Choosing principal components: a new graphical method based on Bayesian model selection

Philipp Auer
University of Ulm

Daniel Gervini[1]
University of Wisconsin–Milwaukee

October 31, 2007

[1]Corresponding author. Email: gervini@uwm.edu. Address: P.O. Box 413, Milwaukee, WI 53201, USA.

Abstract

This article approaches the problem of selecting significant principal components from a Bayesian model selection perspective. The resulting Bayes rule provides a simple graphical technique that can be used instead of (or together with) the popular scree plot to determine the number of significant components to retain. We study the theoretical properties of the new method and show, by examples and simulation, that it provides more clear-cut answers than the scree plot in many interesting situations.

Key Words: Dimension reduction; scree plot; factor analysis; singular value decomposition.

MSC: Primary 62H25; secondary 62-09.

1 Introduction

Multivariate datasets usually contain redundant information, due to high correlations among the variables. Sometimes they also contain irrelevant information; that is, sources of variability that can be considered random noise for practical purposes. To obtain uncorrelated, significant variables, several methods have been proposed over the years. Undoubtedly, the most popular is Principal Component Analysis (PCA), introduced by Hotelling (1933); for a comprehensive account of the methodology and its applications, see Jolliffe (2002).

In many practical situations the first few components accumulate most of the variability, so it is common practice to discard the remaining components, thus reducing the dimensionality of the data. This has many advantages both from an inferential and from a descriptive point of view, since the leading components are often interpretable indices within the context of the problem and they are statistically more stable than subsequent components, which tend to be more unstructured and harder to estimate. The problem is, how many components should be retained?

Selection of principal components has been a recurring topic in PCA, and no consensus has emerged yet. Perhaps the most popular method is the scree plot of Cattell (1966), consisting of plotting the eigenvalues in decreasing order of magnitude and looking for an "elbow" in the graph. This method is simple and useful when it works, but the reader probably knows by experience that in many cases there is no recognizable "elbow" in the plot. More sophisticated methods, in a theoretical and computational sense, are those based on cross-validation (Wold, 1978; Krzanowski, 1987), sequential testing (Bartlett, 1950), and refinements of the scree plot like the "broken-stick" method (Legendre and Legendre, 1983). None of these alternatives have gained much popularity, partly because they are not always superior to the simple scree graph (see discussions in Jolliffe, 2002, ch. 6, and Ferré, 1995).

We think that the problem of selecting principal components is best framed as a model-selection problem, as others have pointed out (e.g. Bartlett, 1950; Anderson, 1984; Krzanowski, 1987). We can regard a m dimensional dataset as a combination of d dominant random factors plus unstructured random noise (with variance not necessarily small). The goal, then, is to tell apart the underlying d-dimensional structure from the random noise. The dimension d can be deter-

mined using a Bayesian model-selection procedure. For a certain combination of prior model probabilities and data distributions, we were able to derive a practical graphical method that, to a certain extent, is independent of its Bayesian motivation and can be used as a complement, or even a substitute, of the scree plot.

2 PCA and Bayesian model selection

Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^m$ be a random sample with finite mean $\boldsymbol{\mu}$ and covariance matrix Σ. The covariance matrix admits a spectral decomposition $\Sigma = \Gamma\Lambda\Gamma^T$, where Γ is orthonormal and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$. Without loss of generality we assume that $\lambda_1 \geq \cdots \geq \lambda_m$. We define $\mathcal{M}_d$ as the model with d dominant eigenvalues: $\lambda_1 \geq \cdots \geq \lambda_d$, $\lambda_d > \lambda_{d+1}$ and $\lambda_{d+1} = \cdots = \lambda_m$.

Note that the $m - d$ tail eigenvalues of model $\mathcal{M}_d$ are smaller than the leading ones, but not necessarily negligible; in fact, their accumulated variance can be substantial if m is large. Nevertheless, $\mathcal{M}_d$ formalizes the idea of "underlying d-dimensional structure plus random noise", since a random vector $\mathbf{x}$ with mean $\boldsymbol{\mu}$ and covariance matrix Σ can be decomposed, under $\mathcal{M}_d$, as

$$\mathbf{x} = \boldsymbol{\mu} + \sum_{k=1}^{d} z_k(\lambda_k - \lambda_{d+1})^{1/2}\boldsymbol{\gamma}_k + \lambda_{d+1}^{1/2}\boldsymbol{\varepsilon},$$

where $z_1, \ldots, z_d$ are uncorrelated random variables with zero mean and unit variance, and $\boldsymbol{\varepsilon}$ is random error with zero mean and identity covariance matrix; $\boldsymbol{\gamma}_k$ above denotes the kth column of Γ. Note that $\mathcal{M}_0$ is the unstructured no-component model (i.e. all the variables are uncorrelated with equal variances) and $\mathcal{M}_{m-1}$ is the full model with m distinguishable components; therefore all possible situations are included in the family of models $\{\mathcal{M}_d : d = 0, \ldots, m-1\}$.

The problem of selecting the number of components to retain has then been transformed into the problem of selecting the right model $\mathcal{M}_d$. We will assign a prior probability $p(d)$ to $\mathcal{M}_d$, with $\sum_{d=0}^{m-1} p(d) = 1$. The probability function $p(d)$ will be decreasing in d; the rationale for this is that, in the interest of parsimony, we should choose a model with many components only when the data strongly indicates so.

Let $p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\mathcal{M}_d)$ be the conditional likelihood under model $\mathcal{M}_d$. The Bayes

rule (for square loss) is

$$\delta(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \underset{0 \le d \le m-1}{\operatorname{argmax}} \, p(\mathcal{M}_d | \mathbf{x}_1, \ldots, \mathbf{x}_n). \tag{1}$$

That is, we select the model with largest posterior probability. In practice, we use an estimator $\hat{p}(\mathbf{x}_1, \ldots, \mathbf{x}_n | \mathcal{M}_d)$ of $p(\mathbf{x}_1, \ldots, \mathbf{x}_n | \mathcal{M}_d)$ and denote the resulting estimated rule by $\hat{d}$.

