimate, $\hat{\boldsymbol{\mu}}$, for $\boldsymbol{\mu}$ is the arithmetic mean. Defining $\hat{\boldsymbol{S}} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})^T$ and $\boldsymbol{C} = \boldsymbol{H}\boldsymbol{H}^T + \sigma^2 \boldsymbol{I}_d$ the log-likelihood evaluated at $\hat{\boldsymbol{\mu}}$ is

$$\log p(\mathcal{D} \mid \boldsymbol{H}, \hat{\boldsymbol{\mu}}, \sigma^2) = -\frac{n}{2} \left(d \log(2\pi) + \log |\boldsymbol{C}| + \mathrm{tr}\left[\boldsymbol{C}^{-1} \hat{\boldsymbol{S}} \right] \right). \tag{3}$$

Next we show how the maximum-likelihood estimates for $\boldsymbol{H}$ and σ^2 can be obtained. Using the symmetry of $\boldsymbol{C}$ and $\hat{\boldsymbol{S}}$ the gradient of (3) with respect to $\boldsymbol{H}$ is

$$\begin{aligned} \frac{\partial \log p(\mathcal{D} \mid \boldsymbol{H}, \hat{\boldsymbol{\mu}}, \sigma^2)}{\partial \boldsymbol{H}} &= \frac{n}{2} \left(\left(\boldsymbol{C}^{-1} \hat{\boldsymbol{S}} \boldsymbol{C}^{-1} \right)^T - \boldsymbol{C}^{-T} \right) 2\boldsymbol{H} \\ &= n \left(\boldsymbol{C}^{-1} \hat{\boldsymbol{S}} \boldsymbol{C}^{-1} \boldsymbol{H} - \boldsymbol{C}^{-1} \boldsymbol{H} \right). \end{aligned}$$

Setting the gradient to zero and multiplying with $\boldsymbol{C}$ from the left yields

$$\hat{\boldsymbol{S}} \boldsymbol{C}^{-1} \boldsymbol{H} = \boldsymbol{H}\,. \tag{4}$$

We are interested in the solutions of (4) for $\boldsymbol{H}$. A possible solution would be $\hat{\boldsymbol{H}} = \boldsymbol{0}$ which corresponds to a minimum of the likelihood. A second solution is $\boldsymbol{C} = \hat{\boldsymbol{S}}$, where the covariance model is exact. In this case, $\boldsymbol{H}\boldsymbol{H}^T = \hat{\boldsymbol{S}} - \sigma^2 \boldsymbol{I}_d$, which has the solution $\hat{\boldsymbol{H}} = \boldsymbol{U}(\boldsymbol{L} - \sigma^2 \boldsymbol{I}_d)^{1/2} \boldsymbol{R}$, where $\boldsymbol{U} \in \mathbb{R}^{d \times d}$ is the orthogonal matrix with the eigenvectors of $\hat{\boldsymbol{S}}$ in its columns, $\boldsymbol{L} \in \mathbb{R}^{d \times d}$ is the diagonal matrix containing the corresponding eigenvalues and $\boldsymbol{R} \in \mathbb{R}^{d \times d}$ is an arbitrary orthogonal matrix. As Tipping and Bishop (1999)

note, having an exact covariance model is generally undesirable because this information is effectively discarded in the dimensionality reduction process anyway.

For the general and more interesting case consider all solutions of (4), but $\hat{\boldsymbol{H}} \neq \boldsymbol{0}$ and $\boldsymbol{C} \neq \hat{\boldsymbol{S}}$. Using the singular value decomposition we know that the estimator can be written as $\hat{\boldsymbol{H}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{R}^T$, where now $\boldsymbol{U} \in \mathbb{R}^{d\times k}$ is a matrix with orthonormal columns $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k)$, $\boldsymbol{D} \in \mathbb{R}^{k\times k}$ is the diagonal matrix containing the singular values $d_1, \ldots, d_k$, and $\boldsymbol{R} \in \mathbb{R}^{k\times k}$ is again an arbitrary orthogonal matrix. We deduce from (4) that

$$\hat{\boldsymbol{S}}\left(\boldsymbol{U}\boldsymbol{D}^2\boldsymbol{U}^T + \sigma^2\boldsymbol{I}_d\right)^{-1}\boldsymbol{U}\boldsymbol{D}\boldsymbol{R}^T = \boldsymbol{U}\boldsymbol{D}\boldsymbol{R}^T \quad \Longrightarrow \quad \hat{\boldsymbol{S}}\boldsymbol{U}\boldsymbol{D} = \boldsymbol{U}\left(\boldsymbol{D}^2 + \sigma^2\boldsymbol{I}_k\right)\boldsymbol{D}\,.$$

If $d_i = 0$, $i = 1, \ldots, k$, the vector $\boldsymbol{u}_i$ can be chosen arbitrarily such that $\boldsymbol{U}$ has orthonormal columns. If $d_i \neq 0$, we have $\hat{\boldsymbol{S}}\boldsymbol{u}_i = (\sigma^2 + d_i^2)\boldsymbol{u}_i$ which means that each column of $\boldsymbol{U}$ has to be an eigenvector of $\hat{\boldsymbol{S}}$ corresponding to the eigenvalue $l_i = \sigma^2 + d_i^2$. Therefore, $d_i = (l_i - \sigma^2)^{1/2}$ and all maximum-likelihood solutions for $\boldsymbol{H}$ must have the form

$$\hat{\boldsymbol{H}} = \boldsymbol{U}(\boldsymbol{L} - \sigma^2\boldsymbol{I}_k)^{1/2}\boldsymbol{R}\,, \tag{5}$$

where now $\boldsymbol{L} \in \mathbb{R}^{k\times k}$ is the diagonal matrix with entries l_i and $\boldsymbol{U}$ contains the corresponding eigenvectors. With this, we can simplify

$$\begin{aligned}\left|\hat{\boldsymbol{H}}\hat{\boldsymbol{H}}^T + \sigma^2\boldsymbol{I}_d\right| &= \left|\boldsymbol{U}\left(\boldsymbol{L} - \sigma^2\boldsymbol{I}_k\right)\boldsymbol{U}^T + \sigma^2\boldsymbol{I}_d\right| = \sigma^{2(d-k)}\left|(\boldsymbol{L} - \sigma^2\boldsymbol{I}_k)\boldsymbol{U}^T\boldsymbol{U} + \sigma^2\boldsymbol{I}_k\right| \\ &= \sigma^{2(d-k)}\left|\boldsymbol{L} - \sigma^2\boldsymbol{I}_k + \sigma^2\boldsymbol{I}_k\right| = \sigma^{2(d-k)}\left|\boldsymbol{L}\right| = \sigma^{2(d-q)}\prod_{i=1}^{q} l_i\,,\end{aligned} \tag{6}$$

where q is the number of non-zero d_i and equals the rank of $\hat{\boldsymbol{H}}$. Therefore, the eigenvalues $l_1, \ldots, l_q$ correspond to those components retained in PCA and $l_{q+1}, \ldots, l_d$ correspond to those discarded. Using the Woodbury matrix identity we find that

$$\begin{aligned}\left(\hat{\boldsymbol{H}}\hat{\boldsymbol{H}}^T + \sigma^2\boldsymbol{I}_d\right)^{-1} &= \sigma^{-2}\boldsymbol{I}_d - \sigma^{-2}\boldsymbol{U}\left(\left(\boldsymbol{L} - \sigma^2\boldsymbol{I}_k\right)^{-1} + \sigma^{-2}\boldsymbol{I}_k\right)^{-1}\boldsymbol{U}^T\sigma^{-2} \\ &= \sigma^{-2}\boldsymbol{I}_d + \boldsymbol{U}\left(\boldsymbol{L}^{-1} - \sigma^{-2}\boldsymbol{I}_k\right)\boldsymbol{U}^T\,.\end{aligned} \tag{7}$$

Since $\boldsymbol{U}^T\hat{\boldsymbol{S}}\boldsymbol{U} = \boldsymbol{L}$, we get that

$$\begin{aligned}\operatorname{tr}\left[\left(\hat{\boldsymbol{H}}\hat{\boldsymbol{H}}^T + \sigma^2\boldsymbol{I}_d\right)^{-1}\hat{\boldsymbol{S}}\right] &= \operatorname{tr}\left[\sigma^{-2}\hat{\boldsymbol{S}}\right] + \operatorname{tr}\left[\boldsymbol{U}\left(\boldsymbol{L}^{-1} - \sigma^{-2}\boldsymbol{I}_k\right)\boldsymbol{U}^T\hat{\boldsymbol{S}}\right] \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{d} l_i + \sum_{i=1}^{q}\left(\frac{1}{l_i} - \frac{1}{\sigma^2}\right)l_i = q + \frac{1}{\sigma^2}\sum_{i=q+1}^{d} l_i\,.\end{aligned}$$

