Dimension Selection for Feature Selection and Dimension Reduction with Principal and Independent Component Analysis

Inge Koch  
inge@maths.unsw.edu.au  
Department of Statistics, School of Mathematics, University of New South Wales, Sydney, NSW 2052 Australia

Kanta Naito  
naito@riko.shimane-u.ac.jp  
Department of Mathematics, Faculty of Science and Engineering, Shimane University, Matsue 690-8504 Japan

This article is concerned with the problem of selecting the best or most informative dimension for dimension reduction and feature extraction in high-dimensional data. The dimension of the data is reduced by principal component analysis; subsequent application of independent component analysis to the principal component scores determines the most nongaussian directions in the lower-dimensional space. A criterion for choosing the optimal dimension based on bias-adjusted skewness and kurtosis is proposed. This new dimension selector is applied to real data sets and compared to existing methods. Simulation studies for a range of densities show that the proposed method performs well and is more appropriate for nongaussian data than existing methods.

1 Introduction

This article proposes a dimension selector for the most informative lower-dimensional subspace in feature extraction and dimension-reduction problems. Our methodology is based on separating the informative part of the data from noise in high-dimensional problems: it reduces high-dimensional data and represents them on a smaller number of new directions that contain the interesting information inherent in the original data. Our algorithm combines principal component analysis (PCA) and independent component analysis (ICA) and makes use of objective- and data-driven criteria in the selection of the best dimension.

A combined PC-IC algorithm has been used in a number of papers, including Prasad, Arcot, and Koch (2005), Gilmour and Koch (2004), Koch, Marron, and Chen (2005) and is also implicitly referred to in many ICA algorithms (see Hyvärinen, Karhunen, & Oja, 2001). Our approach is motivated
by Prasad et al. (2005), who proposed a feature selection and extraction method for efficient classification by choosing the best subset of features using PC-IC. Their method resulted in a very efficient algorithm that is also highly competitive in terms of classification accuracy. Their research was driven by finding a smaller number of predictors or features for efficient classification, while our research focuses on the more general problem of dimension reduction in high-dimensional data. Features selection can be regarded as a special case of dimension reduction: representing or approximating high-dimensional data in terms of a smaller number of new directions that span a lower-dimensional space.

PCA alone achieves dimension reduction of high-dimensional data. PCA is well understood (see Anderson, 1985, Mardia, Kent, & Bibby, 1979), easily accessible, and implemented in many software packages. However, the lower-dimensional directions chosen by PCA do not always preserve or represent the structure inherent in the original data. Finding interesting and informative structure in data is one of the two main aims of ICA. ICA can be applied to the original or to PC-reduced lower-dimensional data. In either case, it aims to find new directions, and the number of new directions is at most that of the data that form the input to ICA. A mixture of PCA and ICA seems to be reasonable since considerable computational effort is required to implement ICA, especially for high-dimensional data, and hence dimension reduction without loss of information or with minimal loss is essential as a step preceding ICA. This is the reason we adopt PCA as a first step prior to ICA. Different approaches to ICA and software packages are available (see Cardoso, 1999; Bach & Jordan, 2002; Amari, 2002). We follow the approach of Hyvärinen et al. (2001) and use their software, FastICA2-4 (Hyvärinen, & Oja, 2005). More specifically, we use their skewness and kurtosis approximation to negentropy, since these provide fast computation as well as easy interpretation.

Dimension-reduction techniques generally provide a constructive approach to representing data in terms of new directions in lower-dimensional spaces, with the number of new directions varying from one to the (possibly very high) dimension of the original data. Traditionally in PCA, one is looking for an elbow, or one uses the 95% (resp. 99.9%) rule; that is, one uses the dimension whose contribution to total variance is at least 95% (resp. 99.9%). The latter is the route Prasad et al. (2005) have taken, and this choice of dimension clearly lacks objectivity.

A number of (gaussian) model-based approaches to dimension reduction have been proposed in the literature over the past few years. We review these approaches in section 5. If the model fits the data, then these methods may be exactly what is needed, and good dimension-reduction solutions are expected to be obtained with any of them. Furthermore, in specific applications, prior or additional information can help determine the dimension. A practical adjustment of this kind has been discussed in Beckmann and Smith (2004). However, to our knowledge, there do not appear to exist
nonparametric or model-free objective approaches that will determine
the right number of new directions or the best dimension for a lower-
dimensional space. We shall see in section 6 in particular that the gaussian
model–based approaches are not adequate in many cases and that better
results can be achieved with our nonparametric method.

The main purpose of this article is to propose a generic nonparametric
approach to dimension selection that is based on theory and objective crite-
ria without requiring a model or knowledge of the distribution of the data
and is data driven. Of course, if additional knowledge about the distribution
of the data exists, integrating such knowledge is expected to improve the
performance, but we will not be concerned with special cases here. Closely
related to the selection of this dimension is the choice of the direction vectors
in this lower-dimensional subspace. In an ICA-based framework, the natural
candidates are those IC directions based on the same objective function
as the selector—in our case, skewness or kurtosis. Finding those directions
is the aim of any good ICA algorithm, so we will not be concerned with
this issue here, but will focus instead on the choice of the dimension for the
lower-dimensional subspace.

We use the following notations throughout this article. Let $X$ denote a
d-dimensional random vector with mean $\mu$ and covariance matrix $\Sigma$. For
$k = 3, 4$, define the moments

$$\beta_k(X) = \max_{\alpha \in S^{d-1}} B_k(\alpha|X),$$

where

$$B_k(\alpha|X) = \left| E \left[ \frac{\alpha^T (X - \mu)}{\sqrt{\alpha^T \Sigma \alpha}} \right]^k \right| - 3 \delta_{k, 4}.$$ 

$S^{d-1}$ is the unit sphere in $\mathbb{R}^d$, and $\delta_{k, 4}$ is the Kronecker delta. The moments
$\beta_3(X)$ and $\beta_4(X)$ are the multivariate skewness and kurtosis of $X$, re-
spectively, in the sense of Malkovich and Afifi (1973). Note that multivariate
skewness was originally defined as the square of $\beta_3(X)$ above, but many
authors since then have used it in the form given here, as this is more
in keeping with the one-dimensional version. The multivariate skewness
and kurtosis are both defined such that they are zero for gaussian random
vectors.

For $N$ independent observations $X_1, \ldots, X_N$, the sample skewness and
sample kurtosis are defined as

$$\hat{\beta}_k(X_1, \ldots, X_N) = \max_{\alpha \in S^{d-1}} \hat{B}_k(\alpha|X_1, \ldots, X_N)$$
for $k = 3, 4$, where

\[
\hat{B}_k(\alpha|X_1, \ldots, X_N) = \left| \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{\alpha^T (X_i - \bar{X})}{\sqrt{\alpha^T S \alpha}} \right\}^k - 3\delta_{k,4} \right| .
\] (1.2)

Here, $\bar{X}$ denotes the sample mean and $S$ the sample covariance matrix. We will base our feature extraction and dimension reduction for high-dimensional data on these $\hat{B}_k$’s. It should be noted that $\beta_k(X)$ is not a function of $X$ but just a property of the distribution of $X$, while $\hat{B}_k(X_1, \ldots, X_N)$ is a random variable. Expressing $X$ and $(X_1, \ldots, X_N)$ in terms of $\beta_k(X)$ and $\hat{B}_k(X_1, \ldots, X_N)$, respectively, is necessary to explain our technique in the subsequent discussion. This type of statistic, based on the projections and summary statistics, is commonly used in the statistical science community. A typical and powerful tool based on projections is projection pursuit (see Huber, 1985; Jones & Sibson, 1987; Friedman, 1987), in which the nonlinear structure of high-dimensional data is explored through low-dimensional projections. The optimal low-dimensional directions are determined via a projection index that aims to detect nongaussianity. Hence, statistics for testing gaussianity could be based on the projection index, and typical instances are the $\hat{B}_k$’s above. The projection index proposed in Jones and Sibson (1987) is nothing other than a weighted quadratic combination of $\hat{B}_k$’s. A close relation between ICA and projection pursuit has been discussed by several authors (see, e.g., Cardoso, 1999).

This article is organized as follows. Section 2 describes our methodology and algorithm. Details are given in both the population version and the empirical (sample) version. A criterion for selecting the best or most informative dimension is proposed in section 3. The idea to construct well-known Akaike information criteria (AIC) is exploited to develop the proposed criterion. We argue that the gaussian distribution could be seen as the null structure for developing our criterion and present the necessary theoretical results. In section 4, we report the results of applying our selection method to real data sets. Section 5 discusses and appraises some model-based approaches to dimension reduction. The reports of Monte Carlo simulations and comparisons with the model-based approached discussed in section 5 are provided in section 6. All proofs of theoretical results are summarized in appendix A.

2 PC-IC Algorithm

Our algorithm for dimension reduction and feature extraction is described in this section. It combines PCA and ICA.
2.1 PCA and ICA. ICA can be used to extract features in multivariate structure (Hyvärinen et al., 2001). Feature extraction and testing for gaussianity (normality) with the skewness and kurtosis approximations in ICA were discussed in Prasad et al. (2005) and Koch et al. (2005). The relevant skeleton of the algorithm in those articles contains two steps: reduce and maximize:

**Reduce.** For $2 \leq p < d$, determine the $p$-dimensional subspace consisting of the first $p$ principal component directions obtained via PCA.