A neat expression for $\hat{d}$ can be obtained if we assume that the data is Normal and we take as prior probabilities

$$p(d) \propto \exp(-\frac{n}{2}\theta d), \;\; d = 0, \ldots, m-1,$$

for a given $\theta > 0$. Let $\{\hat{\lambda}_k\}$ and $\{\hat{\boldsymbol{\gamma}}_k\}$ be the eigenvalues and eigenvectors of the sample covariance matrix. The maximum likelihood estimators of $\{\lambda_k\}$ under model $\mathcal{M}_d$ are $\hat{\lambda}_1, \ldots, \hat{\lambda}_d$ for the leading d eigenvalues and $\sum_{k=d+1}^{m} \hat{\lambda}_k/(m-d)$ for the common $m - d$ eigenvalues. Therefore, the estimated conditional likelihood comes down to

$$\hat{p}(\mathbf{x}_1, \ldots, \mathbf{x}_n | \mathcal{M}_d) = (2\pi)^{-\frac{nm}{2}} \left\{ \left(\prod_{k=1}^{d} \hat{\lambda}_k \right) \left(\frac{\sum_{k=d+1}^{m} \hat{\lambda}_k}{m-d} \right)^{m-d} \right\}^{-\frac{n}{2}} \exp(-\frac{nm}{2}).$$

Since

$$\begin{aligned} \underset{0 \le d \le m-1}{\operatorname{argmax}} \, \hat{p}(\mathcal{M}_d | \mathbf{x}_1, \ldots, \mathbf{x}_n) &= \underset{0 \le d \le m-1}{\operatorname{argmax}} \, \frac{\hat{p}(\mathcal{M}_d | \mathbf{x}_1, \ldots, \mathbf{x}_n)}{\hat{p}(\mathcal{M}_{m-1} | \mathbf{x}_1, \ldots, \mathbf{x}_n)} \\ &= \underset{0 \le d \le m-1}{\operatorname{argmax}} \, \frac{\hat{p}(\mathbf{x}_1, \ldots, \mathbf{x}_n | \mathcal{M}_d) p(d)}{\hat{p}(\mathbf{x}_1, \ldots, \mathbf{x}_n | \mathcal{M}_{m-1}) p(m-1)}, \end{aligned}$$

after simple algebra we obtain

$$\hat{d} = \underset{0 \le d \le m-1}{\operatorname{argmax}} \left\{ \left(\frac{\hat{G}_d}{\hat{A}_d} \right)^{m-d} e^{\theta(m-1-d)} \right\}^{\frac{n}{2}}, \tag{2}$$

where $\hat{G}_d$ and $\hat{A}_d$ are the geometric and the arithmetic means of $\hat{\lambda}_{d+1}, \ldots, \hat{\lambda}_m$, respectively. We will use the notation $\hat{d} = \hat{d}(\theta)$ to stress the dependence of $\hat{d}$ on the subjective parameter θ. The properties of this Bayes rule are studied next.

3 Characterization and properties of the Bayes rule

For reasons both theoretical and computational, it is more practical to maximize the natural logarithm of the objective function (2) and get rid of the factor $n/2$. Then we have

$$\hat{d}(\theta) = \underset{0 \le d \le m-1}{\operatorname{argmax}} \{\hat{F}(d) + \theta(m-1-d)\},$$

where

$$\hat{F}(d) = (m-d) \log \left(\frac{\hat{G}_d}{\hat{A}_d} \right). \tag{3}$$

Note that $\hat{F}(d)$ is equivalent to the likelihood-ratio test statistic for the hypothesis $\lambda_{d+1} = \cdots = \lambda_m$, which is the basis of Bartlett's (1950) sequential method for selection of principal components. The Bayes rule, then, maximizes a penalized version of $\hat{F}(d)$; the term $\theta(m-1-d)$ penalizes high dimensions, and the penalty increases as θ increases.

The following properties of $\hat{F}(d)$ and $\hat{d}(\theta)$ are proved in the Appendix.

Proposition 1 *The function $\hat{F} : \{0, \ldots, m-1\} \longrightarrow (-\infty, 0]$ defined in (3) is:*

(a) *non-decreasing, and strictly increasing unless $\hat{\lambda}_q = \hat{\lambda}_{q+1} = \cdots = \hat{\lambda}_m$ for some q, in which case $\hat{F}(d) = 0$ for all $d \ge q$;*

(b) *invariant under translation, rotation and rescaling of $\mathbf{x}_1, \ldots, \mathbf{x}_n$, and therefore, so is $\hat{d}(\theta)$.*

Proposition 2 *The function $\hat{d} : [0, \infty) \longrightarrow \{0, \ldots, m-1\}$ is non-increasing.*

Proposition 3 *For each $d \in \{0, \ldots, m-1\}$, let*

$$\begin{aligned} \theta_{(d)} &= \max \left\{ \frac{\hat{F}(k) - \hat{F}(d)}{k-d} : k > d \right\}, \\ \theta^{(d)} &= \min \left\{ \frac{\hat{F}(d) - \hat{F}(k)}{d-k} : k < d \right\}, \end{aligned}$$

with $\theta_{(m-1)} = 0$ and $\theta^{(0)} = \infty$. Then:

(a) *$\hat{d}(\theta) = d$ only if $\theta \in [\theta_{(d)}, \theta^{(d)}]$. If $\theta_{(d)} > \theta^{(d)}$, $\hat{d}(\theta) \ne d$ for all θ.*

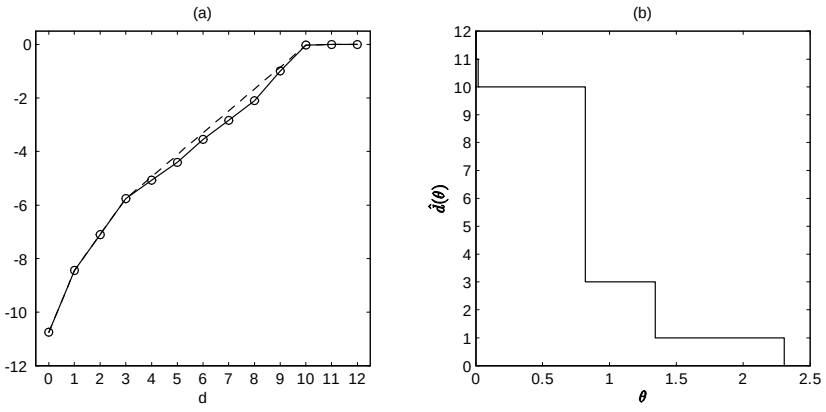

Figure 1: Pit Props Example. (a) Functions $\hat{F}(d)$ (solid line) and $\check{F}(d)$ (dashed line). (b) Function $\hat{d}(\theta)$.

(b) $\theta_{(d)} \leq \theta^{(d)}$ *if and only if* $\hat{F}(d) = \check{F}(d)$*, where* $\check{F}$ *is the concave majorant of* $\hat{F}$ *(i.e.* $\check{F}$ *is the smallest concave function such that* $\check{F} \geq \hat{F}$*).*

Proposition 3 characterizes the Bayes rule $\hat{d}(\theta)$ and provides an efficient way to compute it. It turns out that $\hat{d}(\theta)$ is a step function that starts at $\hat{d}(0) = m - 1$, jumps down at each value of $\theta_{(d)}$ which is less than $\theta^{(d)}$, and is zero for all $\theta \geq \theta_{(0)}$. Note that if $\theta_{(d)} > \theta^{(d)}$, model $\mathcal{M}_d$ is not selected for any θ.