Plugging the latter result and (6) into the log-likelihood (3) we get

$$\log p\left(\mathcal{D} \mid \hat{\boldsymbol{H}}, \hat{\boldsymbol{\mu}}, \sigma^2\right) = -\frac{n}{2}\left(\sum_{i=1}^{q}\log l_i + (d-q)\log\sigma^2 + d\log 2\pi + q + \frac{1}{\sigma^2}\sum_{i=q+1}^{d} l_i\right). \tag{8}$$

The maximum-likelihood estimate for σ^2 turns out to be the average of the discarded eigenvalues:

$$\frac{n(d-q)}{\sigma} + \frac{n}{\sigma^3} \sum_{i=q+1}^{d} l_i = 0 \quad \Longrightarrow \quad \hat{\sigma}^2 = \frac{1}{d-q} \sum_{i=q+1}^{d} l_i \,. \tag{9}$$

Substituting $\hat{\sigma}^2$ into the log-likelihood (8) we arrive at

$$\log p\left(\mathcal{D} \mid \hat{\boldsymbol{H}}, \hat{\boldsymbol{\mu}}, \hat{\sigma}^2\right) = -\frac{n}{2}\left(\sum_{i=1}^{q} \log l_i + (d-q) \log\left(\frac{1}{d-q} \sum_{i=q+1}^{d} l_i\right) + d \log 2\pi + d\right). \tag{10}$$

In order to find those eigenvalues of $\hat{\boldsymbol{S}}$ that correspond to discarded components we have to maximize the above log-likelihood with respect to their choice. The sum of all eigenvalues is constant implying that maximization of (10) with respect to the eigenvalues is equivalent to minimizing

$$\min_{\substack{l_i \\ i=q+1,\ldots,d}} \log\left(\frac{1}{d-q} \sum_{i=q+1}^{d} l_i\right) - \frac{1}{d-q} \sum_{i=q+1}^{d} \log l_i \,.$$

The above function is minimized when all l_i, $i = q+1, \ldots, d$, have the same value. Hence, the discarded eigenvalues are adjacent in the ordered spectrum of $\hat{\boldsymbol{S}}$. From the fact that $l_i > \sigma^2$, $\forall i = 1, \ldots, q$ and (9) we know that the smallest eigenvalue of $\hat{\boldsymbol{S}}$ is definitely discarded. This implies that the smallest $d-q$ eigenvalues of $\hat{\boldsymbol{S}}$ are discarded and the top q eigenvalues are retained. Seen from this angle, the estimate for σ^2 can be interpreted as the average information loss when reducing the dimension of the data.

3 Choosing the Number of Principal Components Using Bayesian Model Selection

For the probabilistic PCA model with positive definite covariance matrix Minka (2001) developed a fast and efficient criterion to select the optimum number of principal components. It is based on Bayesian model selection and is therefore sometimes called the MIBS criterion (*Mi*nka *B*ayesian model *S*election) in the literature. A variety of books refer to this criterion, for example Cichocki and Amari (2002), Smidl and Quinn (2005) and Bishop (2008). It showed good results for PCA and also for independent component analysis. Although the MIBS criterion is widely used, we found some errors in its proof. In this section we give a corrected proof and propose a modified criterion for selecting the optimum number of principal components.

3.1 Choice of the Prior

In contrast to maximum-likelihood estimation performed in Section 2 we have to assign priors to the parameters $\boldsymbol{\mu}$, $\boldsymbol{H}$ and σ^2 in the probabilistic PCA model (1) when we want

to go the Bayesian way. For the mean $\boldsymbol{\mu}$ we use a non-informative, flat prior, $p(\boldsymbol{\mu}) \propto 1$. Integrating the likelihood (2) with respect to $\boldsymbol{\mu}$ yields

$$p(\mathcal{D} \mid \boldsymbol{H}, \sigma^2) = n^{-\frac{d}{2}} (2\pi)^{-\frac{d(n-1)}{2}} \left|\boldsymbol{H}\boldsymbol{H}^T + \sigma^2 \boldsymbol{I}_d\right|^{-\frac{n-1}{2}} \exp\left\{-\frac{n}{2}\mathrm{tr}\left[\left(\boldsymbol{H}\boldsymbol{H}^T + \sigma^2 \boldsymbol{I}_d\right)^{-1} \hat{\boldsymbol{S}}\right]\right\}. \tag{11}$$

The prior for $\boldsymbol{H}$ is constructed similar to decomposition (5), $\boldsymbol{H} = \boldsymbol{U}(\boldsymbol{\Lambda} - \sigma^2 \boldsymbol{I}_k)^{1/2}\boldsymbol{R}$. In contrast to Section 2, here the orthogonal matrix $\boldsymbol{R}$ and the matrix $\boldsymbol{U}$, which has orthonormal columns, are model parameters. $\boldsymbol{\Lambda}$ is a diagonal matrix containing parameters λ_i, $i = 1, \ldots, k$. Minka (2001) found a conjugate prior for $\boldsymbol{U}$, $\boldsymbol{\Lambda}$, $\boldsymbol{R}$ and σ^2. This prior is parameterized by a single parameter $\alpha > 0$ and by applying simplifications similar to (6) and (7) it can be written as

$$\begin{aligned} p(\boldsymbol{H}, \sigma^2) &= p(\boldsymbol{U}, \boldsymbol{\Lambda}, \boldsymbol{R}, \sigma^2) \\ &\propto \left|\boldsymbol{H}\boldsymbol{H}^T + \sigma^2 \boldsymbol{I}_d\right|^{-\frac{\alpha+2}{2}} \exp\left\{-\frac{\alpha}{2}\mathrm{tr}\left[\left(\boldsymbol{H}\boldsymbol{H}^T + \sigma^2 \boldsymbol{I}_d\right)^{-1}\right]\right\} \\ &\propto |\boldsymbol{\Lambda}|^{-\frac{\alpha+2)}{2}} \sigma^{-(\alpha+2)(d-k)} \exp\left\{-\frac{\alpha}{2}\left(\mathrm{tr}\left[\boldsymbol{\Lambda}^{-1}\right] + \frac{d-k}{\sigma^2}\right)\right\}. \end{aligned} \tag{12}$$

With this definition of the prior the parameters are a-priori independent since their joint prior density (12) factors into the product of the marginal prior densities for each of the parameters:

$$p(\boldsymbol{U}, \boldsymbol{\Lambda}, \boldsymbol{R}, \sigma^2) = p(\sigma^2)p(\boldsymbol{U})p(\boldsymbol{R}) \prod_{i=1}^{k} p(\lambda_i)\,,$$

where $p(\sigma^2) \sim \chi^{-2}(\alpha(d-k), (\alpha+2)(d-k) - 2)$ and $p(\lambda_i) \sim \chi^{-2}(\alpha, \alpha)$ are scaled inverse-χ^2 priors. To be least informative the prior for $\boldsymbol{U}$ is chosen as the reciprocal of the area of the orthonormally constrained subset of the Cartesian product of k d-dimensional unit hyperballs, known as the Stiefel manifold (see Khatri and Mardia, 1977 and James, 1954):

$$p(\boldsymbol{U}) = 2^{-k}\pi^{\frac{1}{4}k(k-1-2d)} \prod_{i=1}^{k} \Gamma\left(\frac{d-i+1}{2}\right) \propto 1\,.$$

The prior for $\boldsymbol{R}$ can also be assumed constant but as the variable does not appear in the likelihood (11) we can integrate it out.