**Maximize.** Find the most nongaussian direction vector for the $p$-dimensional data using ICA.

The usual PCA is applied to high-dimensional data in the reduce step, and the reduced $p$-dimensional data are obtained as the $p$-dimensional principal component scores. ICA is applied to these $p$-dimensional principal component scores in the maximize step to find the most prominent structure. We shall refer to this algorithm as the **PC-IC algorithm**.

2.2 Algorithm in Detail: Population Version. We describe here the PC-IC algorithm for the population setting. Assume that all components of $X$ have fourth moment. Consider the spectral decomposition of

$$\Sigma = \Gamma \Lambda \Gamma^T,$$

where $\Lambda = diag(\lambda_1, \ldots, \lambda_d)$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d > 0$ are the eigenvalues of $\Sigma$, and $\Gamma = [\gamma_1, \cdots, \gamma_d]$ is an orthogonal matrix whose column $\gamma_{\ell}$ is the eigenvector of $\Sigma$ belonging to $\lambda_{\ell}$ for $\ell = 1, \ldots, d$. For simplicity of notation, we assume that $\Sigma$ has rank $d$. Note that our arguments below are not affected if the rank of $\Sigma$ is $r < d$. The only change required for this situation is to consider values for $p \leq r$ instead of $p < d$.

Let $\Gamma_p = [\gamma_1, \cdots, \gamma_p]$, and let $\Lambda_p = diag(\lambda_1, \ldots, \lambda_p)$ for $p < d$. We obtain a dimension reduction of $X$ to $p$-dimensions via a transformation defined by

$$Y = \Gamma_p^T (X - \mu).$$

The $Y$ vector is known as the $p$-dimensional PC score.

To extract features inherent in the original structure of $X$, the PC-IC algorithm applies FastICA (Hyvärinen et al., 2001) to this $Y$ in the following way. First, the sphering of $Y$ is accomplished by

$$Z(p) = \Lambda_p^{-1/2} Y = \Lambda_p^{-1/2} \Gamma_p^T (X - \mu).$$

Note that $Y$ and hence $Z(p)$ are $p$-dimensional random vectors. The PC-IC algorithm finds the direction $\alpha_{s,k}$, which results in $\beta_k(Z(p))$ for $k = 3, 4$, as
in equation 1.1:

\[ \alpha_{o,k} = \arg \max_{\alpha \in \mathbb{S}^{p-1}} B_k(\alpha|Z(p)). \]  

(2.1)

2.3 Algorithm in Detail: Empirical Version. The empirical version is based on the sample covariance matrix \( S \) of the data \( X_1, \ldots, X_N \). The spectral decomposition \( S = \hat{\Gamma} \hat{\Lambda} \hat{\Gamma}^T \) with descending order of eigenvalues naturally provides submatrices \( \hat{\Gamma}_p \) and \( \hat{\Lambda}_p \) defined similarly as \( \Gamma_p \) and \( \Lambda_p \), respectively. From these we obtain the \( p \)-dimensional sphered data, \n
\[ Z(p) = \hat{\Lambda}_p^{-1/2} \hat{\Gamma}_p^T (X_i - \hat{\mu}), \]  

for \( i = 1, \ldots, N \), where \( \hat{\mu} = \overline{X} \). In the empirical version corresponding to equation 2.1, we let IC1\(_k\) denote the PC-IC direction which gives rise to \( \hat{\beta}_k(Z(p)_1, \ldots, Z(p)_N) \). So for \( k = 3, 4 \),

\[ \text{IC1}_k = \hat{\alpha}_{o,k} = \arg \max_{\alpha \in \mathbb{S}^{p-1}} \hat{\beta}_k(\alpha|Z(p)_1, \ldots, Z(p)_N). \]  

(2.3)

with \( \hat{\beta}_k(\alpha|Z(p)_1, \ldots, Z(p)_N) \), as in equation 1.2.

We explore the underlying structure of the original data and extract informative features from the projection \( (\hat{\alpha}_{o,k}^T Z(p)_1, \ldots, \hat{\alpha}_{o,k}^T Z(p)_N) \) of the reduced data \( Z(p) \) \((i = 1, \ldots, N)\) onto IC1\(_k\) \((k = 3, 4)\). In Prasad et al. (2005), a similar algorithm, which is based on the first \( p \) directions IC1-IC\(_p\), was shown to be efficient in extracting features of high-dimensional data. A crucial—and still unanswered—question is that of the choice of \( p \), the dimension onto which the data are projected from their original \( d \) dimensions. Prasad et al.’s heuristic procedure for selecting the dimension \( p \) is based on the variance contribution in the PCA step: they use the cut-off dimension \( p_{95\%} \) which results in a contribution to variance of at least 95\%. This criterion may be reasonable if we analyze multivariate data by PCA only and do not have any other relevant information. However, a criterion based on the variance contribution does appear to be rather arbitrary and does not seem appropriate in a combined PC-IC, since ICA results in directions based on maximizing skewness and kurtosis. Our claim is that the dimension should be selected in accordance with the objective functions skewness and kurtosis if dimension reduction and feature extraction are implemented through these quantities.

3 Selection of the Dimension

This section presents the procedure for choosing the best dimension.

3.1 Methodology. The first direction, IC1\(_k\) \((k = 3, 4)\), is the most nongaussian direction, at least in theory. This direction and its degree of
nongaussianity will depend on the input data. If the dimension of the input data is too small, ICA may not find interesting directions, since valuable information may have been lost. On the other hand, a large $p$ requires a higher computational effort and the data may still contain noise or irrelevant information. A compromise between data complexity and available information is therefore required.

A naive idea is to select the dimension $p$ as

$$p_k = \arg \max_{2 \leq \ell \leq p^*} \hat{\beta}_k(Z(\ell)_1, \ldots, Z(\ell)_N),$$

for $k = 3, 4$, where $p^* = \min\{N, d\}$ is the upper bound (or $p^* = r$, where $r$ denotes the rank of the sample covariance matrix). It is easy to see that this selection rule has the trivial solution $p_k = p^*$ for both $k = 3, 4$. This implies that this criterion is not meaningful. Therefore, we need to formulate a criterion that includes a penalty term for large dimensions. We exploit the idea of including the AIC in our problem.

It is known that the AIC is based on a bias-adjusted version of the log likelihood. Since $\hat{\beta}_k(Z(p)_1, \ldots, Z(p)_N)$ is an estimator of $\beta_k(Z(p))$, it is natural to reduce the bias of the estimator, and we therefore aim to evaluate

$$\text{Bias}_k(p) = E[\hat{\beta}_k(Z(p)_1, \ldots, Z(p)_N)] - \beta_k(Z(p)).$$

for $k = 3, 4$. The evaluation of this bias now raises the question of which population should be employed.

3.2 **Null Structure.** What distribution should we assume as the null structure in our setting? We recall that our objective in the ICA step is to pursue a nongaussian structure as measured by skewness or kurtosis. To judge whether the obtained nongaussian structure is real, it is necessary to evaluate the performance of $\hat{\beta}_k(Z(p)_1, \ldots, Z(p)_N)$ under the structure with null skewness and kurtosis. This suggests that the multivariate gaussian distribution itself is one of the right null structures.

This kind of argument has been employed by Sun (1991) in conjunction with the significance level in projection pursuit. In what follows, all distributional results will be addressed under the assumption of gaussianity of $X$. We therefore note that $\beta_k(Z(p)) = 0$ for $k = 3, 4$ and for all $p \leq p^*$ under the gaussianity assumption. Further, the gaussianity assumption implies that $\hat{\beta}_k(Z(p)_1, \ldots, Z(p)_N)$ is an estimator of zero, and so

$$\text{Bias}_k(p) = E[\hat{\beta}_k(Z(p)_1, \ldots, Z(p)_N)].$$

3.3 **Theoretical Results.** Our problem requires the evaluation of $\text{Bias}_k(p)$ under the assumption that the $X_1, \ldots, X_N$ are independently and identically distributed random vectors from the $d$–dimensional gaussian
distribution \( N_d(\mu, \Sigma) \). Since skewness and kurtosis are location-scale invariant, without loss of generality we can assume that \( \mu = 0 \) and \( \Sigma = I_d \), the identity matrix. In this setting, as \( N \to \infty \),

\[
\sqrt{\frac{N}{k!}} \beta_k(Z(p)_1, \ldots, Z(p)_N) \Rightarrow T_k(p)
\]

for any \( p \leq p^* \), where \( \Rightarrow \) denotes convergence in distribution and \( T_k(p) \) is the maximum of a zero-mean gaussian random field on \( S^{p-1} \) with a suitable covariance function (see Machado, 1983; Kuriki & Takemura, 2001). Hence, for large \( N \),

\[
E \left[ \sqrt{\frac{N}{k!}} \beta_k(Z(p)_1, \ldots, Z(p)_N) \right] \simeq E[T_k(p)]
\]

for \( k = 3, 4 \). Kuriki and Takemura (2001) consider the probability \( P(T_k(p) > a) \) for \( a > 0 \). As this tail probability is difficult to calculate exactly, they give upper and lower bounds for this probability. We make use of this idea and develop upper and lower bounds for the expected value of \( T_k(p) \) for \( k = 3, 4 \) by noting that

\[
E[T_k(p)] = \int_0^\infty P(T_k(p) > a) \, da.
\]

This leads to our main result.