Fig. 1 illustrates how the method works, using the pitprops data (which will be analyzed in more detail in Section 4). The dataset is 13-dimensional, so the possible models range from $\mathcal{M}_0$ to $\mathcal{M}_{12}$. However, $\check{F}(d) = \hat{F}(d)$ only for $d \in \{0, 1, 3, 10, 11, 12\}$ (Fig. 1(a)), so these are actually the only models with chances of being selected. This is more easy to visualize in the $\hat{d}$-plot (Fig. 1(b)). Since $\theta_{(d)} = \check{F}(d+1) - \check{F}(d)$ and $\theta^{(d)} = \check{F}(d) - \check{F}(d-1)$, any $\theta \in [\theta_{(d)}, \theta^{(d)}]$ is a supra-gradient of the concave function $\check{F}$ at d, so the step width $\theta^{(d)} - \theta_{(d)}$ measures, in a certain way, the magnitude of the speed change of $\check{F}$ at d. The precise dependence of $\theta_{(d)}$ and $\theta^{(d)}$ on the estimated eigenvalues is complicated, but it is clear that the bigger the difference between $\hat{\lambda}_d$ and $\hat{\lambda}_{d+1}, \ldots, \hat{\lambda}_m$, the larger $\theta^{(d)} - \theta_{(d)}$ will be.

A strict Bayesian approach would start with a completely subjective choice of

θ and simply select model $\mathcal{M}_{\hat{d}(\theta)}$. However, we have a different approach in mind: we suggest to plot the step function $\hat{d}(\theta)$ for $\theta \in [0, \theta_{(0)}]$ and select the highest dimension d for which the step is significantly large (in Fig. 1(b) this occurs at $d = 10$.) A large step length $\theta^{(d)} - \theta_{(d)}$ means that d is optimal under a wide range of prior model probabilities, even under relatively large values of θ that heavily penalize high dimensions, indicating that d is strongly suggested by the data irrespective (to a large extent) of the choice of prior.

Of course, the problem now is to decide whether a step length $\theta^{(d)} - \theta_{(d)}$ is "significantly large" or not. Our simulations (Section 5) show some scenarios where the longest step corresponds to the true dimension, but there are other situations where step length increases as d decreases, making model choice less clear-cut. This may sound similar to those situations where there is no recognizable "elbow" in the scree plot, but in fact we have found that the right model is usually easier to spot in the $\hat{d}$-plot than in the scree plot. This is partly due to the fact that $\hat{d}(\theta)$ skips some models entirely, and partly due to the next asymptotic result.

Proposition 4 *If $\mathcal{M}_q$ is the true model, with $q < m - 1$, then $P\{\hat{d}(\theta) > q\} \underset{n\to\infty}{\longrightarrow} 0$ for any $\theta > 0$.*

This property shows that, asymptotically speaking, the Bayes rule never overestimates the number of significant components. For finite samples, the consequence of this is that $\hat{d}(\theta)$ decreases very steeply as θ increases. Therefore, large dimensions will be chosen only for very small values of θ, that is, for prior probabilities that practically do not penalize high-dimensional models. In practice, we do want to penalize high dimensional models, so larger values of θ (which yield smaller values of $\hat{d}(\theta)$) will be selected.

The rest of the article illustrates the behavior of the $\hat{d}$-plot with real and simulated data.

4 Examples

Wooden Pitprops

A study was carried out to determine the strength of wooden props used in mining (Jeffers, 1967; see also Seber, 1984, p.190). Thirteen different variables were measured on each of 180 pitprops. The aim of the study was to analyze the relationship

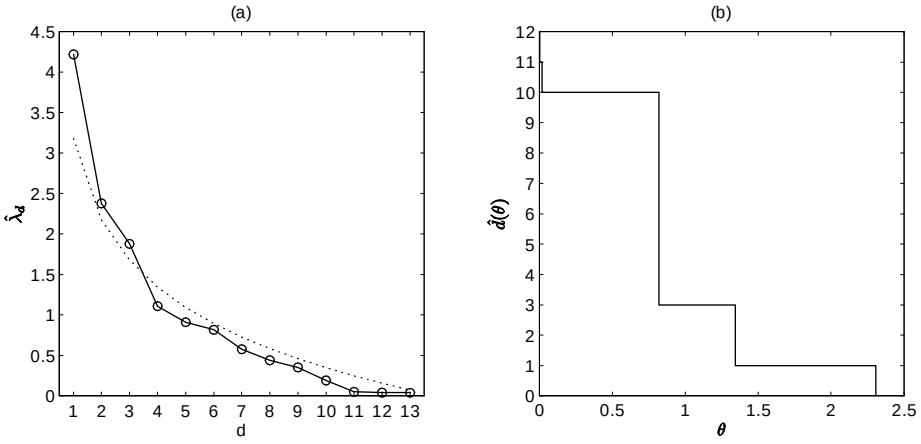

Figure 2: Pitprops Example. (a) Scree plot (solid line) and broken-stick plot (dotted line). (b) $\hat{d}$-plot.

between these variables and maximum compressive strength, but an initial dimension reduction was attempted. The variables had different scales, so the PCA was carried out on the correlation matrix instead of the covariance matrix. The scree plot (Fig. 2(a)) shows a moderate elbow at $d = 4$, and the first three components are above the broken-stick line, so a three-component model seems reasonable. The rule of Kaiser (1960) would keep the first 4 components, which are greater than 1. However, since the first 3 and 4 components represent only 65% and 74% of the variability, respectively, Jeffers (1967) and Seber (1984) settle for 6 components, which account for 87% of the variability. This choice is completely arbitrary, though. In fact, the $\hat{d}$-plot (Fig. 2(b)) shows that a six-component model is not selected for any θ; neither is a 4 component model. The $\hat{d}$-plot indicates that either 1, 3 or 10 components should be retained. Although the ten-component model has a wider step than the three-component model, in the interest of dimension reduction we would keep 3 components in this example. If one is unwilling to sacrifice potentially useful information, a ten-component model is also be justifiable, but intermediate choices between 3 and 10 are not.

Blood Chemistry

This data consists of measurements on 8 blood chemistry variables for 72 patients and was used by Jolliffe (2002, p.133) to illustrate the problems involved in PC selection. Since the scales of the original variables are different, the PCA is done on the correlation matrix. The scree plot (Fig. 3(a)) does not show any sharp elbows. Two moderate elbows occur at $d = 2$ and $d = 4$, and the eigenvalues $\hat{\lambda}_1$ and $\hat{\lambda}_3$ exceed the broken-stick line, so a one-component model or a three-component model are plausible choices. Since the first eigenvalue accounts for only 35% of the variance and the first three account for 70%, the latter choice seems preferable. Kaiser's rule would also keep the first three eigenvalues, since the others are less than 1. The $\hat{d}$-plot (Fig. 3(b)) also suggests a three-component model.