3.2 The Corrected MIBS Criterion

Among the well-known statistical model selection criteria are Akaike's information criterion (AIC) and the minimum description length (MDL) criterion. These can be easily used for finding the optimum number of principal components. Assume we observe samples $\mathcal{D} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ from a zero-mean normal random vector. An appropriate number of principal components is the value of $k = 1, \ldots, d$ for which

$$\mathrm{AIC}(k) = -2n(d-k)\log\rho(k) + 2k(2d-k)\,, \tag{13}$$

$$\mathrm{MDL}(k) = -n(d-k)\log\rho(k) + \frac{k}{2}(2d-k)\log n\,, \tag{14}$$

is minimized. In (13) and (14) the function $\rho(k)$ is

$$\rho(k) = \frac{(l_{k+1} l_{k+2} \cdots l_d)^{\frac{1}{d-k}}}{\frac{1}{d-k}(l_{k+1} + l_{k+2} + \cdots + l_d)} .$$

The criterion that was recommended in Minka (2001) has shown reasonable performance in various applications and is fast to evaluate (see Cichocki and Amari, 2002 and Leonowicz, Karvanen, Tanaka, and Rezmer, 2004):

$$p(\mathcal{D} \mid k) \approx 2^{\frac{m-k}{2}} \rho(k)^{-n(d-k)} |\boldsymbol{A}|^{-\frac{1}{2}} n^{-\frac{k}{2}} \prod_{i=1}^{k} \Gamma\left(\frac{d-i+1}{2}\right) \prod_{i=1}^{k} l_i^{-\frac{n}{2}} \prod_{i=k+1}^{d} l_i^n , \qquad (15)$$

where $m = dk - k(k+1)/2$ and $|\boldsymbol{A}|$ is defined as

$$|\boldsymbol{A}| = n^m \prod_{i=1}^{k} \prod_{j=i+1}^{d} \left(\tilde{l}_j^{-1} - \tilde{l}_i^{-1}\right)(l_i - l_j) \quad \text{where} \quad \tilde{l}_i = \begin{cases} l_i, & \text{if } i \le k , \\ \frac{1}{d-k}\sum_{j=k+1}^{d} l_j, & \text{if } k < i \le d . \end{cases}$$

The number of principal components to retain is taken to be the value of $k = 1, \ldots, d$ for with $p(\mathcal{D} \mid k)$ is maximized. Dropping all the terms that do not grow with n a BIC approximation was also proposed in Minka (2001)

$$p(\mathcal{D} \mid k) \approx n^{-\frac{m+k}{2}} \rho(k)^{-n(d-k)} \prod_{i=1}^{k} l_i^{-\frac{n}{2}} \prod_{i=k+1}^{d} l_i^n . \qquad (16)$$

The following theorem corrects the existing MIBS criterion (15) and serves as a guideline for easy implementation of the novel criterion proposed therein.

Theorem: Consider the probabilistic principal component model (1) with a prior for the parameters $\boldsymbol{H}$, $\boldsymbol{\mu}$, and σ^2 as discussed in Section 3.1

$$p(\boldsymbol{H}, \boldsymbol{\mu}, \sigma^2) \propto \left|\boldsymbol{H}\boldsymbol{H}^T + \sigma^2 \boldsymbol{I}_d\right|^{-\frac{\alpha+2}{2}} \exp\left\{-\frac{\alpha}{2}\mathrm{tr}\left[\left(\boldsymbol{H}\boldsymbol{H}^T + \sigma^2 \boldsymbol{I}_d\right)^{-1}\right]\right\} , \qquad (17)$$

where $\alpha > 0$ is the prior parameter. By applying Laplace approximation the marginal likelihood of $\mathcal{D} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ for a model with $k < d$ principal components is

$$p(\mathcal{D} \mid k) \approx 2^k c_k \left|\hat{\boldsymbol{\Lambda}}\right|^{-\frac{N}{2}+1} \hat{\sigma}^{-N(d-k)+2} \exp\left\{-\tfrac{Nd}{2}+k+1\right\} (2\pi)^{\frac{m+k+1}{2}} (|\boldsymbol{A}_{\boldsymbol{U}}||\boldsymbol{A}_{\boldsymbol{\Lambda}}||\boldsymbol{A}_{\sigma^2}|)^{-\frac{1}{2}} , \qquad (18)$$

where $m = dk - k(k+1)/2$, $N = n + 1 + \alpha$, $\hat{\boldsymbol{\Lambda}}$ is a diagonal matrix with entries $\hat{\lambda}_i$, $i = 1, \ldots, k$, and

$$c_k = \frac{n^{-\frac{d}{2}} (2\pi)^{-\frac{(n-1)d}{2}} \pi^{\frac{1}{4}k(k-1-2d)}}{2^k \Gamma\left(\frac{(\alpha+2)(d-k)}{2} - 1\right) \Gamma\left(\frac{\alpha}{2}\right)^k} \left(\frac{\alpha(d-k)}{2}\right)^{\frac{(\alpha+2)(d-k)-2}{2}} \left(\frac{\alpha}{2}\right)^{\frac{k\alpha}{2}} \prod_{i=1}^{k} \Gamma\left(\frac{d-i+1}{2}\right) . \qquad (19)$$

Moreover, if $l_1, \ldots, l_d$ denote the eigenvalues of $\hat{\boldsymbol{S}} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})^T$, where $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$, we have

$$|\boldsymbol{A}_{\sigma^2}| = \frac{N(d-k)-2}{2}, \qquad \hat{\sigma}^2 = \frac{n\sum_{i=k+1}^{d} l_i}{N(d-k)-2},$$
$$|\boldsymbol{A}_{\boldsymbol{\Lambda}}| = \left(\frac{N}{2}-1\right)^k, \qquad \hat{\lambda}_i = \frac{nl_i+\alpha}{N-2},$$

$$|\boldsymbol{A}_{\boldsymbol{U}}| = n^m \prod_{i=1}^{k}\prod_{j=i+1}^{d}\left(\tilde{\lambda}_j^{-1}-\tilde{\lambda}_i^{-1}\right)(l_i-l_j) \quad \text{where} \quad \tilde{\lambda}_i = \begin{cases} \hat{\lambda}_i, & \text{if } i \le k, \\ \hat{\sigma}^2, & \text{if } k < i \le d. \end{cases}$$

The value for α is typically chosen very small to make the prior less informative. Large values for α may lead to inferior results especially when the number of samples is small.

3.3 Proof of the Corrected MIBS Criterion

The marginal likelihood can be written as

$$p(\mathcal{D} \mid k) = c_k \int \left|\boldsymbol{H}\boldsymbol{H}^T + \sigma^2\boldsymbol{I}_d\right|^{-\frac{N}{2}} \times \exp\left\{-\frac{n}{2}\mathrm{tr}\left[\left(\boldsymbol{H}\boldsymbol{H}^T+\sigma^2\boldsymbol{I}_d\right)^{-1}\left(\hat{\boldsymbol{S}}+\frac{\alpha}{n}\boldsymbol{I}_d\right)\right]\right\} \mathrm{d}\boldsymbol{U}\mathrm{d}\boldsymbol{\Lambda}\mathrm{d}\sigma^2, \quad (20)$$

where $N = n+1+\alpha$. The multiplier c_k defined in (19) combines the constants of the prior (12) and the integrated likelihood (11). To evaluate the integral in (20) we use Laplace approximation as described in Lindley (1980). The matrix $\boldsymbol{U} \in \mathbb{R}^{d\times k}$ has $m = dk - k(k+1)/2$ free parameters since we have $k(k+1)$ constraints because of the orthonormality condition for the columns of $\boldsymbol{U}$. Laplace approximation can therefore be written as

$$\int p(\mathcal{D}, \boldsymbol{U}, \boldsymbol{\Lambda}, \sigma^2 \mid k)\mathrm{d}\boldsymbol{U}\mathrm{d}\boldsymbol{\Lambda}\mathrm{d}\sigma^2 \approx p(\mathcal{D}, \hat{\boldsymbol{U}}, \hat{\boldsymbol{\Lambda}}, \hat{\sigma}^2 \mid k)(2\pi)^{\frac{k+m+1}{2}} \left|\boldsymbol{A}_{\boldsymbol{U}\boldsymbol{\Lambda}\sigma^2}\right|^{-\frac{1}{2}}, \quad (21)$$