**Theorem 1.** For \( k = 3, 4 \), we have

\[
LB_k(p) \leq E[T_k(p)] \leq UB_k(p).
\]

For each dimension \( p \),

\[
LB_k(p) = \sum_{e=0, e \text{ even}}^{p-1} \omega_{p-e,k} \Lambda_{p-e,p+e}(\tan^2 \theta_k),
\]

\[
UB_k(p) = LB_k(p) + \left( \frac{2k - 2}{3k - 2} \right)^{1/2} E[X] \left[ 1 - \Psi(\theta_k, p) \right],
\]

where \( X \) denotes a chi-distributed random variable with \( \ell \) degrees of freedom,

\[
\rho = \rho(p, k) = \binom{p + k - 1}{k}.
\]
Dimension Selection with PC-IC

\[ \theta_k = \cos^{-1} \left[ \frac{2k - 2}{3k - 2} \right]^{1/2}, \]

\[ \omega_{p-e,k} = (-1)^{\omega} k^{(p-1)/2} \left( \frac{k-1}{k} \right)^{\omega/2} \frac{\Gamma[(p+1)/2]}{\Gamma[(p+1-e)/2]} \]

\[ \Psi(\theta_k, p) = \sum_{e=0 \text{ even}}^{p-1} \omega_{p-e,k} \bar{B}_{(p-e)/2,(p-p+e)/2} \left( \frac{2k - 2}{3k - 2} \right) \]

and \( \bar{B}_{\alpha,\beta}(\cdot) \) denotes the upper-tail probability of the beta distribution given by

\[ \bar{B}_{\alpha,\beta}(a) = \int_a^1 \frac{\xi^{\alpha-1} (1-\xi)^{\beta-1}}{B(\alpha, \beta)} d\xi \]

with the usual beta function \( B(\alpha, \beta) \). Further, for \( b > 0 \) and positive integers \( m \) and \( n \),

\[ \Lambda_{m,n}(b) = \Theta_m - \Omega_{m,n}(b) \]

with

\[ \Theta_m = E[\chi_m] = \sqrt{2} \frac{\Gamma[(m+1)/2]}{\Gamma(m/2)}, \]

\[ \Omega_{m,n}(b) = \int_0^\infty \int_0^\infty g_m(\xi) \bar{G}_n(b, \xi) d\xi da, \]

where \( g_m(\xi) \) is the density function of \( \chi_m^2 \) and \( \bar{G}_n(t) = P(\chi_n^2 > t) \) for \( t \geq 0 \). Expressions for \( \Omega_{m,n}(b) \) are

\[ \Omega_{m,2s}(b) = E[\chi_m] \left( \frac{1}{\sqrt{b+1}} \right)^{m+1} \left[ 1 + \sum_{i=1}^{\nu-1} \frac{1}{i} \left( \frac{b}{b+1} \right)^i \frac{1}{B(i, (m+1)/2)} \right], \]

\[ \Omega_{m,2n-1}(b) = \Omega_{m,1}(b) + E[\chi_m] \left( \frac{1}{\sqrt{b+1}} \right)^{m+1} \]

\[ \times \left[ \sum_{i=0}^{\nu-2} \frac{1}{i+1/2} \left( \frac{b}{b+1} \right)^{i+1/2} \frac{1}{B(i+1/2, (m+1)/2)} \right] \]

3.4 Practical Calculations. Exact evaluation of the upper and lower bounds \( \Gamma B_k(p) \) and \( L B_k(p) \) in theorem 1 is not easy. Even with Mathematica, the terms \( \Omega_{p-e,\rho-p+e}(b) \) in \( \Lambda_{p-e,\rho-p+e}(b) \) are not easily calculated for some combinations of \( \rho s \) and \( p s \) (in particular, for the cases \( \rho - p + e = 3 \text{mod} 4 \),...
while the terms $\Omega_{p-\epsilon, k}$ are not a problem. The reason these expressions are difficult to evaluate is that a summation of a large number of terms is required; the range of the individual terms is very large, and the weights are positive and negative. We therefore do not evaluate the formulas for $\Omega_{p-\epsilon, p-\epsilon}(b)$ directly but aim to obtain an approximate value.

From theorem 3.1 in Kuriki and Takemura (2001), a general form of our target can be expressed as

$$\Lambda_{m,n}(b) = \int_0^\infty \int_{a^2}^\infty g_m(\xi) G_n(b \xi) d \xi da,$$

where $G_n(\cdot)$ is the cumulative distribution function of $\chi^2_n$.

To obtain an approximation, we consider a truncated version of $\Lambda_{m,n}(b)$, defined as

$$\Lambda_{m,n}(b|T) = \int_0^T \int_{a^2}^\infty g_m(\xi) G_n(b \xi) d \xi da.$$

Consider a random variable $U$, uniformly distributed on an interval $[0, T]$, and let $F_U$ denote the cumulative distribution function of $U$. Furthermore, let $Y \sim \chi^2_m$. Assume that $Y$ and $U$ are mutually independent. Then we have

$$\Lambda_{m,n}(b|T) = T E_Y \left[ 1\{Y \geq a^2\} G_n(b Y) \right] da$$

$$= T E_U Y [1\{Y \geq U^2\} G_n(b Y)]$$

$$= T E_Y [G_n(b Y) E_U [1\{Y \geq U^2\}]]$$

$$= T E_Y [F_U(\sqrt{Y}) G_n(b Y)].$$

and hence $\Lambda_{m,n}(b|T)$ can be expressed as the expectation of $Y$ only. Therefore, a Monte Carlo estimate for $\Lambda_{m,n}(b|T)$ is

$$\tilde{\Lambda}_{m,n}(b|T) = \frac{T}{N} \sum_{j=1}^N F_U(\sqrt{Y_j}) G_n(b Y_j).$$

where the $Y_1, \ldots, Y_N$ are independently and identically distributed random variables from the $\chi^2_m$ distribution.

Obviously $\tilde{\Lambda}_{m,n}(b|T)$ is an unbiased estimate of $\Lambda_{m,n}(b|T)$, but not of $\Lambda_{m,n}(b)$. Moreover, it underestimates $\Lambda_{m,n}(b)$; this follows from the definition of the truncated version. The mean squared error (MSE) of $\Lambda_{m,n}(b|T)$, $\text{MSE} \left[ \tilde{\Lambda}_{m,n}(b|T) \right]$, is given as follows:
Dimension Selection with PC-IC

Theorem 2. As $T \to \infty$ and $N \to \infty$,

$$MSE \left[ \tilde{\Lambda}_{m,n}(b|T) \right] = O \left[ \exp(-T)T^{m-2} + \frac{T^2}{N} \right].$$

For practical purposes, we suggest to using the $100(1-\alpha)$% point of the $\chi^2$ distribution as $T_e$, which we denote by $\chi^2_{m,\alpha}$ for each $m$. (Recall that $m = p-e$ as in theorem 1.)

Then, using the expression of the squared bias (see the proof of theorem 2), it follows that

$$MSE \left[ \tilde{\Lambda}_{m,n}(b|\chi^2_{m,\alpha}) \right] = O \left[ (\frac{\alpha}{b})^2 + \frac{1}{N} \right].$$

From theorem 1, the actual value of $b$ is $\tan^2 \theta_k$, that is, $3/4$ for $k = 3$ and $2/3$ for $k = 4$. If we choose $\alpha = 10^{-10}$, $(\alpha/b)^2 \approx 1.33 \times 10^{-10}$ for $k = 3$, $(\alpha/b)^2 \approx 1.5 \times 10^{-10}$ for $k = 4$. This $(\alpha/b)^2$ is the quantity that will not vanish, but it can be as small as required, so will be negligible for practical purposes.

Now let

$$L B^T_k(p) = \sum_{e=0,e\text{even}}^{p-1} \omega_{p-e,k} \Lambda_{p-e,p-p+e}(\tan^2 \theta_k|T_e)$$

be a truncated version of $L B_k(p)$, where the $T_e$s are chosen differently for each $e$. Then combining $\tilde{\Lambda}_{m,n}(b|\chi^2)$ appropriately yields the estimate

$$\tilde{L} B^T_k(p) = \sum_{e=0,e\text{even}}^{p-1} \omega_{p-e,k} \tilde{\Lambda}_{p-e,p-p+e}(\tan^2 \theta_k|T_e)$$

of $L B_k(p|T)$. Hence, if we use independent samples distributed as $\chi^2_{p-e}$ for each $\tilde{\Lambda}_{p-e,p-p+e}$, it follows that

$$MSE \left[ \tilde{L} B_k(p|T) \right] = \sum_{e=0,e\text{even}}^{p-1} \omega_{p-e,k}^2 MSE \left[ \tilde{\Lambda}_{p-e,p-p+e}(\tan^2 \theta_k|T_e) \right].$$

3.5 The Proposed Criterion in Practice. The second term on the right-hand side of equation 3.1 can be calculated with sufficient accuracy, for example, with Mathematica. So we define

$$\tilde{L} B_k(p) = \tilde{L} B_k(p|T) + \left( \frac{2k-2}{3k-2} \right)^{1/2} E \left[ \chi^2_{p} [1 - \Psi(\theta_k, p)] \right].$$  (3.2)
The actual values of $\hat{ UB}_k(p)$ are tabulated in Table 3 in appendix B.