Handwritten Digits

Automatic machine recognition of handwritten ZIP codes is important for efficient sorting of mail envelopes. Hastie *et al.* (2001, p.488) analyze a sample of 658 digitalized 16×16 grayscale images of handwritten "threes". The images were

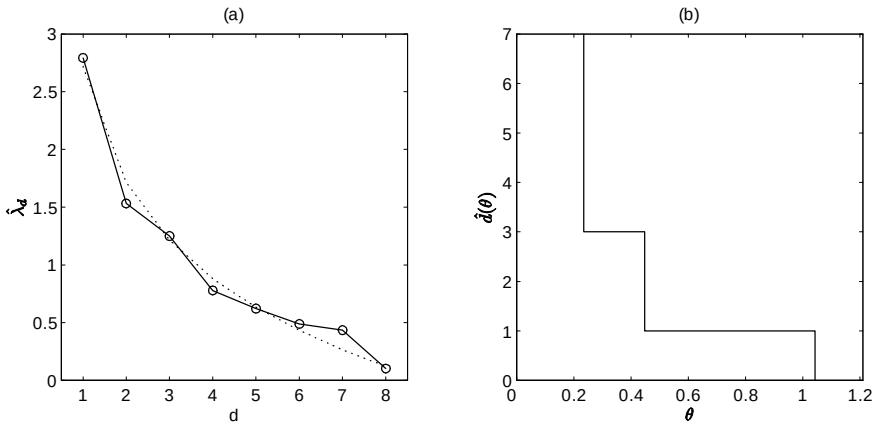

Figure 3: Blood Chemistry Example. (a) Scree plot (solid line) and broken-stick plot (dotted line). (b) $\hat{d}$-plot.

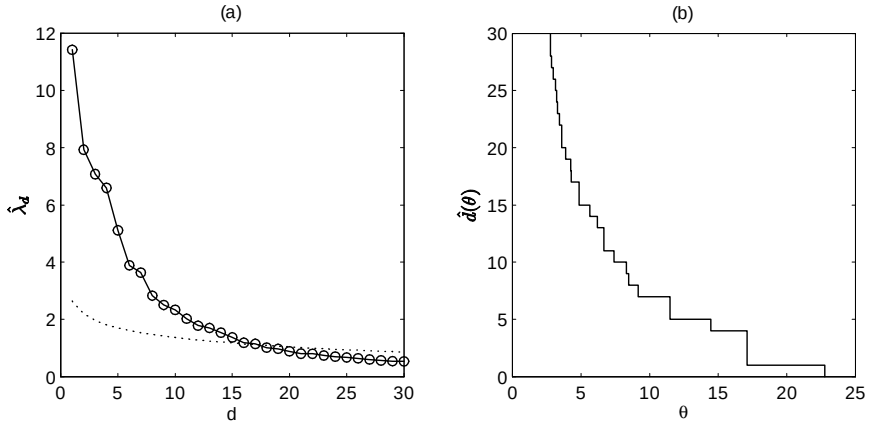

Figure 4: Handwritten Digits Example. (a) Scree-plot of eigenvalues of the covariance matrix. (b) $\hat{d}$-plot.

normalized to have approximately the same size and orientation. Each image can be represented as a 256-dimensional vector of pixels. Since most of the pixels are blank, substantial data compression can be attained with PCA. Fig. 4(a) shows the scree graph and the broken-stick line for the leading 30 components, which account for 82% of the total variance. An "elbow" is discernible at $d = 6$, but the value of $\hat{\lambda}_7$ is closer to $\hat{\lambda}_6$ than to $\hat{\lambda}_8$, so 7 components may be a reasonable choice. The broken-stick method and Kaiser's rule, however, indicate that many more components should be kept: 17 and 34, respectively. The $\hat{d}$-plot (Fig. 4(b)) is more clear-cut than the scree plot, since the first significant step clearly occurs at $d = 7$. The first seven components account for only 51% of the variability, but this is typical of high-dimensional problems where most of the tail components correspond to high-frequency noise.

5 Simulations

5.1 $\hat{d}$-plot performance

The examples in Section 4 show that the behavior of the $\hat{d}$-plot is reasonable, but for real datasets we never really know what the "true" answer is. Therefore, we ran a few simulations to assess our method in situations where the right answer is known. First, we simulated Normal datasets of dimension $m = 20$ with $q = 5$ distinct eigenvalues. Specifically, we took $\lambda_k = 1/k$ for $k = 1, \ldots, 5$ and $\lambda_k = \lambda$ for $k = 6, \ldots, 20$. Three different values of λ were considered: $1/10$, $1/20$ and $1/40$; thus, the ratios λ_6/λ_5 were $1/2$, $1/4$ and $1/8$, respectively. We also considered three sample sizes: 100, 300 and 900. This represents a total of 9 simulated models, ranging from a "difficult" case (large ratio λ_6/λ_5 and small sample size) to an "easy" case (small ratio λ_6/λ_5 and large sample size). Each model was replicated 1000 times.

Since the proposed method is of graphical nature, it is difficult to give a summary description of the simulation results. Fig. 5 shows boxplots of step lengths for each dimension (as expected from Proposition 4, the step lengths are negligible for dimensions larger than 7, so we only show the first 10.) Our method will be judged as successful if the first significantly long step (as θ increases) occurs at $d = 5$; this does not necessarily mean that the step at $d = 5$ has to be the longest, it only