where $\left(\hat{\boldsymbol{U}}, \hat{\boldsymbol{\Lambda}}, \hat{\sigma}^2\right) = \arg\max_{\boldsymbol{U},\boldsymbol{\Lambda},\sigma^2} p(\mathcal{D}, \boldsymbol{U}, \boldsymbol{\Lambda}, \sigma^2 \mid k)$ and $\boldsymbol{A}_{\boldsymbol{U}\boldsymbol{\Lambda}\sigma^2}$ is the negative Hessian of the integrand at $\left(\hat{\boldsymbol{U}}, \hat{\boldsymbol{\Lambda}}, \hat{\sigma}^2\right)$. It is important to know that Laplace approximation is a basis dependent method, so we have to choose an appropriate parameterization, $\left(\boldsymbol{Z}, \acute{\boldsymbol{\Lambda}}, \acute{\sigma}^2\right)$, for $(\boldsymbol{U}, \boldsymbol{\Lambda}, \sigma^2)$. In the case of $\boldsymbol{\Lambda}$ and σ^2 it is easy to guess a good parameterization, since Laplace approximation works better when using integrals over the whole real axis than just over the positive real numbers. Therefore, one can use $\acute{\sigma}^2 = \log\sigma^2$ and $\acute{\lambda}_i = \log\lambda_i$ for all $i = 1, \ldots, k$, where $\acute{\lambda}_i$ are the diagonal elements of $\acute{\boldsymbol{\Lambda}}$. The Jacobian matrices for the transformations are

$$\boldsymbol{J}_{\boldsymbol{\Lambda}} = \frac{\partial(\lambda_1, \ldots, \lambda_k)}{\partial\left(\acute{\lambda}_1, \ldots, \acute{\lambda}_k\right)} = \begin{pmatrix} \exp\{\acute{\lambda}_1\} & & \\ & \ddots & \\ & & \exp\{\acute{\lambda}_k\} \end{pmatrix} = \boldsymbol{\Lambda}, \qquad \boldsymbol{J}_{\sigma^2} = \sigma^2.$$

The matrix $\boldsymbol{U}$ is expressed in Euler vector coordinates

$$\boldsymbol{U} = \boldsymbol{U}_d \exp\{\boldsymbol{Z}\} \begin{bmatrix} \boldsymbol{I}_k \\ \boldsymbol{0} \end{bmatrix}, \tag{22}$$

where $\boldsymbol{Z} \in \mathbb{R}^{d\times d}$ is a skew-symmetric matrix of parameters and $\boldsymbol{U}_d \in \mathbb{R}^{d\times d}$ is a fixed orthogonal matrix. The free parameters of $\boldsymbol{Z}$ are the first k rows of the upper triangle: $d(d-1)/2 - (d-k)(d-k-1)/2 = m$, as desired. If $\boldsymbol{\Lambda}$ and σ^2 are fixed, we can use the result in (7) and write $p(\mathcal{D}, \boldsymbol{U} \mid k, \boldsymbol{\Lambda}, \sigma^2)$ as

$$p(\mathcal{D}, \boldsymbol{U} \mid k, \boldsymbol{\Lambda}, \sigma^2) \propto \exp\left\{-\frac{n}{2}\mathrm{tr}\left[\left(\boldsymbol{\Lambda}^{-1} - \sigma^{-2}\boldsymbol{I}_k\right)\boldsymbol{U}^T\hat{\boldsymbol{S}}\boldsymbol{U}\right]\right\},$$

which is known to be the von Mises-Fisher matrix distribution (see Khatri and Mardia, 1977). This density is maximized for $\boldsymbol{U}$ at $\hat{\boldsymbol{U}}$, which contains the first k eigenvectors of $\hat{\boldsymbol{S}}$. In the parameterization (22) this corresponds to the case where $\hat{\boldsymbol{Z}} = \boldsymbol{0}$ and $\boldsymbol{U}_d$ is equal to the matrix of eigenvectors of $\hat{\boldsymbol{S}}$. Note, that the density has the same value if the sign of a column of $\boldsymbol{U}$ is changed. This can happen 2^k times, so the density has 2^k extreme points with the same function values and we need to multiply (21) with 2^k. The determinant of the Jacobian for the transformation (22) at $\hat{\boldsymbol{Z}}$ is obviously equal to 1.

Using the same argument as in (7) it follows that

$$\mathrm{tr}\left[\left(\boldsymbol{H}\boldsymbol{H}^T + \sigma^2\boldsymbol{I}_d\right)^{-1}\left(\hat{\boldsymbol{S}} + \frac{\alpha}{n}\boldsymbol{I}_d\right)\right] = \frac{1}{\sigma^2}\mathrm{tr}\left[\hat{\boldsymbol{S}} - \boldsymbol{U}^T\hat{\boldsymbol{S}}\boldsymbol{U}\right] + \mathrm{tr}\left[\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^T\hat{\boldsymbol{S}}\boldsymbol{U} + \frac{\alpha}{n}\boldsymbol{\Lambda}^{-1}\right], \tag{23}$$

and if $\boldsymbol{L} = \mathrm{diag}(l_1, \ldots, l_k)$ denotes the diagonal matrix containing the first k eigenvalues of $\hat{\boldsymbol{S}}$, we can rewrite $\log p\big(\mathcal{D}, \hat{\boldsymbol{Z}}, \acute{\boldsymbol{\Lambda}}, \acute{\sigma}^2 \mid k\big)$ as

$$\begin{aligned}\log p\left(\mathcal{D}, \hat{\boldsymbol{Z}}, \acute{\boldsymbol{\Lambda}}, \acute{\sigma}^2 \mid k\right) \propto{}& \left(-\frac{N}{2} + 1\right)\log|\boldsymbol{\Lambda}| + \left(-\frac{N(d-k)}{2} + 1\right)\log\sigma^2 \\ & -\frac{n}{2}\left(\frac{1}{\sigma^2}\left(\mathrm{tr}\big[\hat{\boldsymbol{S}}\big] - \mathrm{tr}\big[\boldsymbol{L}\big]\right) + \mathrm{tr}\left[\boldsymbol{\Lambda}^{-1}\boldsymbol{L}\right] + \frac{\alpha}{n}\mathrm{tr}\left[\boldsymbol{\Lambda}^{-1}\right]\right).\end{aligned}$$

The estimate for $\boldsymbol{\Lambda}$ can be calculated by differentiation of the latter equation:

$$\begin{aligned}\frac{\partial \log p\left(\mathcal{D}, \hat{\boldsymbol{Z}}, \acute{\boldsymbol{\Lambda}}, \acute{\sigma}^2 \mid k\right)}{\partial \acute{\lambda}_i} &= -\frac{N}{2} + 1 + \frac{nl_i}{2\exp\{\acute{\lambda}_i\}} + \frac{\alpha}{2\exp\{\acute{\lambda}_i\}} \\ &= -\frac{N}{2} + 1 + \frac{nl_i + \alpha}{2\lambda_i} \quad \forall i = 1, \ldots, k\,.\end{aligned}$$

If we set this derivative to zero, we get that $\exp\{\acute{\lambda}_i\} = \frac{nl_i+\alpha}{n-1+\alpha}$ or equivalently

$$\hat{\lambda}_i = \frac{nl_i + \alpha}{N - 2} \quad \forall i = 1, \ldots, k\,.$$

The cross derivatives are zero and the second derivatives with respect to $\acute{\lambda}_i$ are

$$\frac{\partial^2 \log p\left(\mathcal{D}, \hat{\boldsymbol{Z}}, \acute{\boldsymbol{\Lambda}}, \acute{\sigma}^2 \mid k\right)}{\partial \acute{\lambda}_i^2} = -\frac{nl_i + \alpha}{2\exp\{\acute{\lambda}_i\}} = -\frac{nl_i + \alpha}{2\lambda_i} \quad \forall i = 1, \ldots, k\,.$$

The negative Hessian of $\log p\big(\mathcal{D}, \hat{\boldsymbol{Z}}, \acute{\boldsymbol{\Lambda}}, \acute{\sigma}^2 \mid k\big)$ at $\boldsymbol{\Lambda} = \hat{\boldsymbol{\Lambda}} = \operatorname{diag}\big(\hat{\lambda}_1, \ldots, \hat{\lambda}_k\big)$ can now be calculated as

$$|\boldsymbol{A}_{\boldsymbol{\Lambda}}| = \left(\frac{N}{2} - 1\right)^k . \tag{24}$$