Our criterion for choosing the best or most informative dimension is based on the bias-adjusted version of $\hat{\beta}_k$ given by

$$\hat{I}_k(p) = \sqrt{\frac{N}{k!}} \hat{\beta}_k(Z(p)_1, \ldots, Z(p)_N) - \hat{ UB}_k(p).$$

(3.3)

For $k = 3, 4$ the dimension selector is thus given by

$$\hat{p}_k = \arg \max_{2 \leq p \leq p^*} \hat{I}_k(p), \ k = 3, 4.$$  (3.4)

Here, we have chosen $ UB_k(p)$ as an appropriate substitute for $E[T_k(p)]$. Our calculations show that $ LB_k(p)$ vanishes very rapidly with increasing dimension, and thus it contains little information, while the extra term in $ UB_k(p)$ dominates the behavior of $ UB_k(p)$. Since $ UB_k(p)$ is an upper bound for $E[T_k(p)]$, $\hat{I}_k(p)$ is smaller than the corresponding expression with $E[T_k(p)]$ instead of $ UB_k(p)$. It is not clear whether this affects the determination of $\hat{p}_k$ in equation 3.4.

4 Implementation: Real Data

4.1 The Data Sets. We apply the bias-adjusted selector of the best or most informative dimension to a number of real data sets. Apart from the last two, all our data sets are available on the UCI Machine Learning Repository (see Blake & Merz, 1998). References and more information about these data sets are available on their Web site. Some of these data sets have also been used in Prasad et al. (2005) in their feature selection. We calculate only the first IC from the $p$-dimensional scores, while Prasad et al. obtained and used all $p$ directions $IC_1, \ldots, IC_p$ for their efficient and accurate classification.

The Heroin Market data consists of 17 key indicator data series with observations (counts) collected over 66 consecutive months from January 1997 to December 2002. These data series form part of a broader project funded by the National Drug Law Enforcement Research Foundation (NDLERF). (For a detailed description of these series, see Gilmour & Koch, 2004.) This data set is different from all others used here in that it contains only 17 independent observations, but each observation has 66 features or dimensions. Thus, we are dealing with a HDLSS (high-dimension, low-sample-size) problem. As we shall see, the skewness-based criterion was able to find the most informative dimension, while the kurtosis-based dimension did not.

Finally, the Swiss Bank Notes data set stems from Flury and Riedwyl (1988). It deals with six measurements of 100 real and 100 counterfeit old
Swiss 1000 franc bank notes. A preliminary PCA showed that the first principal component scores are bimodal, and PC1 contributes about 67% to the total variance.

In the data sets Bupa Liver, Abalone, Glass, Wine, and Breast Cancer, the different features vary greatly in their range. Features on a large scale dominate the contribution to variance and would normally be picked as outliers in a PC analysis. As our primary interest is the determination of the optimal dimension rather than the specific derived features (see Hastie & Tibshirani, 2002), we used most data sets in their raw and standardized version in our PC-IC analysis. The “Comment” column in Table 1 indicates whether the raw data or the standardized data were used. Here, standardization refers to centering with the sample mean and normalizing with the standard deviation, usually across features, with the exception of the Heroin data set, where we standardized each time series, as the time series (rather than the features) were on different scales.

4.2 Calculation of the Best Dimension. Theorem 1 gives an expression for the upper bounds $UB_k$, which are used in the bias-adjusted selector, equation 3.4. The calculations of the optimal dimension were carried out in Matlab 7.0.4. For the PC part, we used pcaSM (see Marron, 2004), and for the IC part we used FastICA2.4 (see Havarinen & Oja, 2005). The data sets vary in their dimensions, from 6 to 33, and number of independent samples. For most of the data sets, we considered all samples as well as smaller subsets of the samples. Obviously the performance of $\hat{p}_k$ depends on the sample size. We shall see in this and the next section that the larger the sample size, the higher the value of $\hat{p}_k$.

For each data combination, that is, for a data set in its raw or standardized version, with all samples or with a subset of samples, we calculate the $p$-dimensional sphered data $Z(p), (i = 1, \ldots, N)$ as in equation 2.2 for $p = 2, \ldots, p^*,$ with $p^* = r \leq \min\{N, d\}$ where $r$ denotes the rank of the covariance matrix $S$.

For the $p$-dimensional sphered data, we calculate the most nongaussian direction with FastICA and the absolute skewness and absolute kurtosis. FastICA is an iterative technique with randomly chosen starting positions. We observed that different runs of ICA (even for the same input data) can converge to different components or different local optima. We found empirically that 10 runs of IC1 for each data combination were sufficient to guarantee a global skewness or kurtosis maximum of IC1. Using equations 3.3 and 3.4 we determine the best dimension.

4.3 Results and Interpretation. The results of the best dimension selection are given in Table 1. The second column details the dimension of each data combination: $N$ denotes the size of the sample, and $d$ denotes the dimension of the original feature space. The relative contributions to
variance due to the most nongaussian subspace are given in the Percentage Variance column. The last two columns contain the skewness and kurtosis dimension $\hat{p}_3$ and $\hat{p}_4$. Figures in parentheses in the column $\hat{p}_4$ denote a local maximum at the dimension given in parentheses. If the values for $\hat{p}_k$ differed for skewness and kurtosis, we reported two numbers in the variance column: the first one refers to the variance contribution obtained for the skewness dimension and the second to the contribution due to kurtosis.

Table 1 highlights the following:

- For the same size of the sample, the raw data resulted in a lower $\hat{p}_k$ value than the standardized data.
- Smaller samples result in a smaller $\hat{p}_k$ value.
- The $\hat{p}_3$ and $\hat{p}_4$ values are mostly identical.
- $\hat{p}_k$ does not appear to be related to the variance contribution of 95%.

How can we interpret the results presented in Table 1? Consider the Swiss Bank Notes. Measurements of six different quantities were made for 200 bank notes, and these six quantities correspond to properties of the bank notes such as the length of the bank notes. Our method selected the first four PC dimensions. This means that these PC combinations are regarded as “best” in capturing the nongaussian nature of the data. Since they are linear combinations of the original measurements, we learn how to combine the original six measurements in such a way that the dimensionality of the problem is reduced while the most informative information is preserved. For nongaussian data such as the Swiss Bank Notes, our criterion takes advantage of the nature of the data. For gaussian data, other dimension-selection criteria exist, and we briefly describe two such methods in section 5. Different criteria lead to different solutions and therefore have different interpretations.

To conclude this section, we return to the dimension-selection results of Table 1.

If the raw data contain outliers or observations on a different scale, a high contribution to variance is obtained with a small number of PC directions. Outlier directions are highly nongaussian, and a low-dimensional space will therefore be the most informative, that is, ICA will find the most nongaussian directions within a low-dimensional space.

Subsets of the original data set resulted in a lower $\hat{p}_k$ value than the total sample. This is consistent with the test for gaussianity carried out inside ICA. In the next section, we consider this aspect of the dimension selector further.

FastICA mostly results in very similar values for $\hat{p}_3$ and $\hat{p}_4$. When the $\hat{p}_4$ value was clearly bigger than $\hat{p}_3$, the calculations showed that the kurtosis approximation had a local maximum at $p = \hat{p}_3$. 
A comparison with the $\hat{p}_k$ values we obtained and the dimension $\hat{p}_95\%$ chosen in Prasad et al. (2005) (which is based on explaining 95% of variance) indicates that contribution to variance is not a good predictor for the best nongaussian dimension. Indeed, other effects, such as the differing scales of measurements, outliers, and sample size, affect the dimension of the most nongaussian space more strongly.

5 PCA-Based Approaches

We describe some model-based approaches to dimension reduction that have been used in the recent literature and examine their applicability. These approaches fall into two main groups:

- Probabilistic principal component analysis
- Bayesian principal component analysis
The approaches have the same starting point: a common underlying model,

\[ X = AV + \mu + \epsilon, \quad (5.1) \]

where \( X \) denotes a \( d \)-dimensional vector of observations with mean \( \mu \), \( V \) denotes a \( p \)-dimensional vector of latent variables, and \( \epsilon \sim N(0, \sigma^2 I_d) \) represents independent gaussian noise. In this framework, \( p < d \), \( A \) is a \( d \times p \) matrix, and both \( A \) and \( p \) are assumed to be unknown. The aim is to determine the unknown dimension \( p \) of the latent variables \( V \).

Probabilistic principal component analysis is described in Tipping and Bishop (1999) and Bishop (1999). Their key idea is that the dimension of the latent variables is that value of \( p \) that minimizes the estimated prediction error. We will denote this solution by \( \hat{p}(\mathcal{P}) \).