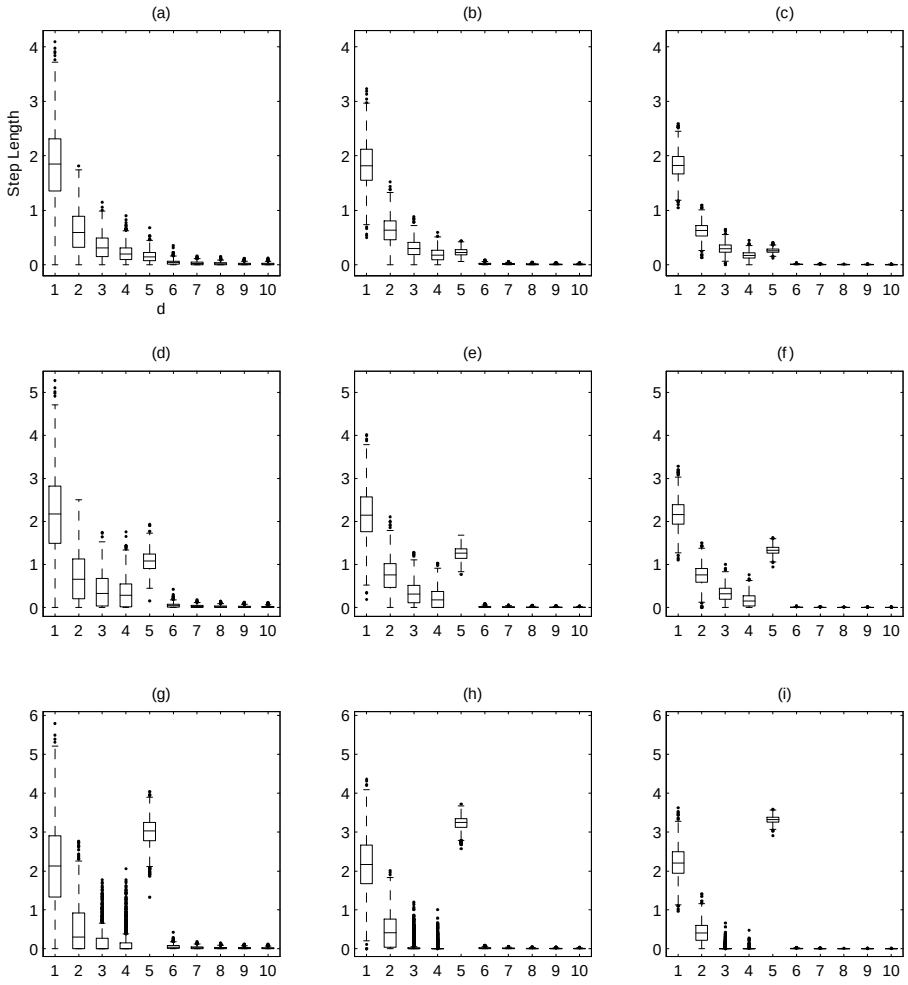

Figure 5: Simulation Results for Normal Data. Step lengths of simulated $\hat{d}$-plots for models with 5 significant components and varying ratios λ_6/λ_5 and sample sizes n. (a,b,c) $\lambda_6/\lambda_5 = 1/2$, (d,e,f) $\lambda_6/\lambda_5 = 1/4$, (g,h,i) $\lambda_6/\lambda_5 = 1/8$; (a,d,g) $n = 100$, (b,e,h) $n = 300$, (c,f,i) $n = 900$.

means that it should be recognizably longer that the steps at higher dimensions, and in particular than the step at $d = 6$. In Fig. 5 we see that this is so, even for the worst-case scenario (Fig. 5(a)). In general, we see that the relative step length at $d = 5$ depends on the ratio λ_6/λ_5, as expected: the smaller λ_6 is with respect to λ_5, the longer the step at $d = 5$ is. Overall, for each ratio λ_6/λ_5 we observe that median step lengths remain stable as n increases, while the variability decreases. This means that increasing the sample size makes detection more precise, not because the step length at the true model gets larger, but because the relative step lengths tend to stabilize.

The last claim is better understood with the help of Fig. 6(a–c). These boxplots are log-ratios of consecutive step lengths for the three models with $\lambda_6/\lambda_5 = 1/2$ (we added 10^{-10} to each step length to avoid logarithm-of-zero situations). We expect the step at $d = 5$ to be wider than the step at $d = 6$, thus the log-ratios at $d = 5$ should be positive. To a large extent this is what happens, although for $n = 100$ (Fig. 6(a)) the $\hat{d}$-plots are very variable and often the steps at $d = 7$ or $d = 8$ are longer than the steps at $d = 5$. As n increases the behavior of the $\hat{d}$-plot stabilizes, and for $n = 300$ (Fig. 6(b)) we already see that the step ratio at $d = 5$ is significantly larger than at preceding and subsequent dimensions, so the user would almost always select the right model.

The preceding results correspond to normally distributed data. To assess the performance of the method under skewed distributions, we generated random vectors with standardized independent marginal Gamma distributions. Specifically, $\mathbf{X}$ is such that $X_k = \sqrt{\lambda_k}(Y_k - 10)/\sqrt{50}$, with $Y_k \sim \Gamma(2, 5)$. The sequences of eigenvalues, the dimension m and the sample size n were the same as above. Again, each scenario was replicated 1000 times. The results are shown in Fig. 7. Although the step lengths are now more variable than for normally distributed data, the relative behavior is the same: the step length at the true dimension $d = 5$ is significantly larger than the step lengths at higher dimensions (except for a small overlap in the worst-case scenario of Fig. 7(a)), and this becomes more evident as n increases or as the ratio λ_6/λ_5 decreases. In conclusion, we can say that the $\hat{d}$-plot performs equally well for this highly skewed distribution as for the Normal distribution.

A different kind of deviation from normality that is worth studying is provided by heavy-tailed distributions. We generated random vectors with standardized independent t distributions with 3 degrees of freedom; that is, $X_k = \sqrt{\lambda_k}Y_k/\sqrt{3}$, with

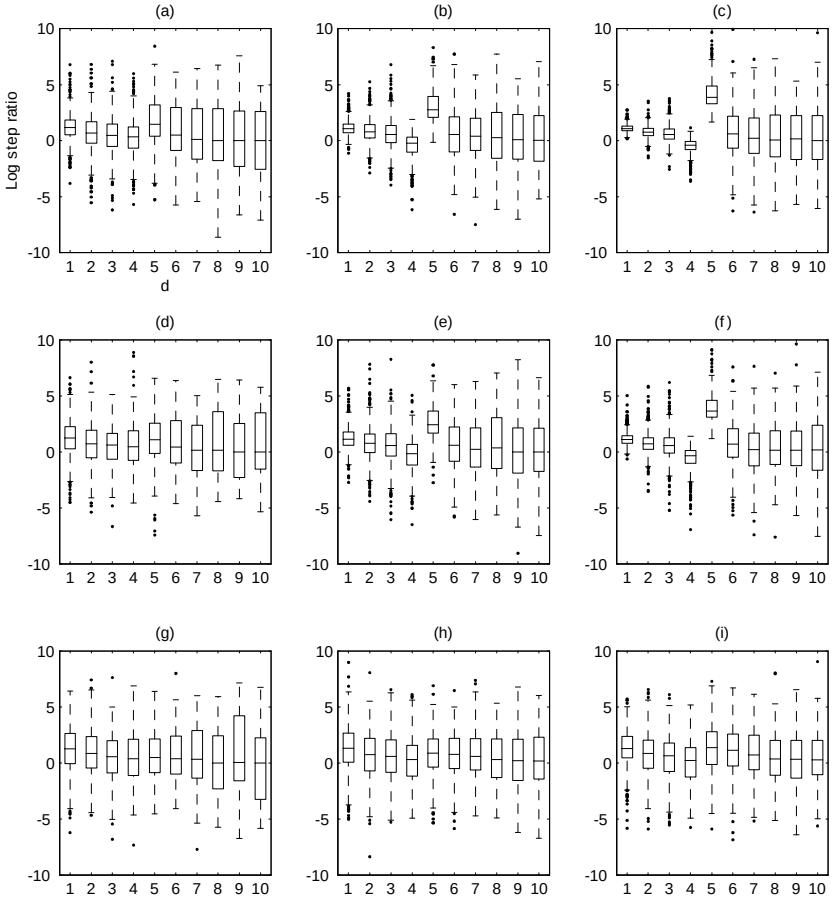

Figure 6: Simulation Results. Log-ratios of consecutive $\hat{d}$-plot step lengths for models with 5 dominant components and eigenvalue ratio $\lambda_6/\lambda_5 = 1/2$. Sample sizes are $n = 100$ [(a,d,g)], $n = 300$ [(b,e,h)] and $n = 900$ [(c,f,i)]. Data distribution is Normal [(a–c)], marginal Gamma [(d–f)] and marginal t [(g–i)].