The estimate of σ^2 is obtained analogously:

$$\begin{aligned}
\frac{\partial \log p\left(\mathcal{D}, \hat{\boldsymbol{Z}}, \hat{\boldsymbol{\Lambda}}, \acute{\sigma}^2 \mid k\right)}{\partial \acute{\sigma}^2} &= -\frac{N(d-k)}{2} + 1 + \frac{n}{2\acute{\sigma}^2}\left(\operatorname{tr}\big[\hat{\boldsymbol{S}}\big] - \operatorname{tr}\big[\boldsymbol{L}\big]\right) , \\
\hat{\sigma}^2 &= \frac{n\left(\operatorname{tr}\big[\hat{\boldsymbol{S}}\big] - \operatorname{tr}\big[\boldsymbol{L}\big]\right)}{N(d-k)-2} = \frac{n\sum_{i=k+1}^{d} l_i}{N(d-k)-2} \\
\frac{\partial^2 \log p\left(\mathcal{D}, \hat{\boldsymbol{Z}}, \hat{\boldsymbol{\Lambda}}, \acute{\sigma}^2 \mid k\right)}{\partial \acute{\sigma}^4} &= -\frac{n}{2\acute{\sigma}^2}\left(\operatorname{tr}\big[\hat{\boldsymbol{S}}\big] - \operatorname{tr}\big[\boldsymbol{L}\big]\right) , \\
\left.\frac{\partial^2 \log p\left(\mathcal{D}, \hat{\boldsymbol{Z}}, \hat{\boldsymbol{\Lambda}}, \acute{\sigma}^2 \mid k\right)}{\partial \acute{\sigma}^4}\right|_{\sigma^2=\hat{\sigma}^2} &= -\frac{N(d-k)-2}{2} = -|\boldsymbol{A}_{\sigma^2}| .
\end{aligned} \tag{25}$$

Substituting the estimates $\hat{\boldsymbol{Z}}$, $\hat{\boldsymbol{\Lambda}}$ and $\hat{\sigma}^2$ into (23) we get that

$$(23)\,|_{\hat{\boldsymbol{Z}},\hat{\boldsymbol{\Lambda}},\hat{\sigma}^2} = \frac{N(d-k)-2}{n} + \sum_{i=1}^{k}\frac{l_i(N-2)}{nl_i+\alpha} + \frac{\alpha}{n}\sum_{i=1}^{k}\frac{N-2}{nl_i+\alpha} = \frac{Nd-2k-2}{n} ,$$

which leads to the following equation for $p\big(\mathcal{D}, \hat{\boldsymbol{Z}}, \hat{\boldsymbol{\Lambda}}, \hat{\sigma}^2 \mid k\big)$:

$$p\left(\mathcal{D}, \hat{\boldsymbol{Z}}, \hat{\boldsymbol{\Lambda}}, \hat{\sigma}^2 \mid k\right) = c_k\big|\hat{\boldsymbol{\Lambda}}\big|^{-\frac{N}{2}+1}\hat{\sigma}^{-N(d-k)+2}\exp\left\{-\frac{Nd}{2}+k+1\right\} .$$

Next, we have to evaluate the Hessian of $\log p\big(\mathcal{D}, \boldsymbol{Z}, \hat{\boldsymbol{\Lambda}}, \hat{\sigma}^2 \mid k\big)$ at $\hat{\boldsymbol{Z}}$. Using

$$\exp\{\boldsymbol{Z}\} = \sum_{i=0}^{\infty}\frac{\boldsymbol{Z}^i}{i!} = \boldsymbol{I}_d + \boldsymbol{Z} + \frac{\boldsymbol{Z}^2}{2} + \frac{\boldsymbol{Z}^3}{6} + \cdots$$

and certain matrix differential rules we obtain the first and second differential of $\boldsymbol{U}$:

$$\begin{aligned}
\mathrm{d}\boldsymbol{U} &= \boldsymbol{U}_d\big(\mathrm{d}\boldsymbol{Z} + \frac{1}{2}\left(\mathrm{d}\boldsymbol{Z}\boldsymbol{Z} + \boldsymbol{Z}\mathrm{d}\boldsymbol{Z}\right) + \cdots\big)\begin{bmatrix}\boldsymbol{I}_k \\ \boldsymbol{0}\end{bmatrix} , \\
\mathrm{d}\boldsymbol{U}\,|_{\boldsymbol{Z}=\boldsymbol{0}} &= \boldsymbol{U}_d\mathrm{d}\boldsymbol{Z}\begin{bmatrix}\boldsymbol{I}_k \\ \boldsymbol{0}\end{bmatrix} ,
\end{aligned} \tag{26}$$

$$\begin{aligned}
\mathrm{d}^2\boldsymbol{U} &= \boldsymbol{U}_d\big(\mathrm{d}\left(\mathrm{d}\boldsymbol{Z}\right) + \frac{1}{2}\left(\mathrm{d}\boldsymbol{Z}\mathrm{d}\boldsymbol{Z} + \mathrm{d}\boldsymbol{Z}\mathrm{d}\boldsymbol{Z}\right) + \cdots\big)\begin{bmatrix}\boldsymbol{I}_k \\ \boldsymbol{0}\end{bmatrix} , \\
\mathrm{d}^2\boldsymbol{U}\,|_{\boldsymbol{Z}=\boldsymbol{0}} &= \boldsymbol{U}_d\left(\mathrm{d}\boldsymbol{Z}\right)^2\begin{bmatrix}\boldsymbol{I}_k \\ \boldsymbol{0}\end{bmatrix} .
\end{aligned} \tag{27}$$

The differential of $\log p\big(\mathcal{D}, \boldsymbol{Z}, \hat{\boldsymbol{\Lambda}}, \hat{\sigma}^2 \mid k\big)$ can be evaluated as

$$\begin{aligned}
\mathrm{d}\log p\left(\mathcal{D}, \boldsymbol{Z}, \hat{\boldsymbol{\Lambda}}, \hat{\sigma}^2 \mid k\right) &= -\frac{n}{2}\mathrm{tr}\left[\mathrm{d}\left(\left(\hat{\boldsymbol{\Lambda}}^{-1} - \hat{\sigma}^{-2}\boldsymbol{I}_k\right)\boldsymbol{U}^T\hat{\boldsymbol{S}}\boldsymbol{U}\right)\right] \\
&= -\frac{n}{2}\mathrm{tr}\left[\left(\hat{\boldsymbol{\Lambda}}^{-1} - \hat{\sigma}^{-2}\boldsymbol{I}_k\right)\left(\mathrm{d}\boldsymbol{U}^T\hat{\boldsymbol{S}}\boldsymbol{U} + \boldsymbol{U}^T\hat{\boldsymbol{S}}\mathrm{d}\boldsymbol{U}\right)\right] \\
&= -n\mathrm{tr}\left[\left(\hat{\boldsymbol{\Lambda}}^{-1} - \hat{\sigma}^{-2}\boldsymbol{I}_k\right)\boldsymbol{U}^T\hat{\boldsymbol{S}}\mathrm{d}\boldsymbol{U}\right]. \qquad (28)
\end{aligned}$$

The second differential is obtained by taking the differential of (28):

$$\mathrm{d}^2\log p\left(\mathcal{D}, \boldsymbol{Z}, \hat{\boldsymbol{\Lambda}}, \hat{\sigma}^2 \mid k\right) = -n\mathrm{tr}\left[\left(\hat{\boldsymbol{\Lambda}}^{-1} - \hat{\sigma}^{-2}\boldsymbol{I}_k\right)\left(\mathrm{d}\boldsymbol{U}^T\hat{\boldsymbol{S}}\mathrm{d}\boldsymbol{U} + \boldsymbol{U}^T\hat{\boldsymbol{S}}\mathrm{d}^2\boldsymbol{U}\right)\right]. \qquad (29)$$

Since $\boldsymbol{U}_d^T\hat{\boldsymbol{S}}\boldsymbol{U}_d = \boldsymbol{L}$ and $\hat{\boldsymbol{U}}^T\hat{\boldsymbol{S}}\boldsymbol{U}_d = \left[\begin{smallmatrix}\boldsymbol{I}_k\\ \boldsymbol{0}\end{smallmatrix}\right]^T\boldsymbol{L}$, just containing the top k rows of $\boldsymbol{L}$, we can rewrite (29) evaluated at $\hat{\boldsymbol{Z}}$ by applying (26) and (27) as

$$-n\mathrm{tr}\left[\begin{bmatrix}\boldsymbol{I}_k\\ \boldsymbol{0}\end{bmatrix}\left(\hat{\boldsymbol{\Lambda}}^{-1} - \hat{\sigma}^{-2}\boldsymbol{I}_k\right)\begin{bmatrix}\boldsymbol{I}_k\\ \boldsymbol{0}\end{bmatrix}^T\left(\mathrm{d}\boldsymbol{Z}^T\boldsymbol{L}\mathrm{d}\boldsymbol{Z} + \boldsymbol{L}\mathrm{d}\boldsymbol{Z}\mathrm{d}\boldsymbol{Z}\right)\right]. \qquad (30)$$