In addition to the assumptions of equation 5.1, we assume the following:

\begin{enumerate}
  \item [P1:] The latent vectors are independent and normal with \( V \sim N(0, I_p) \).
  \item [P2:] The observations are normal with \( X \sim N(\mu, AA^T + \sigma^2 I_d) \).
  \item [P3:] The conditional distributions \( X|V \) and \( V|X \) are normal with mean and variances derived from the means and variances of \( X, V, \) and \( \epsilon \).
\end{enumerate}

Under the assumptions P1–P3 the estimated prediction error is the negative log likelihood. For a sample of size \( N \) from this model, the log likelihood is given by

\[ \mathcal{L} = -\frac{N}{2} \left\{ d \log(2\pi) + \log |C| + \text{tr}(C^{-1}S) \right\}, \quad (5.2) \]

where \( C = AA^T + \sigma^2 I_d \).

The second approach, Bayesian principal component analysis, is described in Penny, Roberts, and Everson (2001) and Beckmann and Smith (2004). It is based on the key idea of a Bayesian framework proposed in Minka (2000). We first describe the main ideas of Minka’s method and then discuss how these ideas are used in Penny et al. and Beckmann and Smith.

Similar to Tipping and Bishop (1999), Minka (2000) starts with the gaussian model of equation 5.1, which satisfies P1 and P2. Instead of employing the maximum likelihood solution for \( A, \sigma \), and for the mean \( \mu \), Minka proposes to use a Bayesian framework that requires evaluation of the conditional probabilities \( \text{pr}(X_1, \ldots, X_N| A, \mu, \sigma^2) \).

Minka uses a noninformative prior for the mean \( \mu \) and Laplace’s method to approximate the conditional density. The final probability is conditioned
Dimension Selection with PC-IC

on the dimension $p$ of the latent variables $V$ only:

$$
pr(X_1, \ldots, X_N | p) \approx pr(U) \left( \prod_{j=1}^{p} \lambda_j \right)^{-N/2} \\
\times \hat{\sigma}^{-N(d-p)}(2\pi)^{(m+p)/2}|A_Z|^{-1/2}N^{-p/2},
$$

(5.3)

with $m = dp - p(p + 1)/2$,

$$
pr(U) = 2^{-p} \prod_{j=1}^{p} \Gamma \left( (d - j + 1)/2 \right) \pi^{-(d-j+1)/2},
$$

$$
|A_Z| = N \prod_{i=1}^{p} \prod_{j=i+1}^{d} (\hat{\lambda}_i^{-1} - \hat{\lambda}_j^{-1}) (\lambda_i - \lambda_j),
$$

$$
\hat{\lambda}_i = \lambda_i \text{ for } i \leq p, \quad \text{and} \quad \lambda_j = \hat{\sigma}^2 \text{ for } j > p.
$$

The product on the right-hand side of equation 5.3 is used to estimate the likelihood, and the value $\hat{p}^{(B)}$ that maximizes this product is the solution to the latent variable dimension selection.

Penny et al. (2001) use the linear mixing framework of ICA, described in equation 5.1, which deviates from the classical ICA framework in that it includes additive noise. In their model, $\mu = 0$ and assumptions P1 to P3 are replaced by the weaker assumptions B1 and B2:

B1: The latent vectors are independent.
B2: The observations $x$ conditional on $A$ and $V$ are normal with $(X | A, V) \sim N(AV, \sigma^2 I_d)$.

For this framework, they extend Tipping and Bishop’s (1999) approach in their section 12.3 by including a rotation matrix $Q$ in the expression for the mixing matrix $A$, as is usual in ICA. In section 12.4 the authors state that “for gaussian sources the ICA model reduces to PCA”. From there on, they propose the use of Minka’s method to determine the dimension $p$, but they also discuss “flexible” source models. We will return to this point.

The essence of Beckmann and Smith (2004) is Minka’s method; however, the authors take into account the limited data structure and the structure of noise as found in their FMRI applications. The latter refers to preprocessing of the noise, which includes standardizing the noise. For gaussian sources $V$, they propose the use of Minka’s $\hat{p}^{(B)}$, but they differ from Minka by using the mean as an estimate for $\mu$ instead of Minka’s prior distribution. After determining $\hat{p}^{(B)}$, the authors calculate the ICA rotation matrix $Q$ and the independent sources within a standard ICA framework. This last step
does not appear to relate to the choice of dimension, but it is relevant to the discussion of the applicability of their methodology. Tipping and Bishop (1999) and Minka (2000) propose clear, easy-to-use, and theoretically well-founded methods for selecting a lower dimension for the latent variables. Both articles make it very clear that the sources as well as the noise are gaussian. These properties enable them to combine PCA ideas with gaussian likelihood maximization and result in elegant solutions. The two solutions will, of course, be different, as different problems are solved (probabilistic and Bayesian, respectively).

Penny et al. (2001) and Beckmann and Smith (2004) have extended these algorithms to ICA frameworks, that is, to nongaussian sources. Penny et al. offer some discussion regarding a number of other source distributions and flexible source models and give two such examples where reference is made to specific (nongaussian) source distributions and how to accommodate them into Minka’s gaussian framework. This extension will no doubt work in the specific cases but does not lead to a general methodology. In contrast, Beckmann and Smith combine two approaches with inconsistent assumptions: finding the best \( \hat{p}^{(B)} \) by assuming that the sources are gaussian, and then finding \( \hat{p}^{(B)} \) nongaussian sources with standard ICA methodology. Since real data are never purely gaussian, and since, in practice, ICA finds nongaussian structure in gaussian randomly generated data, the method could lead to reasonable results for some data.

The question of interest here is whether PCA-based methods and their extensions described in Tipping and Bishop (1999) and Minka (2000) can result in the right dimension reduction without knowledge of the underlying model or distribution. Like the initial approaches of Tipping and Bishop and Minka, the method we propose is generic and supported by theory. We are not interested in finding modifications for specific cases as suggested by Penny et al. (2001) and Beckmann and Smith (2004).

In the following section we compare our method with those of Tipping and Bishop (1999) and Minka (2000). As the simulations will demonstrate, for the realistic nongaussian sources we use, neither of these methods is appropriate, and both are inferior to our method in terms of finding the correct dimension. This result is not surprising, since the assumptions of their models have been violated. These comparisons alone strongly indicate the need for our method, which does not replace the existing methods but complements them.

6 Implementation: Simulation Study

6.1 Nongaussianity, Sample Size, and PC Projections. We test the performance of our approach in simulations based on a mixture of gaussian and nongaussian directions. We determine the number of nongaussian directions with the PC-IC algorithm and FastICA. For comparison, we also
include results obtained with the probabilistic and the Bayesian approaches discussed in section 5.

Our theoretical arguments are based on the limit case \( N \to \infty \), while in practice, we deal with small and moderate sample sizes. The simulation results are therefore not expected to give a precise answer, especially for smaller sample sizes. A powerful test for non-gaussianity is given by the negentropy: for gaussian data, the negentropy is zero, while positive negentropy indicates deviations from gaussianity. The notion of negentropy relies on the joint probability density of the data. Density estimation for multivariate or high-dimensional data becomes more complex as the dimension increases, and good estimates in high dimensions require a large number of data. Even then, current density estimation methods are not always very accurate. For this reason, approximations to the negentropy are commonly used. FastICA uses the skewness and kurtosis as approximations to the negentropy (refered to as nonlinearities in their framework). Below we discuss the capability of skewness and kurtosis to distinguish gaussian from non-gaussian models. This analysis informs our choice of simulation models.

6.1.1 Small Sample Sizes. If the sample size is small, the empirical distribution of non-gaussian data may not differ sufficiently from that of a gaussian sample of the same size. This means that in a test for gaussianity, the data appear consistent with the gaussian model and will not be rejected as non-gaussian. Thus, a technique such as ICA, which tests the deviation from gaussianity in the new directions, tends to underestimate the number of non-gaussian directions, and hence also the dimension of the most non-gaussian subspace. For small samples, \( \hat{p}_k \) is therefore likely to be smaller than the true number of non-gaussian dimensions.

6.1.2 Large Sample Sizes. If the sample size is large, the model is a better approximation to the theory, and the dimension \( \hat{p}_k \) should therefore give an accurate description of the presence of non-gaussian structure in the data. In the large sample case, non-gaussianity tests are powerful and easily find deviations from gaussianity. Thus, non-gaussian directions will appear as non-gaussian.

FastICA employs a very sensitive test for non-gaussianity: indeed, even for samples from the multivariate gaussian distribution, FastICA will find non-gaussian directions (see Koch et al., 2005). Thus, for large samples, the number of non-gaussian directions ICA finds could be an overestimate. Consequently, \( \hat{p}_k \) is likely to be bigger than the true number of non-gaussian dimensions.

6.1.3 Skewness as a Test for Non-gaussianity. The skewness is based on the third moment of a random vector or sample. Symmetric distributions have a zero skewness. As the skewness is location and scale invariant, the
skewness approximation to negentropy will not be able to differentiate between the gaussian distribution and a distribution that is symmetric about its mean. In particular, ICA-based skewness tests will not be able to differentiate between gaussian and uniform data, or between data from the gaussian and the double exponential distributions, and \( \hat{p}_k \) will severely underestimate the nongaussian structure.