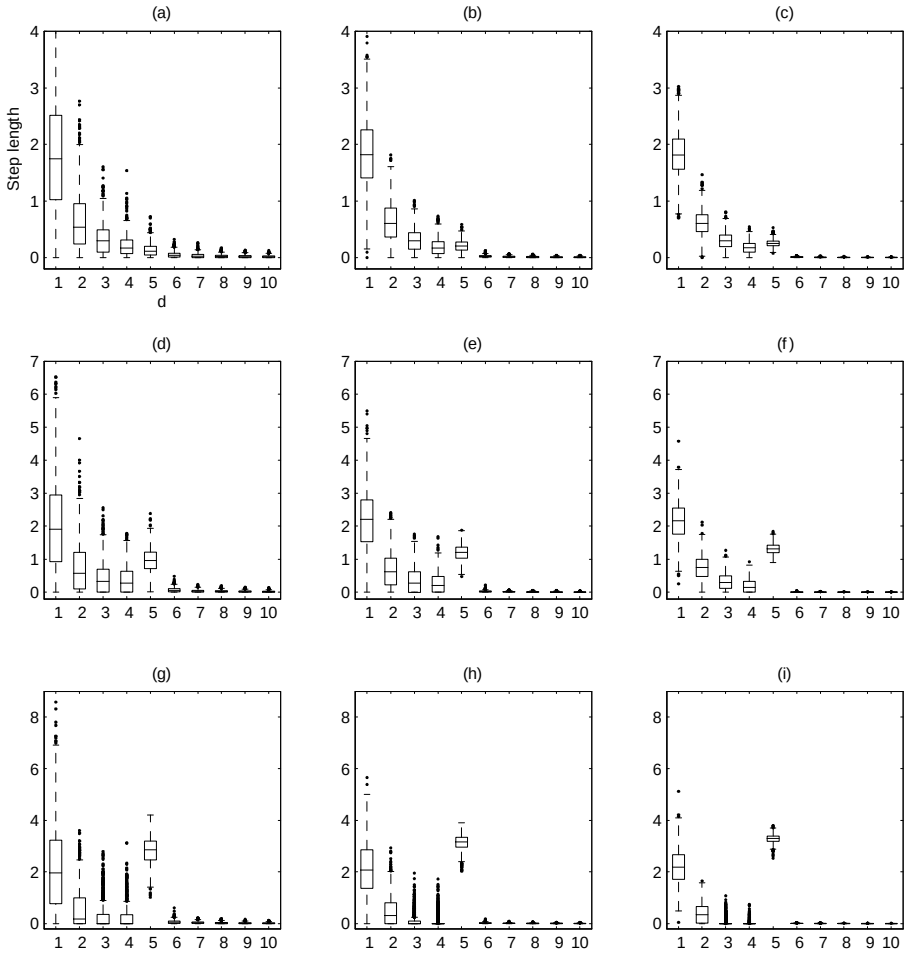

Figure 7: Simulation Results for Marginal Gamma Data. Step lengths of simulated $\hat{d}$-plots for models with 5 significant components and varying ratios λ_6/λ_5 and sample sizes n. (a,b,c) $\lambda_6/\lambda_5 = 1/2$, (d,e,f) $\lambda_6/\lambda_5 = 1/4$, (g,h,i) $\lambda_6/\lambda_5 = 1/8$; (a,d,g) $n = 100$, (b,e,h) $n = 300$, (c,f,i) $n = 900$.

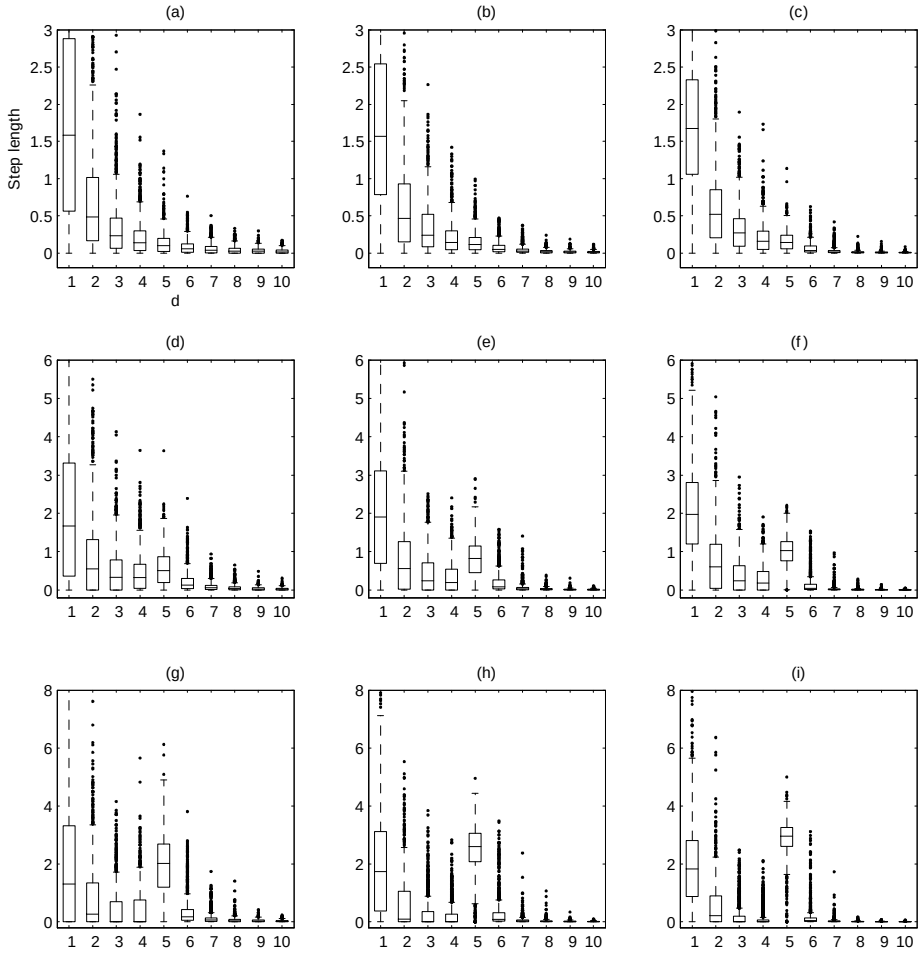

Figure 8: Simulation Results for Marginal t Data. Step lengths of simulated $\hat{d}$-plots for models with 5 significant components and varying ratios λ_6/λ_5 and sample sizes n. (a,b,c) $\lambda_6/\lambda_5 = 1/2$, (d,e,f) $\lambda_6/\lambda_5 = 1/4$, (g,h,i) $\lambda_6/\lambda_5 = 1/8$; (a,d,g) $n = 100$, (b,e,h) $n = 300$, (c,f,i) $n = 900$.