If we define the diagonal matrix $\boldsymbol{B} = \left[\begin{smallmatrix}\boldsymbol{I}_k\\ \boldsymbol{0}\end{smallmatrix}\right]\left(\hat{\boldsymbol{\Lambda}}^{-1} - \hat{\sigma}^{-2}\boldsymbol{I}_k\right)\left[\begin{smallmatrix}\boldsymbol{I}_k\\ \boldsymbol{0}\end{smallmatrix}\right]^T$ and $\boldsymbol{T} = \boldsymbol{B}\mathrm{d}\boldsymbol{Z}^T + \mathrm{d}\boldsymbol{Z}\boldsymbol{B}$, we can simplify (30). Using the skew-symmetry of $\boldsymbol{Z}$ we get $\boldsymbol{T} = \mathrm{d}\boldsymbol{Z}\boldsymbol{B} - \boldsymbol{B}\mathrm{d}\boldsymbol{Z}$ and

$$\mathrm{d}^2\log p\left(\mathcal{D}, \boldsymbol{Z}, \hat{\boldsymbol{\Lambda}}, \hat{\sigma}^2 \mid k\right)\Big|_{\boldsymbol{Z}=\boldsymbol{0}} = -n\mathrm{tr}\left[\boldsymbol{T}\boldsymbol{L}\mathrm{d}\boldsymbol{Z}\right]. \qquad (31)$$

If we define a diagonal matrix $\tilde{\boldsymbol{\Lambda}}$ with entries $\tilde{\lambda}_i$, $i = 1, \ldots, d$, as

$$\tilde{\boldsymbol{\Lambda}} = \begin{pmatrix}\hat{\boldsymbol{\Lambda}} & \boldsymbol{0}\\ \boldsymbol{0} & \hat{\sigma}^2\boldsymbol{I}_k\end{pmatrix},$$

we get from the definition of $\boldsymbol{B}$ and $\boldsymbol{T}$ that the ij-th element of $\boldsymbol{T}$ can be written as $t_{ij} = \big(\tilde{\lambda}_j^{-1} - \tilde{\lambda}_i^{-1}\big)\mathrm{d}z_{ij}$. The next step is to show that (31) can be simplified to

$$\mathrm{d}^2\log p\left(\mathcal{D}, \boldsymbol{Z}, \hat{\boldsymbol{\Lambda}}, \hat{\sigma}^2 \mid k\right)\Big|_{\boldsymbol{Z}=\boldsymbol{0}} = -\sum_{i=1}^{k}\sum_{j=i+1}^{d} n\left(\tilde{\lambda}_j^{-1} - \tilde{\lambda}_i^{-1}\right)(l_i - l_j)\,\mathrm{d}z_{ij}^2. \qquad (32)$$

To prove (32) we note that the ij-th element of $\boldsymbol{T}\boldsymbol{L}$ is $\big(\boldsymbol{T}\boldsymbol{L}\big)_{ij} = \big(\tilde{\lambda}_j^{-1} - \tilde{\lambda}_i^{-1}\big)l_j\mathrm{d}z_{ij}$. This term is zero if $i, j > k$ or $i = j$. Hence, the elements of $\boldsymbol{T}\boldsymbol{L}\mathrm{d}\boldsymbol{Z}$ can be written as $(\boldsymbol{T}\boldsymbol{L}\mathrm{d}\boldsymbol{Z})_{ij} = \sum_{u=1}^{d}(\boldsymbol{T}\boldsymbol{L})_{iu}\mathrm{d}z_{uj}$ if $i \le k$ and $(\boldsymbol{T}\boldsymbol{L}\mathrm{d}\boldsymbol{Z})_{ij} = \sum_{u=1}^{k}(\boldsymbol{T}\boldsymbol{L})_{iu}\mathrm{d}z_{uj}$ if $i, j > k$. Taking into account the skew-symmetry of $\mathrm{d}\boldsymbol{Z}$ the trace of $\boldsymbol{T}\boldsymbol{L}\mathrm{d}\boldsymbol{Z}$ is

$$\begin{aligned}
\mathrm{tr}\left[\boldsymbol{T}\boldsymbol{L}\mathrm{d}\boldsymbol{Z}\right] &= -\sum_{i=1}^{k}\sum_{j=1}^{d}\left(\tilde{\lambda}_j^{-1} - \tilde{\lambda}_i^{-1}\right)l_j\mathrm{d}z_{ij}^2 - \sum_{i=k+1}^{d}\sum_{j=1}^{k}\left(\tilde{\lambda}_j^{-1} - \tilde{\lambda}_i^{-1}\right)l_j\mathrm{d}z_{ij}^2 \\
&= \sum_{i=1}^{k}\left(\sum_{j=k+1}^{d}\left(\tilde{\lambda}_j^{-1} - \tilde{\lambda}_i^{-1}\right)l_i\mathrm{d}z_{ij}^2 - \sum_{j=1}^{d}\left(\tilde{\lambda}_j^{-1} - \tilde{\lambda}_i^{-1}\right)l_j\mathrm{d}z_{ij}^2\right).
\end{aligned}$$

The latter equation can be rewritten as

$$\operatorname{tr}\left[\boldsymbol{T}\boldsymbol{L}\mathrm{d}\boldsymbol{Z}\right] = \sum_{i=1}^{k}\sum_{j=i+1}^{d}\left(\tilde{\lambda}_j^{-1}-\tilde{\lambda}_i^{-1}\right)(l_i-l_j)\,\mathrm{d}z_{ij}^2 - \sum_{i=1}^{k}\left(\sum_{j=1}^{i-1}\left(\tilde{\lambda}_j^{-1}-\tilde{\lambda}_i^{-1}\right)l_j\mathrm{d}z_{ij}^2 + \sum_{j=i+1}^{k}\left(\tilde{\lambda}_j^{-1}-\tilde{\lambda}_i^{-1}\right)l_i\mathrm{d}z_{ij}^2\right). \quad (33)$$

It remains to show that the second term of (33) is zero and we prove it using induction. Let $B\,(k)$ denote the value of this second term in (33). It is clear that $B(1)$ and $B(2)$ are both equal to zero. Let us assume that the claim holds for $B(s)$ with a certain $s \geq 2$. For $s+1$ simple calculation yields

$$B(s+1) = B(s) + \sum_{j=1}^{s}\left(\tilde{\lambda}_j^{-1}-\tilde{\lambda}_{s+1}^{-1}\right)l_j\mathrm{d}z_{s+1,j}^2 + \sum_{i=1}^{s}\left(\tilde{\lambda}_{s+1}^{-1}-\tilde{\lambda}_i^{-1}\right)l_i\mathrm{d}z_{s+1,i}^2\,,$$

where the first term on the right side of the equation is zero due to the induction hypothesis and the sum of the last two terms is clearly also equal to zero.

With this we have shown that the cross derivatives are zero and therefore the Hessian of $\log p\big(\mathcal{D}, \boldsymbol{Z}, \hat{\boldsymbol{\Lambda}}, \hat{\sigma}^2 \mid k\big)$ is diagonal at $\boldsymbol{Z} = \boldsymbol{0}$. Using (32), the determinant of the negative Hessian with respect to $\boldsymbol{Z}$, required in (21), is

$$|\boldsymbol{A}_{\boldsymbol{U}}| = n^m \prod_{i=1}^{k}\prod_{j=i+1}^{d}\left(\tilde{\lambda}_j^{-1}-\tilde{\lambda}_i^{-1}\right)(l_i-l_j)\,. \quad (34)$$

Since all the cross derivatives between $\hat{\boldsymbol{\Lambda}}$, $\hat{\sigma}^2$ and $\boldsymbol{Z}$ are also zero, the negative Hessian of $\log p\big(\mathcal{D}, \boldsymbol{Z}, \hat{\boldsymbol{\Lambda}}, \hat{\sigma}^2 | k\big)$ is block diagonal. Its determinant equals $|\boldsymbol{A}_{\boldsymbol{U}\boldsymbol{\Lambda}\sigma^2}| = |\boldsymbol{A}_{\boldsymbol{U}}||\boldsymbol{A}_{\boldsymbol{\Lambda}}||\boldsymbol{A}_{\sigma^2}|$, where the factors are given in (24), (25), and (34), respectively.