6.1.4 Kurtosis as a Test for nongaussianity. The kurtosis is based on the fourth moment of a random vector or sample. Kurtosis as in equation 1.2 is adjusted to yield to zero kurtosis for multivariate gaussian random vectors. For supergaussian distributions, kurtosis is positive, while subgaussian distributions lead to negative kurtosis values. However, the kurtosis test in FastICA is biased toward positive kurtosis, and it is therefore harder to differentiate subgaussian distributions from the gaussian distribution. Consequently, for subgaussian data, \( \hat{p}_k \) is smaller than the true number of subgaussian data directions. Indeed, our simulations with the uniform distribution confirmed this bias in the kurtosis estimates.

6.1.5 Support of the Distribution. The gaussian distribution is characterized by its infinite support. Distributions with compact support are therefore clearly distinct from those with infinite support. Neither skewness nor kurtosis tests are able to differentiate on the basis of the support of the distribution. Specifically, the beta distribution with positive and equal parameter values has skewness and kurtosis similar to those of the gaussian distribution. FastICA is not able to distinguish between them. So, again, \( \hat{p}_k \) is smaller than the true number of nongaussian data directions.

6.1.6 PC Projections. The first step in the PC-IC algorithm consists in reducing the \( d \)-dimensional original data to \( p \)-dimensional scores for \( p = 2, \ldots, r \), where \( r = \min\{\min(d, N), \text{rank}(S)\} \). The PC scores are weighted averages of the original data. As the dimension \( d \) becomes large (\( d \geq 30 \)), a central limit effect seems to be noticeable in the projected data; thus, the scores are more gaussian than the original data, and \( \hat{p}_k \) will be smaller than the true number of nongaussian directions.

6.2 Simulation Models and Results

6.2.1 Simulation Strategy. The above discussion informs our choice of distributions and our simulation strategy. Since the accuracy of \( \hat{p}_k \) depends on the type of distribution as well as the sample size, we demonstrate its performance across a number of distributions and sample sizes and also show its relative accuracy by comparing it for models that differ only by the number of nongaussian dimensions but are the same otherwise.
Dimension Selection with PC-IC

Table 2: Dimensions \( d \) for simulated models with \( d_{ng} \) nongaussian dimensions.

<table>
<thead>
<tr>
<th>( d )</th>
<th>5</th>
<th>5</th>
<th>7</th>
<th>7</th>
<th>10</th>
<th>10</th>
<th>20</th>
<th>20</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>( d_{ng} )</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

6.2.2 Nongaussian Models and the Mixed Data. The development of our approach and theoretical arguments requires \( N \) independent observations \( X_1, \ldots, X_N \) without reference to any model. The only assumption we require is that the observations have fourth moments. Such \( d \)-dimensional data correspond best to the generic ICA description:

\[
X = AV, \tag{6.1}
\]

with \( A \) the \( d \times d \) mixing matrix and \( V \) a \( d \)-dimensional vector of independent sources. This ICA model is different from that in equation 5.1, which forms the bases of the PCA-type approaches.

We choose three nongaussian families of distributions, which set them apart from the gaussian distribution in ways that skewness or kurtosis should be able to discern:

- The beta distribution with parameters 5 and 1.2. This distribution has compact support and is asymmetric.
- A mixture of two gaussians consisting of one-quarter of the samples from the gaussian distribution \( N(-1, 0.49) \) and three-quarters of the data from the gaussian \( N(4, 1.44) \). This distribution is strongly bimodal and asymmetric.
- The exponential distribution with parameter \( \lambda = 2 \). This is a supergaussian that is also asymmetric.

For each of these models, we generate independent source data. We generate \( N \times d_{ng} \) independent random samples from the distribution \( F \), where \( F \) denotes any one of the three distributions above and \( N \times d_g \) independent samples from the multivariate gaussian distribution with zero mean and covariance matrix \( I_{d_g} \). Then \( d = d_{ng} + d_g \) denotes the dimension of the independent source data \( V \). In the simulations, we vary \( d \), and \( d_{ng} \) within \( d \). For the sample size \( N \), we use the values \( N = 20, 50, 100, 200, 500, 1000, 2000 \). Table 2 summaries the combinations we use.

We apply a \( d \times d \) mixing matrix \( A \) to the independent source data \( V_1, \ldots, V_N \) to obtain the observed or mixed data \( X_1, \ldots, X_N \) given in equation 6.1. We use a class of Higham test matrices (see Davies & Higham, 2000). These matrices are generated in Matlab with the command \( A = \text{gallery}(\text{randcolu}, d) \). The columns in the random matrix \( A \) have unit length. The singular values of this matrix can be specified, but we do not use
this option. Matrices of this form are invertible. For all simulations based on a specific dimension $d$, the same matrix $A = A(d)$ is used.

6.2.3 Approaches for Calculating the Dimension $\hat{p}$. We apply the PC-IC algorithm to mixed data $X_1, \ldots, X_N$ obtained from sources $V = V(N, d, dng, F)$. The independent sources are characterized by $d, dng$, the nongaussian distribution $F$, and the sample size $N$. For $p = 2, \ldots, d$ we determine the most nongaussian IC1 direction by taking the maximum over 10 IC1 repetitions obtained with FastICA2.4. From these values for each $p$, we determine the $\hat{p}_k$ ($k = 3, 4$) as in equation 3.3.

Our method is based on skewness and kurtosis. We carry out separate simulations for skewness and kurtosis, which result in the values $\hat{p}_3$ for skewness and $\hat{p}_4$ for kurtosis.

In addition we apply the probabilistic PCA approach of Tipping and Bishop (1999), which we refer to as PPCA, and the Bayesian approach of Minka (2000), referred to as BPCA, to the data $X_1, \ldots, X_N$. For the probabilistic PCA, we calculate the log likelihood of equation 5.2 and find its maximizer $\hat{p}_P$. For the Bayesian approach, we find the maximizer $\hat{p}_B$ of equation 5.3. In the PCA-based simulations, $p$ ranges over the values $\{2, \ldots, d-1\}$. Our approach considers values $p \in \{2, \ldots, d\}$, since ICA requires at least a two-dimensional subspace to find a nongaussian direction. As our approach is nonparametric and treats all observations the same way, the flexible sources referred to in Penny et al. (2001) are not appropriate, nor is the preprocessing of the noise as described in Beckmann and Smith (2004) applicable, since there is no additive noise in the model of equation 6.11. Thus, the algorithms we can compare our method with are the probabilistic PCA given in equation 5.2 and the Bayesian PCA given in equation 5.3.

6.2.4 Simulation Results. For 1000 runs with the sources $V(N, d, dng, F)$, we recorded the proportion of times that dimension $\hat{p}$ was selected.

Figures 1 to 3 show typical performances for our method and the two PCA-based methods. The pattern is the same in each figure. Each row of subplots presents the results of one method. In the first row, we present our skewness results, referred to as PC-IC skewness in the figures. In Figures 1 and 2, the second row shows our kurtosis results (PC-IC kurtosis). The second last row displays the results of PPCA, and the last row shows the results of BPCA. We display the results for the sample sizes $N = 50, 100, 200, 500$, and 1000. Each column of subplots displays the results of a fixed sample size; the first column has results for $N = 50$, the next for $N = 100$, and the last for $N = 1000$. The $x$-axis of each subfigure displays the range of dimensions $1, \ldots, d$, and the $y$-axis shows the percentage that a particular dimension $p$ was the maximizer in 1000 runs. The correct dimension, namely, the number of nongaussian dimensions, is indicated by a square shape in each plot, for example, dimension 4 in Figure 1.
Figure 1: Proportion $\hat{p}$ against dimension for five-dimensional distribution. Sample sizes $N = 50, 100, 200, 500, 1000$ vary along rows. $\hat{p}$ with PC-IC skewness (top row), with PC-IC kurtosis (second row), with PPCA (third row), with BPCA (bottom row). Correct dimension indicated by a square, here at $p = 4$.

Figure 1 displays results for low dimensions. For five-dimensional data, we use the beta distribution for four of the independent sources. The results are displayed in the 20 subfigures of Figure 1. The PC-IC skewness criterion (top row) finds the correct dimension for $N = 100$ and 200. For larger values of the sample size, $\hat{p}_3$ is a little bigger than the true value. This is consistent with our comments about large samples. The results obtained with kurtosis are not as good as those for skewness. Although kurtosis finds the correct dimension for the two largest sample sizes, the percentages for all values of $\hat{p}$ are quite high. Further, skewness is computationally preferable, as kurtosis calculations take two to four times longer. The third row shows the results from PPCA: for all sample sizes, the maximum occurs at the correct dimension. Finally, the results for BPCA indicate that BPCA finds the correct dimension for the smallest sample size ($N = 50$), but for all other sample sizes, BPCA has a maximum at $\hat{p}_B = 1$.

Figure 2 shows the result for 10-dimensional data. Here we use the bimodal distribution in three dimensions. For this moderate dimension, PC-IC skewness (top row) finds the correct dimension for the larger sample sizes and underestimates it by one for the smaller values. This is expected: $\hat{p}_3$ tends to increase with sample size. The kurtosis results are
very similar in that the correct dimension is underestimated by one for the smaller sample sizes, but again, the skewness results are superior to the kurtosis results. The PPCA plots all have a maximum at 1. Unlike the results for PC-IC skewness, the dimension of the maximum does not change with sample size and indeed is wrong in each case. BPCA shows a trend that was not so clear for the five-dimensional data of Figure 1: for low sample sizes, the maximum of BPCA occurs at the last value of its range, \( d - 1 \), and for higher sample sizes, the maximum occurs at the 1.