$Y_k \sim t_3$. Note that a t_3 distribution has finite variance but not finite moments of higher order, so the sample covariance matrix of $\mathbf{X}$ and its eigenvalues are consistent but not asymptotically Normal (Bilodeau and Brenner, 1999, ch. 6). The sequences of eigenvalues and the parameters m and n were the same as before, and 1000 replications were run for each model. The results are shown in Fig. 8 (some boxplots were trimmed for better visualization). Clearly, the $\hat{d}$-plot will have trouble selecting the right dimension when the ratio λ_6/λ_5 is small, due to the high variability of the step lengths. For $\lambda_6/\lambda_5 = 1/2$, only when $n = 900$ (Fig. 8(c)) is the step at $d = 5$ significantly longer than the step at $d = 6$ with little overlap of the boxplots. From the point of view of relative step lengths the situation is not better (Fig. 6(d–f)). The case $\lambda_6/\lambda_5 = 1/4$ is more favorable to our method, although for $n = 100$ (Fig. 8(d)) there is still a small overlap between the distributions of step lengths at $d = 5$ and $d = 6$. The rest of the cases (Figs. 8(e–f)) are more clear-cut but the distribution of the step length at $d = 5$ still shows larger variability than for the Normal or marginal Gamma models. So it is clear that heavy-tailed distributions, by increasing the variability of the sample eigenvalues, are more problematic for the $\hat{d}$-plot than skewed distributions.

5.2 Comparison with the scree plot

Comparing the performance of the $\hat{d}$-plot and the scree plot by simulation is complicated because neither method provides an objective way to select the optimal dimension d. For example, although the choice $d = 7$ was obvious in Fig. 4(b), the step length at $d = 7$ is neither the longest in the absolute sense nor the one with highest ratio between consecutive steps (some of the higher dimensions show larger ratios between consecutive steps, even though the step lengths are insignificantly short; this behavior of the $\hat{d}$-plot is also apparent in the simulation results of Fig. 6). The scree plot has a similar drawback: a significant "elbow" that is obvious for the human eye is not necessarily the only one in the plot or the largest one in a quantifiable way. This precludes a realistic comparative Monte Carlo study of the $\hat{d}$-plot and the scree plot except for rather trivial situations.

Nevertheless, some insight on the behavior of the scree plot can be gained by plotting the magnitude of the "elbows" at different dimensions. An "elbow" occurs at d when the difference between $\hat{\lambda}_{d-1}$ and $\hat{\lambda}_d$ is much bigger than the difference between $\hat{\lambda}_d$ and $\hat{\lambda}_{d+1}$, indicating that a $d-1$ dimensional model may be appropriate.

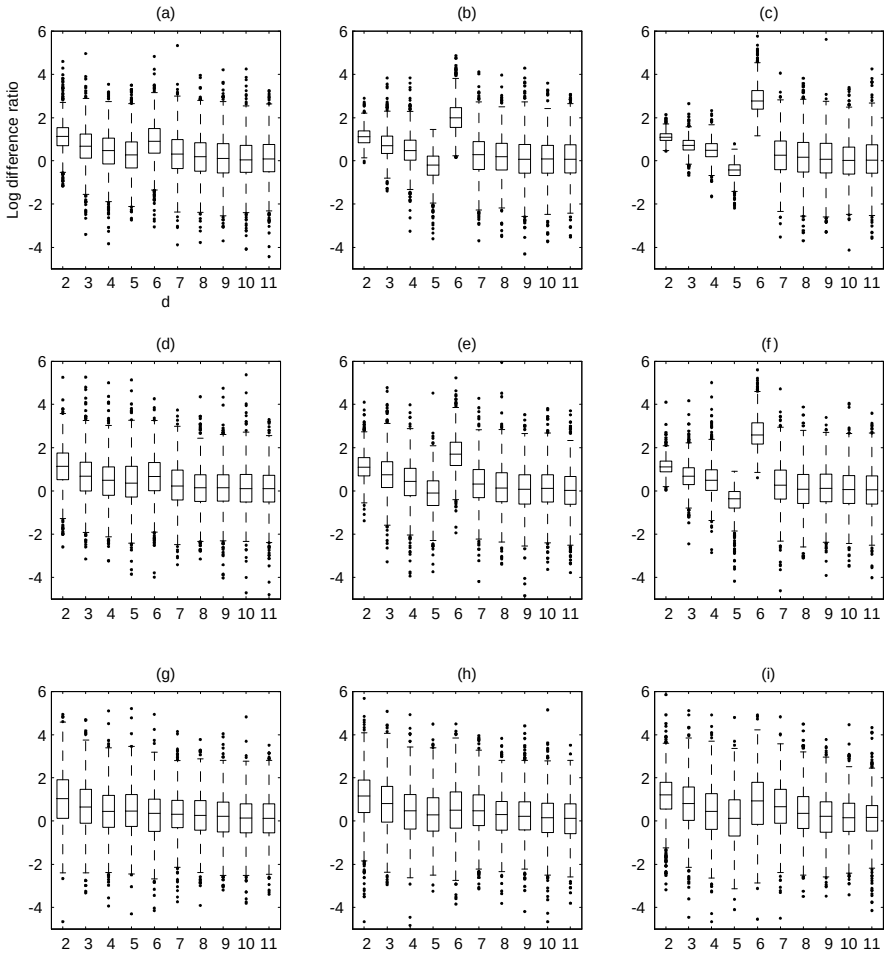

Figure 9: Simulation Results. Logarithm of difference ratios of eigenvalues for models with 5 dominant components and ratio $\lambda_6/\lambda_5 = 1/2$. Sample sizes are $n = 100$ [(a,d,g)], $n = 300$ [(b,e,h)] and $n = 900$ [(c,f,i)]. Data distribution is Normal [(a–c)], marginal Gamma [(d–f)] and marginal t [(g–i)].

So we simulated the distributions of $\log\{(\hat{\lambda}_{d-1} - \hat{\lambda}_d)/(\hat{\lambda}_d - \hat{\lambda}_{d+1})\}$ for the same sampling situations as in Section 5.1. To save space, we show only the results for $\lambda_6/\lambda_5 = 1/2$ in Fig. 9. Under the Normal and marginal Gamma distributions, the scree plot will show easily detectable "elbows" for $n = 300$ and $n = 900$, but for $n = 100$ the situation is less clear cut. For the marginal t_3 distribution we see that not even for $n = 900$ are there significant "elbows" at $d = 6$; therefore, the scree plot will perform poorly when the ratio λ_6/λ_5 is small and the distribution is heavy tailed. Overall, the patterns in Fig. 9 mimic those of Fig. 6, so we can conclude that the behavior of the scree plot and the $\hat{d}$-plot are comparable.

6 Discussion

By framing principal-component selection as a Bayesian model selection problem, we were able to derive a new graphical method that can be used as an alternative or as a complement to the widely used scree plot. The method is still subjective, but has certain advantages. From a theoretical point of view, there is no ambiguity as to what the $\hat{d}$-plot indicates: for a given prior parameter θ, $\hat{d}(\theta)$ is the dimension with highest posterior probability. In contrast, from a scree plot one gets an "elbow" (if there is one) that cannot be translated into any formal statistical concept. From a more practical point of view, we think the $\hat{d}$-plot is easier to read than the scree plot; the user has to determine whether a step length is significantly larger than others, which is easier to do than finding an "elbow" in the scree plot. Moreover, we have seen that the method is consistent (Proposition 4) and gives the right answer for simulated data. Nevertheless, given our comparative simulation results and the fact that a plot of the individual eigenvalues is always instructive, we suggest that our method be used as a complement rather than a substitute of the scree plot.