Now that we have all the terms and estimates which are needed in (21), Laplace approximation of the integral directly leads to

$$p(\mathcal{D} \mid k) \approx 2^k c_k \left|\hat{\boldsymbol{\Lambda}}\right|^{-\frac{N}{2}+1} \hat{\sigma}^{-N(d-k)+2} \exp\left\{-\tfrac{Nd}{2}+k+1\right\} (2\pi)^{\frac{m+k+1}{2}} \left(|\boldsymbol{A}_{\boldsymbol{U}}||\boldsymbol{A}_{\boldsymbol{\Lambda}}||\boldsymbol{A}_{\sigma^2}|\right)^{-\frac{1}{2}},$$

which is our recommended criterion. In this formula a few errors of Minka (2001) have been corrected. In the original paper the term $\exp\{k+1\}$ and the Jacobians of the transformations were missing.

4 Simulation Study

In this section we want to test the performance of the different criteria for selecting the optimum number of principal components using synthetic data. In a simulation study we compare four criteria that make use of the probabilistic PCA model: the original MIBS criterion (15), its corrected version (18), the BIC (16) and the orthogonal variational Bayes approximation, called OVPCA (see Smidl and Quinn, 2005). Additionally we consider the AIC (13) and MDL (14) criteria.

The prior parameter α for the corrected MIBS criterion is chosen to be $\alpha = 0.01$ to make the prior diffuse. The subsequently presented results for the simulation study would not change much as long as the the value for α is kept small, say smaller than 1. Large values for α, however, would adversely affect the results, especially for small samples sizes when the prior information dominates the information obtained from the likelihood.

In the first experiment the data is sampled from a 10-dimensional zero-mean Gaussian distribution with variances [10 8 6 4 2] in the first five directions. The remaining five directions have a variance equal to 1. The results for samples sizes varying between 5 and 350 are shown in Figure 1(a). For every sample size we simulate 1000 data sets and count how often the true dimensionality is recovered by a certain criterion. One can see that AIC is the most accurate criterion for sample sizes smaller than 75, however, fails to increase the recovery rate when the number of samples gets large. The AIC is an inconsistent estimator for the true number of dimensionality which is visible in nearly all subsequent experiments. The corrected MIBS criterion, abbreviated by *corMIBS* in Figures 1(a) - (h), has excellent recovery rates for both small and large number of sampled data. Notably, it gives far better results than the original MIBS criterion for all different sample sizes. MDL turns out to be less accurate than the corrected and the original MIBS criterion but outperforms the BIC. The OVPCA shows the worst performance when there are less than 100 data samples but catches up to BIC and even MDL when we further increase the sample size.

The second experiment, displayed in Figure 1(b), is the same as the first except that the noise dimension is changed from 5 to 10. Again, the corrected MIBS criterion performs best and gives larger recovery rates than the original MIBS criterion. Compared to the first experiment, the MDL criterion and BIC need much more samples to reach a high recovery rate. This effect is far smaller for both MIBS criteria and also for the OVPCA criterion, which leads to superior results compared to MDL and BIC in this experiment.

The third experiment differs from the second experiment only by the fact that here we have a noise variance of 0.5 instead of 1. Compared to the second experiment, all criteria need less data to reach high recovery rates. The corrected MIBS criterion is the top performer for all sample sizes, however, the difference to the original MIBS criterion is smaller than in the first two experiments. Again, OVPCA is weak when the number of samples is small. Detailed results are shown in Figure 1(c).

In the fourth experiment the noise variance is reduced even more and set equal to 0.1. As can be seen in Figure 1(d), all criteria yield recovery rates higher than 95% when the sample size exceeds 25. The corrected MIBS criterion shows a far better performance than the original MIBS criterion for the very small sample sizes of 8 and 11. Note that this is the only experiment in which the BIC gives slightly better results than the MDL criterion. OVPCA is as good as the MIBS criteria when the number of samples is 17.

The fifth experiment is the same as the third except that here we have 20 noise dimensions instead of 10. The facts already mentioned in the second experiment also apply here. Figure 1(e) shows that the MDL criterion and BIC have a weak performance compared to the third experiment due to the added noise dimensions. The MIBS criteria and the OVPCA criterion are much less affected. Again, the corrected MIBS criterion gives the best results, especially for small sample sizes.

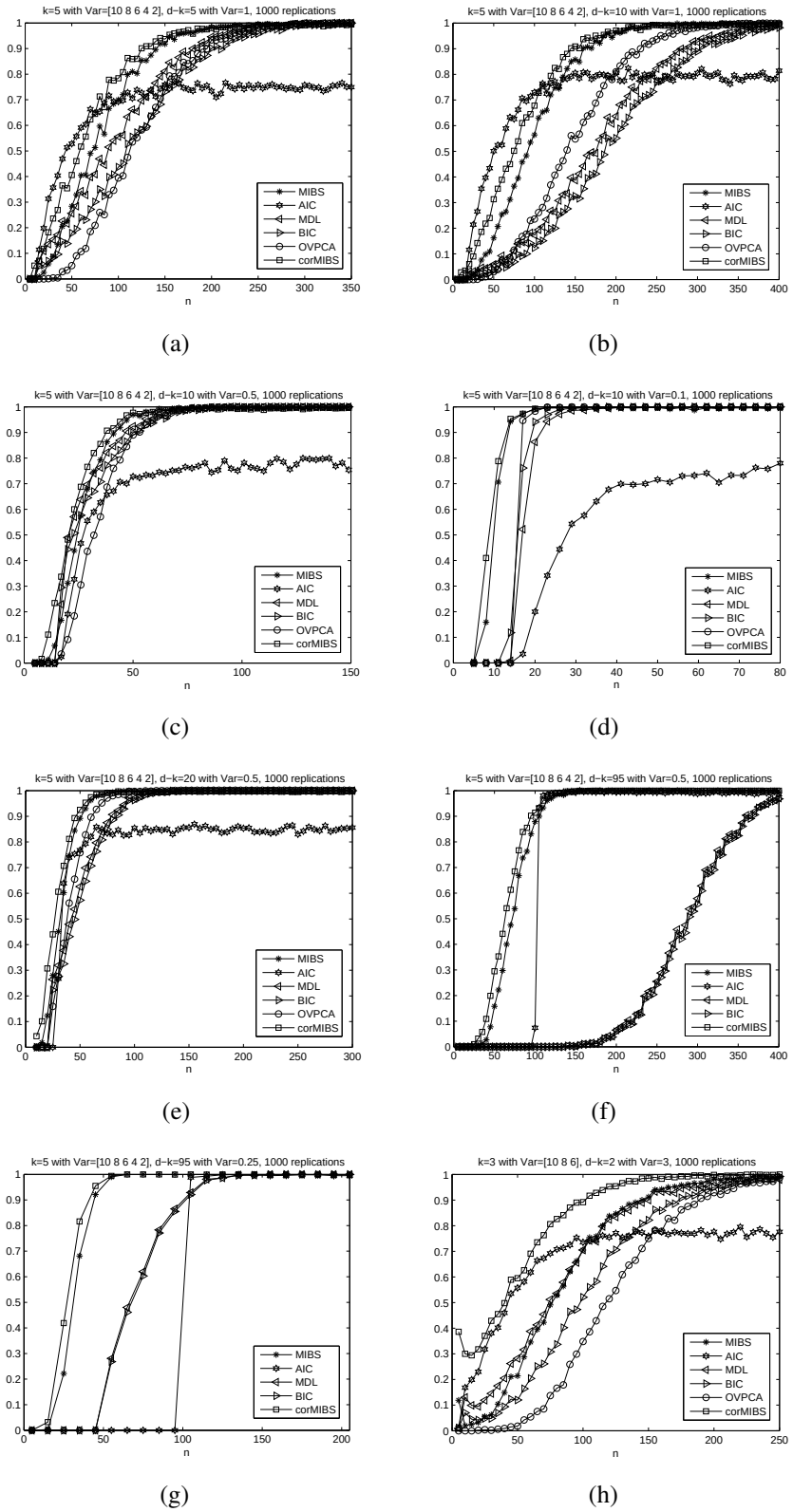

Figure 1: Recovery rates for the different criteria in the simulation study. The results of experiments 1-8 are displayed in (a)-(h).