Figure 3 shows the results for 20-dimensional data with four bimodal dimensions. As the kurtosis results are inferior to the skewness results, we present only the skewness results here. For the three larger samples, PC-IC skewness determined the correct dimension convincingly. For 20-dimensional data, sample sizes of 50 or 100 are very small, and it is therefore not surprising that the skewness criterion underestimates the dimension. The second row of figures shows that PPCA picked \( d - 1 = 19 \) as the dimension of the latent variables for each sample size. This choice is far from the correct choice: four dimensions. As in the results in Figure 2, PPCA appears to be independent of the sample size, but this time, PPCA picks
Figure 3: Proportion $\hat{p}$ against dimension for 20-dimensional distribution with three bimodal dimensions. Sample sizes $N = 50, 100, 200, 500, 1000$ vary along rows. $\hat{p}$ with PC-IC skewness (top row), with PPCA (middle row), with BPCA (bottom row). Correct dimension indicated by a square, here at $p = 4$.

The other extreme dimension, $d - 1$, instead of 1 in the 10-dimensional data. We will comment on this discrepancy between these two results in the last paragraph of this section. BPCA shows the same pattern as for the 10-dimensional data of Figure 2: for the two smallest sample sizes, the maximum occurs at the largest dimension (at $\hat{p}^{(B)} = 19$), and for the two largest sample sizes, the maximum occurs at 1. For $N = 200$, we note that the peak has now moved to $\hat{p}^{(B)} = 3$, so happens to be close to the true value.

These figures show that PC-IC with the skewness criterion produces good results for nongaussian distributions of diverse dimensions and different sample sizes. PC-IC with skewness is mostly superior to PC-IC with kurtosis in determining the correct dimension. This is a little surprising, but the results are very convincing. Since skewness is calculated more efficiently than kurtosis with FastICA, we recommend the use of PC-IC with skewness.

When the dimension of the data is small as in Figure 1, PC-IC with skewness yields correct results for the smaller sample sizes we considered, while for larger samples, the selected dimension can be too large. For small $N$ and moderate or larger dimensions $d \geq 10$, $\hat{p}_3$ tends to be a little smaller than $dN$, but as $N$ increases, the PC-IC selects the correct dimension. These results are consistent with the observations of section 6.1 and the results in Table 1.

A comparison between PC-IC with skewness and the PCA-based methods shows convincingly that our method is clearly superior to either of these two for nongaussian data and models as in equation 6.1. This is not
surprising, since both methods were developed for gaussian data with additive gaussian noise. PPCA appeared to select the correct dimension for the low-dimensional simulations based on the beta distribution. In some sense, this distribution is closer to the normal distribution than the bimodal, so a possible explanation for the good performance of PPCA is that it shows some distributional robustness, that is, PPCA can handle models that are close to the gaussian model for which it has been designed. The bimodal distribution is very different from the gaussian distribution, and it is therefore not surprising that PPCA does not yield such good results. What is curious, however, is the jump from the lowest to the highest dimension that occurred between Figures 2 and 3 or between dimensions 10 and 20. Further simulations (not presented here) showed that PPCA depended strongly on the model, here the class of mixing matrices, so PPCA is not model robust. These simulations indicate that the class of models and distributions for which PPCA works is quite restrictive and thus highlight a serious drawback of PPCA. In contrast, BPCA appears to be model robust and always has a very clear peak where the maximum occurs, but BPCA appears too sensitive to sample size. We observed (but do not show this here) that the selected dimension of the maximum decreases from $d - 1$ to 1 as the sample size increases from 100 to 120. In summary, the results show that our method produces good results, where neither of the PCA-based methods is applicable, so complements these methods.

7 Conclusion

A criterion for choosing the dimension in dimension-reduction problems has been proposed. The criterion is essentially based on an approximation of bias-adjusted skewness and kurtosis for PC scores. Applications to real data sets reveal that the proposed criterion returns a higher dimension than $\hat{p}_{95\%}$ based on PC alone. This seems to be obvious since nongaussianity cannot be captured by the variance of low-dimensional projections alone as PCA aims for. Our proposed criterion includes as a vital part pursuing of nongaussianity, and hence it should yield higher dimensions. We have shown in simulation studies that the criterion performs well for simulated structures and performs better than existing PCA-based methods for non-gaussian data. The ICA part of the criterion seems to make its performance quite similar to that of testing gaussianity based on skewness and kurtosis, and in fact this has been observed in our simulation studies. Therefore, it is desirable to consider simulations of structures that are used in the literature as alternative hypotheses for testing gaussianity. As mentioned in section 6.1.5, more accurate selection of the dimension would be possible provided we could differentiate between compact and noncompact support of the underlying structure. However, we reserve this problem for a future topic.
Appendix A: Proof of Theoretical Results

The following lemma is useful to prove theorem 1.

Lemma 1. (1) For \( b \geq 0 \), we have

\[
\int_0^\infty \overline{G}_m[(b + 1)t^2]dt = \sqrt{\frac{2}{b + 1}} \frac{\Gamma[(m + 1)/2]}{\Gamma(m/2)} = \frac{E[\chi_m]}{\sqrt{b + 1}}.
\]

(2) Fix integers \( i \) and \( m \) and put \( k = 2i + m \). For \( a, b \geq 0 \) the tail probability \( \overline{G}_k \) satisfies

\[
\frac{\overline{G}_k[(b + 1)a^2]}{(b + 1)^{k/2}} = \frac{(2b)^{i/2} \Gamma(k/2)}{\Gamma(m/2)} = \int_0^\infty e^{-bt}(bt)^{(2i+1)/2}g_m(t)dt.
\]

(3) Fix integers \( i \) and \( m \) and put \( \ell = 2i + 1 + m \). For \( a, b \geq 0 \) the tail probability \( \overline{G}_\ell \) satisfies

\[
\frac{\overline{G}_\ell[(b + 1)a^2]}{(b + 1)^{\ell/2}} = \frac{b \Gamma(\ell/2)}{\Gamma(m/2)} = \int_0^\infty e^{-bt}\left(\frac{bt}{2}\right)^{(2i+1)/2}g_m(t)dt.
\]

Proof of Lemma 1. 1. With \( \overline{G}_m(t^2) = P(\chi_m^2 \geq t^2) = P(\chi_m \geq t) \) for \( t \geq 0 \), using the change of variables \( u = t^2 \), we find that

\[
\int_0^\infty \overline{G}_m(t^2)dt = \int_0^\infty \overline{G}_m(u) \frac{du}{2\sqrt{u}}
\]

\[
= \overline{G}_m(u)\sqrt{u}|_0^\infty + \int_0^\infty \sqrt{u}g_m(u)du
\]

\[
= E[\chi_m]
\]

since the first partial integrand equals 0. From this equation and applying the change of variable \( u = (b + 1)t^2 \), we get

\[
\int_0^\infty \overline{G}_m[(b + 1)t^2]dt = \int_0^\infty \overline{G}_m(u) \frac{du}{2\sqrt{u(b + 1)}}
\]
\[ E[\chi_m] = \frac{1}{\sqrt{b+1}} \]
\[ = \sqrt{\frac{2}{b+1}} \Gamma[(m+1)/2] \Gamma(m/2) \]

2. Fix \( i \) and \( m \). For \( a, b \geq 0 \), let \( K \) denote the integral:

\[
K = \int_{a^2}^{\infty} e^{-(b+1)t/2} \frac{t^{m+1}}{2^m \Gamma(m/2)} \frac{\Gamma((2i+1)/2)}{\Gamma((2i+m)/2)} dt
\]

Put \( k = 2i + m \). The change of variable \( r = (b+1)t \), and applying the definition of the tail probability \( G_k \) leads to

\[
K = \frac{(2b)^i \Gamma(k/2)}{\Gamma(m/2)} \int_{a^2}^{\infty} \frac{e^{-(b+1)t/2} r^{k/2-1}}{2^{k/2} \Gamma(k/2)} \frac{\Gamma((2i+1)/2)}{\Gamma((2i+m)/2)} dr
\]

\[
= \frac{(2b)^i \Gamma(k/2)}{(b+1)^{k/2} \Gamma(m/2)} G_k[(b+1)a^2].
\]

3. Fix \( i \) and \( m \). For \( a, b \geq 0 \) let \( L \) denote the integral

\[
L = \int_{a^2}^{\infty} e^{-(b+1)t/2} \frac{t^{2i+1/m}}{2^{m+1+i} \Gamma(m/2)} \frac{\Gamma((2i+1+m)/2)}{\Gamma((2i+m)/2) \Gamma(m/2)} dt
\]
Dimension Selection with PC-IC

With \( \ell = 2i + 1 + m \), and the change of variable \( r = (b + 1)t \), we get

\[
L = \frac{b^{(2i+1)/2} \Gamma(\ell/2)}{\Gamma(m/2)} \int_0^\infty \frac{e^{-(b+1)t/2} r^{\ell/2-1}}{2^{\ell/2} \Gamma(\ell/2)} dr
\]

\[
= \frac{b^{(2i+1)/2} \Gamma(\ell/2)}{(b+1)^{\ell/2} \Gamma(m/2)} \times C_\ell[(b+1)a^2].
\]