Appendix

Proof of Proposition 1. Since

$$\hat{F}(d) - \hat{F}(d-1) = -\log\hat{\lambda}_d - (m-d)\log\hat{A}_d + (m-d+1)\log\hat{A}_{d-1},$$

and the logarithm is a strictly concave function,

$$\log \hat{A}_{d-1} \geq \frac{1}{m-d+1} \log \hat{\lambda}_d + \frac{m-d}{m-d+1} \log \hat{A}_d,$$

with equality holding if and only if $\hat{\lambda}_d = \hat{A}_d$, that is, if and only if $\hat{\lambda}_d = \hat{\lambda}_{d+1} = \cdots = \hat{\lambda}_m$. This proves part (a). Part (b) is a consequence of the translation and rotation invariance of the estimated eigenvalues, and the scale invariance of the ratios $\hat{G}_d/\hat{A}_d$. □

Proof of Proposition 2. Let $\theta_1 < \theta_2$, $d_1 = \hat{d}(\theta_1)$ and $d_2 = \hat{d}(\theta_2)$. Then

$$\begin{aligned} \hat{F}(d_2) + \theta_1(m-1-d_2) &\leq \hat{F}(d_1) + \theta_1(m-1-d_1), \\ \hat{F}(d_1) + \theta_2(m-1-d_1) &\leq \hat{F}(d_2) + \theta_2(m-1-d_2). \end{aligned}$$

This implies

$$\theta_2(d_2 - d_1) \leq \hat{F}(d_2) - \hat{F}(d_1) \leq \theta_1(d_2 - d_1),$$

which cannot happen if $\hat{F}(d_2) - \hat{F}(d_1)$ and $d_2 - d_1$ have the same sign. □

Proof of Proposition 3. (a) Given θ, let $d = \hat{d}(\theta)$. Then $\hat{F}(d) - \hat{F}(k) \geq \theta(d-k)$ for all k. Since $\hat{F}$ is non-decreasing (Proposition 1(a)),

$$\begin{aligned} \theta &\leq \frac{\hat{F}(d) - \hat{F}(k)}{d-k} \text{ for all } k < d, \\ \theta &\geq \frac{\hat{F}(k) - \hat{F}(d)}{k-d} \text{ for all } k > d. \end{aligned}$$

Then $\theta_{(d)} \leq \theta \leq \theta^{(d)}$, and $\hat{d}(\theta)$ cannot be equal to d if $\theta^{(d)} < \theta_{(d)}$.

(b) By definition,

$$\check{F}(d) = \max \left\{ \sum_{k=1}^{m-1} \alpha_k \hat{F}(k) : \alpha_k \geq 0, \sum_{k=1}^{m-1} \alpha_k k = d, \sum_{k=1}^{m-1} \alpha_k = 1 \right\}.$$

If $\theta^{(d)} < \theta_{(d)}$, there are $k_1 < d$ and $k_2 > d$ such that

$$\frac{\hat{F}(d) - \hat{F}(k_1)}{d - k_1} < \frac{\hat{F}(k_2) - \hat{F}(d)}{k_2 - d}.$$

This implies that

$$\hat{F}(d) < \frac{d-k_1}{k_2-k_1}\hat{F}(k_2) + \frac{k_2-d}{k_2-k_1}\hat{F}(k_1),$$

so $\hat{F}(d) < \check{F}(d)$. Conversely, if $\hat{F}(d) < \check{F}(d)$, there exist non-negative $\alpha_1, \ldots, \alpha_{m-1}$ such that $\sum_{k=1}^{m-1} \alpha_k k = d$, $\sum_{k=1}^{m-1} \alpha_k = 1$ and

$$\hat{F}(d) < \sum_{k=1}^{m-1} \alpha_k \hat{F}(k).$$

This implies that

$$\sum_{k=1}^{d-1} \alpha_k \{\hat{F}(d) - \hat{F}(k)\} < \sum_{k=d+1}^{m-1} \alpha_k \{\hat{F}(k) - \hat{F}(d)\}$$

and then

$$\theta^{(d)} \sum_{k=1}^{d-1} \alpha_k (d-k) < \theta_{(d)} \sum_{k=d+1}^{m-1} \alpha_k (k-d).$$

Since $\sum_{k=1}^{d-1} \alpha_k (d-k) = \sum_{k=d+1}^{m-1} \alpha_k (k-d)$, it follows that $\theta^{(d)} < \theta_{(d)}$. □

Proof of Proposition 4. Since $\{\hat{\lambda}_k\}$ are consistent estimators of $\{\lambda_k\}$ for distributions with finite second moments (see e.g. Bilodeau and Brenner, 1999, p.135),

$$\operatorname*{plim}_{n\to\infty} \hat{F}(d) := F(d) = \begin{cases} (m-d)\log(G_d/A_d), & d < q, \\ 0, & d \geq q, \end{cases}$$

where G_d and A_d are the geometric and arithmetic means of $\lambda_{d+1}, \ldots, \lambda_m$, respectively. If $\theta > 0$ and $\hat{d}(\theta) = d > q$, $\hat{F}(d) - \hat{F}(q) > \theta(d-q)$, so

$$\limsup_{n\to\infty} P\{\hat{d}(\theta) = d\} \leq \lim_{n\to\infty} P\{\hat{F}(d) > 0\}.$$

Since $F(d) < 0$ for $d < q$, it follows that $\lim_{n\to\infty} P\{\hat{d}(\theta) = d\}$ exists and is equal to 0 for any $d > q$ when $\theta > 0$. □

References

Anderson, T.W. (1984). Estimating linear relationships. *Ann. Math. Statist.*, 34, 122–148.

Bartlett, M.S. (1950). Test of significance for factor analysis. *Br. J. Psychol. Statis. Sec.*, 3, 77–85.

Bilodeau, M. and Brenner, D. (1999). *Theory of Multivariate Statistics*. New York: Springer.

Cattell, R.B. (1966). The scree test for the number of factors. *J. Mult. Behav. Res.*, 1, 245–276.

Ferré, L. (1995). Selection of components in principal component analysis: A comparison of methods. *Comput. Statist. Data Anal.*, 19, 669–682.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference and Prediction.* New York: Springer.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Ed. Psychol.*, 24, 417–441 and 498–520.

Jeffers, J.N.R. (1967). Two case studies in the application of principal component analysis. *Appl. Statist.*, 16, 225–236.

Jolliffe, I.T. (2002). *Principal Component Analysis* (2nd ed.). New York: Springer.

Kaiser, H.F. (1960). The application of electronic computers to factor analysis. *Educ. Psychol. Meas.,* 20, 141–151.

Krzanowski, W.J. (1987). Cross-validation in principal component analysis. *Biometrics*, 43, 575–584.

Legendre, L. and Legendre, P. (1983). *Numerical Ecology*. Amsterdam: Elsevier.

Seber, G.A.F. (1984) *Multivariate Observations*. New York: Wiley.

Wold, S. (1978). Cross validatory estimation of the number of components in factor and principal component models. *Technometrics*, 20, 397–405.