The sixth and the seventh experiment, displayed in Figure 1(f)-(g), examine the case where the number of noise dimensions is 95 and, therefore, very large compared to the true dimensionality of the data which is equal to 5. When the data dimension is large the OVPCA is a very time-consuming method. This is the reason why we omit it for these two experiments. The corrected MIBS criterion is the most accurate method, followed by the original MIBS criterion. The AIC shows good results for sample sizes larger than 100. As we would expect, both MDL and BIC show a poor performance because of the large data dimensionality.

The last experiment is designed to test the case in which the true dimensionality is larger than the number of noise dimensions. The data is generated from a five-dimensional Gaussian distribution with variances [10 8 6] in the first three directions. The remaining two directions have a variance equal to 3. Again, the corrected MIBS criterion leads to the best results. However, in this experiment the difference in terms of recovery rate between the corrected MIBS criterion and all other criteria is more striking. The original MIBS criterion and the MDL criterion perform equally well. The BIC and especially the OVPCA give poor results.

5 Analysis of Hyper-Spectral Skin Cancer Data

If spectral measurements using hundreds of narrow contiguous wavelength intervals are performed, the resulting image is called a hyper-spectral image and is often represented as a hyper-spectral image cube. In contrast to RGB-images, where every pixel can be represented as a three-dimensional vector with entries corresponding to the red, green and blue channels, a hyper-spectral image contains pixels represented as multidimensional vectors with elements indicating the reflectivity at a specific wavelength. Thus, these vectors correspond to spectra in the physical meaning and are equal to spectra measured with e.g. spectrometers.

A set of 310 hyper-spectral images (171×170 pixels and 270 spectral bands after preprocessing steps) of malign and benign lesions were taken in clinical studies at the Medical University Graz, Austria. They are classified as melanoma or mole by human experts on the basis of a histological examination. Kazianka, Leitner, and Pilz (2008) used this data set and reported on the classification results for unobserved skin cancer images. In their paper they mention that as the dimensionality of the data equals the number of spectral bands, using the full spectral information in classification and clustering approaches leads to computational complexity. Moreover, the spectral bands are highly correlated and contain noise. To overcome the curse of dimensionality PCA is used to reduce the dimensions of the data, and inherently also the unwanted noise. Preceding analysis and inspection of scores and loadings showed that the optimum number of principal components to be retained is 7 (see Kazianka, 2007).

To test the performance of the AIC, BIC, MDL, MIBS and corrected MIBS criterion we selected 500 pixels from the training set. The prior parameter α for the corrected MIBS criterion is again set to $\alpha = 0.01$. The dimensionality picked by each method is shown in Table 1. The original MIBS and the corrected MIBS criterion choose the number of components that was suggested by the preceding analysis.

Table 1: Number of components picked by the different criteria for the skin cancer data.

Estimator	MIBS	corMIBS	MDL	AIC	BIC
No. Features	7	7	6	8	6

The values of the criteria for varying number of retained components are shown in Figure (2). As can be seen in Figure 2(a), the corrected MIBS criterion gives less probability mass for larger dimensionalities than the original formulation.

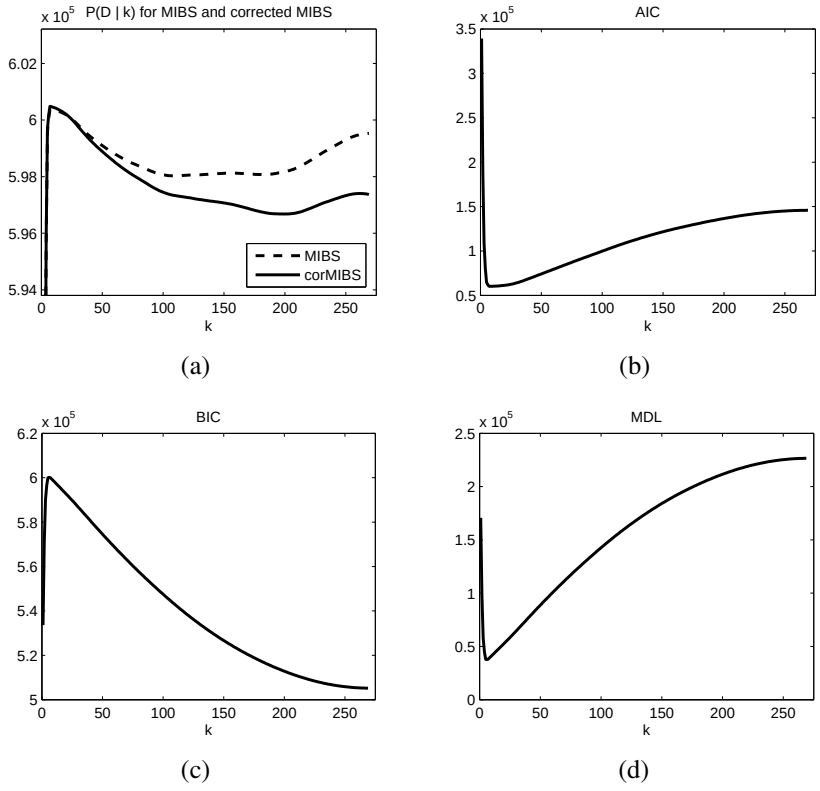

Figure 2: Values of the criteria for varying dimensionality in the skin cancer study.

6 Conclusion

The corrected version of the MIBS criterion for automatically selecting the optimum data dimensionality in the probabilistic PCA model shows excellent results in a simulation study. It is the constant top performer for all eight different experiments and works especially well with small sample sizes. The proposed criterion not only outperforms the original MIBS criterion but also other well-known methods such as BIC, AIC, MDL and OVPCA. The study also reveals that the performance of the BIC and the MDL criterion strongly depends on the data dimension. Both criteria give poor results when the number of dimensions is large. Besides the promising results for synthetic data the application to the hyper-spectral skin cancer images shows that the corrected MIBS criterion is also suitable for real world data which do not necessarily follow a Gaussian distribution.

Acknowledgements

The authors are grateful to Vaclav Smidl from the UTIA Prague for providing the code for the OVPCA method. We would also like to express our gratitude to CTR Carinthian Tech Research for providing the hyper-spectral skin cancer data. Furthermore, our special thanks go to the referee whose comments and suggestions helped a lot in preparing this final version of the paper.

References

Basilevsky, A. (1994). *Statistical Factor Analysis and Related Methods*. New York: Wiley.

Bishop, C. (2008). *Pattern Recognition and Machine Learning*. Berlin: Springer.

Cichocki, A., and Amari, S. (2002). *Adaptive Blind Signal and Image Processing*. Chichester: Wiley.

James, A. (1954). Normal Multivariate Analysis and the Orthogonal Group. *Annals of Mathematical Statistics*, *25*, 40-75.

Kazianka, H. (2007). *Classification Techniques for Hyper-Spectral Medical Image Data*. Unpublished master's thesis, University of Klagenfurt.

Kazianka, H., Leitner, R., and Pilz, J. (2008). Segmentation and Classification of Hyper-Spectral Skin Data. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker (Eds.), *Data Analysis, Machine Learning and Applications* (p. 245-252). Berlin: Springer.

Khatri, C., and Mardia, K. (1977). The von Mises-Fisher Distribution in Orientation Statistics. *Journal of the Royal Statistical Society, Series B*, *39*, 95-106.

Leonowicz, Z., Karvanen, J., Tanaka, T., and Rezmer, J. (2004). Model Order Selection Criteria: Comparative Study and Applications. In *Proceedings of the VIth International Workshop CPEE 2004* (p. 193-196). Warsaw: University of Technology.

Lindley, D. (1980). Approximate Bayesian Statistics Methods. In J. Bernardo, M. de Groot, D. Lindley, and A. Smith (Eds.), *Bayesian Statistics* (p. 223-237). Valencia: University Press.

Minka, T. (2001). Automatic Choice of Dimensionality for PCA. *Advances in Neural Information Processing Systems*, *13*, 598-604.

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, *2*, 559-572.

Smidl, V., and Quinn, A. (2005). *The Variational Bayes Method in Signal Processing*. Berlin: Springer.

Tipping, M., and Bishop, C. (1999). Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society, Series B*, *61*, 611-622.

Authors' address:
Hannes Kazianka and Jürgen Pilz
Institute of Statistics, University Klagenfurt
Universitätsstraße 65-67
9020 Klagenfurt
E-Mails: Hannes.Kazianka@uni-klu.ac.at