**Proof of Theorem 1.** From theorems 3.1 and 3.3 of Kuriki and Takemura (2001), we have

\[
Q_{k,p}^L(a) \leq P(T_k(p) \geq a) \leq Q_{k,p}^H(a)
\]

for \( a > 0 \), where

\[
Q_{k,p}^L(a) = \sum_{e=0}^{p-1} \omega_{p-e,k} Q_{p-e,p-r}(a^2, \tan^2 \theta_k),
\]

\[
Q_{k,p}^H(a) = Q_{k,p}^L(a) + G_p(a^2(1 + \tan^2 \theta_k)) \left(1 - \frac{\text{Vol}(M_0)}{\text{Vol}(S^{p-1})}\right),
\]

\[
Q_{m,n}(a, b) = \int_a^\infty g_n(\xi)(1 - G_n(b \xi))d\xi,
\]

\( \text{Vol}(M_0) \) is the volume of tube \( M_0 \) which is fully described in Kuriki and Takemura (2001), and \( \text{Vol}(S^{p-1}) \) is the total volume of \( S^{p-1} \). Using the expression of \( Q_{k,p}^L \) and \( Q_{k,p}^H \), it follows that

\[
LB_k(p) = \sum_{e=0}^{p-1} \omega_{p-e,k} \int_0^\infty Q_{p-e,p-r}(a^2, b) da
\]

\[
\leq E[T_k(p)]
\]

\[
\leq \sum_{e=0}^{p-1} \omega_{p-e,k} \int_0^\infty Q_{p-e,p-r}(a^2, b) da
\]

\[
+ \left(1 - \frac{\text{Vol}(M_0)}{\text{Vol}(S^{p-1})}\right) \int_0^\infty G_p((1 + \tan^2 \theta_k)a^2) da
\]

\[
= UB_k(p).
\]
The ratio $\text{Vol}(M_\theta)/\text{Vol}(S^p)$ can be expressed using $\Psi(\theta_k, p)$ as in lemmas 3.3 and 3.4 in Kuriki and Takemura (2001). Direct but tedious calculations of the integral $Q_{m,n}(a^2, b)$ by means of formulas 18.18 and 18.19 in Johnson, Kotz, and Barakrishnan (1994), and the previous lemma yield the expressions tabulated in the theorem.

**Proof of Theorem 2.** First, we note that

$$\text{MSE} \left[ \hat{\Lambda}_{m,n}(b|T) \right] = E \left[ \left( \hat{\Lambda}_{m,n}(b|T) - \Lambda_{m,n}(b) \right)^2 \right]$$

$$= \left\{ \text{Bias} \left( \hat{\Lambda}_{m,n}(b|T) \right) \right\}^2 + \text{Var} \left[ \hat{\Lambda}_{m,n}(b|T) \right]$$

$$= \left\{ \Lambda_{m,n}(b|T) - \Lambda_{m,n}(b) \right\}^2 + \text{Var} \left[ \Lambda_{m,n}(b|T) \right].$$

Using expression 18.18 in Johnson et al. (1994) again, we have for even $m$ that

$$|\Lambda_{m,n}(b) - \Lambda_{m,n}(b|T)| = \int_T^\infty \int_\Bbb{R}^2 g_m(\xi)G_n(b\xi)d\xi da$$

$$\leq \int_T^\infty \int_\Bbb{R}^2 g_m(\xi)d\xi da$$

$$= \int_T^\infty G_m(a^2)da$$

$$= \sum_{i=0}^{(m/2)-1} \frac{1}{i!} \int_T^\infty \exp \left( -\frac{a^2}{2} \right) \left( \frac{a^2}{2} \right)^i da$$

$$= \sum_{i=0}^{(m/2)-1} \frac{\Gamma((i + 1)/2)}{\sqrt{2}^{i+1}i!} \hat{G}_{i+1}(T^2)$$

$$\leq \hat{G}_{m/2}(T^2) \sum_{i=0}^{(m/2)-1} \frac{\Gamma((i + 1)/2)}{\sqrt{2}^{i+1}i!}$$

$$= O(\exp(-T/2)T^{(m/2)-1}).$$

The proof for odd $m$ is similar using equation 18.19 in Johnson et al. (1994). The evaluation of the variance is trivial.

**Appendix B: Table of Upper Bounds**

The table here contains the upper bounds $\overline{\Omega}_k(p)$ for dimensions $p = 2, \ldots, 50$ and $k = 3, 4$ as in equation (7) of section 3.5. The kurtosis values
Dimension Selection with PC-IC

Table 3: Skewness and Kurtosis Upper Bounds.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>UB-skewness</th>
<th>UB-kurtosis</th>
<th>Dimension</th>
<th>UB-skewness</th>
<th>UB-kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.5993</td>
<td>1.7536</td>
<td>27</td>
<td>45.6915</td>
<td>128.2290</td>
</tr>
<tr>
<td>3</td>
<td>2.9790</td>
<td>2.8356</td>
<td>28</td>
<td>48.1634</td>
<td>137.4000</td>
</tr>
<tr>
<td>4</td>
<td>3.3430</td>
<td>4.5500</td>
<td>29</td>
<td>50.6783</td>
<td>146.8870</td>
</tr>
<tr>
<td>5</td>
<td>4.4404</td>
<td>6.4576</td>
<td>30</td>
<td>53.2353</td>
<td>156.6900</td>
</tr>
<tr>
<td>6</td>
<td>5.6317</td>
<td>8.6776</td>
<td>31</td>
<td>55.8339</td>
<td>166.8090</td>
</tr>
<tr>
<td>7</td>
<td>6.9225</td>
<td>11.2116</td>
<td>32</td>
<td>58.4734</td>
<td>177.2450</td>
</tr>
<tr>
<td>8</td>
<td>8.2636</td>
<td>14.0606</td>
<td>33</td>
<td>61.1532</td>
<td>187.9970</td>
</tr>
<tr>
<td>9</td>
<td>9.6954</td>
<td>17.2250</td>
<td>34</td>
<td>63.8726</td>
<td>199.0650</td>
</tr>
<tr>
<td>11</td>
<td>12.7727</td>
<td>24.5010</td>
<td>36</td>
<td>69.4285</td>
<td>222.1490</td>
</tr>
<tr>
<td>12</td>
<td>14.4123</td>
<td>28.6129</td>
<td>37</td>
<td>72.2634</td>
<td>234.1660</td>
</tr>
<tr>
<td>13</td>
<td>16.1157</td>
<td>33.0409</td>
<td>38</td>
<td>75.1380</td>
<td>246.4990</td>
</tr>
<tr>
<td>14</td>
<td>17.8806</td>
<td>37.7849</td>
<td>39</td>
<td>78.0458</td>
<td>259.1480</td>
</tr>
<tr>
<td>15</td>
<td>19.7050</td>
<td>42.8451</td>
<td>40</td>
<td>80.9920</td>
<td>272.1130</td>
</tr>
<tr>
<td>16</td>
<td>21.5870</td>
<td>48.2214</td>
<td>41</td>
<td>83.9745</td>
<td>285.3950</td>
</tr>
<tr>
<td>17</td>
<td>23.5251</td>
<td>53.9138</td>
<td>42</td>
<td>86.9630</td>
<td>298.9930</td>
</tr>
<tr>
<td>18</td>
<td>25.5157</td>
<td>59.9224</td>
<td>43</td>
<td>90.0460</td>
<td>312.9070</td>
</tr>
<tr>
<td>19</td>
<td>27.5629</td>
<td>66.2473</td>
<td>44</td>
<td>93.1343</td>
<td>327.1370</td>
</tr>
<tr>
<td>20</td>
<td>29.6600</td>
<td>72.8883</td>
<td>45</td>
<td>96.2571</td>
<td>341.6840</td>
</tr>
<tr>
<td>21</td>
<td>31.8075</td>
<td>79.8455</td>
<td>46</td>
<td>99.7390</td>
<td>356.5460</td>
</tr>
<tr>
<td>22</td>
<td>34.0042</td>
<td>87.1189</td>
<td>47</td>
<td>102.6050</td>
<td>371.7250</td>
</tr>
<tr>
<td>23</td>
<td>36.2491</td>
<td>94.7085</td>
<td>48</td>
<td>105.8290</td>
<td>387.2200</td>
</tr>
<tr>
<td>24</td>
<td>38.5413</td>
<td>102.6140</td>
<td>49</td>
<td>109.0860</td>
<td>403.0320</td>
</tr>
<tr>
<td>25</td>
<td>40.8796</td>
<td>110.8360</td>
<td>50</td>
<td>112.3760</td>
<td>419.1600</td>
</tr>
<tr>
<td>26</td>
<td>43.2633</td>
<td>119.3750</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

increase much faster than skewness bounds, and both increase at a rate that is faster than the linear rate.

Acknowledgments

We gratefully acknowledge the reviewers’ suggestions to include comparisons with probabilistic dimension selection methods.

References


I. Koch and K. Naito


Dimension Selection with PC-IC


Received November 21, 2005; accepted June 16, 2006.