HAL open science

Intrinsic Dimension Estimation by Maximum Likelihood in Isotropic Probabilistic PCA

Charles Bouveyron, Gilles Celeux, Stéphane Girard

To cite this version:

Charles Bouveyron, Gilles Celeux, Stéphane Girard. Intrinsic Dimension Estimation by Maximum Likelihood in Isotropic Probabilistic PCA. Pattern Recognition Letters, 2011, 32 (14), pp.1706-1713. 10.1016/j.patrec.2011.07.017 . hal-00440372v3

HAL Id: hal-00440372

https://hal.science/hal-00440372v3

Submitted on 11 Jul 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Intrinsic Dimension Estimation by Maximum Likelihood in Isotropic Probabilistic PCA

Charles Bouveyron[a], Gilles Celeux[b], Stéphane Girard[c]

[a] *Laboratoire SAMM, EA 4543, University Paris 1 Panthéon–Sorbonne 90 rue de Tolbiac, 75013 Paris, France*
[b] *Select, Inria Saclay-Île de France, Dept. de mathématiques Université Paris-Sud, 91405 Orsay Cedex, France*
[c] *Mistis, Inria Rhône-Alpes & LJK, Inovallée, 655 av. de l'Europe, Montbonnot 38334 Saint-Ismier cedex, France*

Abstract

A central issue in dimension reduction is choosing a sensible number of dimensions to be retained. This work demonstrates the surprising result of the asymptotic consistency of the maximum likelihood criterion for determining the intrinsic dimension of a dataset in an isotropic version of Probabilistic Principal Component Analysis (PPCA). Numerical experiments on simulated and real datasets show that the maximum likelihood criterion can actually be used in practice and outperforms existing intrinsic dimension selection criteria in various situations. This paper exhibits and outlines the limits of the maximum likelihood criterion. It leads to recommend the use of the AIC criterion in specific situations. A useful application of this work would be the automatic selection of intrinsic dimensions in mixtures of isotropic PPCA for classification.

Keywords: Probabilistic PCA, isotropic model, dimension reduction, intrinsic dimension, maximum likelihood, asymptotic consistency.

1. Introduction

The analysis of high-dimensional data has become an important problem in statistical learning and dimension reduction has a central place in such settings. Among all existing methods, Principal Component Analysis (PCA) [18] and its probabilistic version (PPCA) [32, 33] are two popular techniques. A central issue in dimension reduction is choosing a sensible number of dimensions to be retained. We refer to [10] for a review on this topic. Two kind of approaches have been proposed in the last decades for intrinsic dimension estimation.

Local methods. The local approach estimates the topological dimension (defined as the basis dimension of the tangent space of the data manifold) from

the information contained in sample neighborhoods. Fukanaga-Olsen's algorithm [17] consists of estimating the rank of the variance matrix computed locally on a Voronoi tessellation. In [9], the Voronoi tessellation is computed thanks to a topology representing network. The algorithms proposed by Pettis *et al.* [25] and Verver-Duin [35] are based on the analysis of the distances from one point to its nearest neighbors. The main limitation of local approaches is their sensitivity to outliers.

Global methods. The global approach consists of unfolding the whole dataset into a linear subspace. The estimated intrinsic dimension is then the dimension of the resulting subspace. Such methods can be divided into three subfamilies.

- *Projection methods*: The lower dimensional subspace can be estimated by minimizing some projection errors. Examples of such approaches include PCA [18] sometimes associated with Cattell's scree test [12] and its non linear extensions based either on auto-associative models [19, 13] or Mercer kernels [29]. Multidimensional scaling type algorithms aim at finding the projection which (locally) preserve the distances among data. Recent methods include LLE [28] and ISOMAP [31].

- *Fractal-based methods*: These techniques rely on the assumption that the dataset is generated by a dynamic system. Their goal is to estimate the dimension of the attractor associated to this dynamic system. For instance, [20] addresses this problem through the estimation of the box-counting dimension and some heuristic methods are introduced in [11]. Most of these methods are designed for low-dimensional datasets since their complexity grows exponentially with the dimension.

- *Model-based methods*: The use of a parametric model permits to derive a maximum likelihood (ML) estimator of the intrinsic dimension. For instance, in [21], the number of points in a small sphere is modeled by a Poisson process. We also refer to [22] for a bias correction of the previous ML estimator. In a similar spirit, [15] uses a polynomial regression based on a uniformity assumption. Several methods are based on a Bayesian approach: Minka [23] proposes a direct calculation of the Laplace approximation of the marginal likelihood while the Bayesian Information Criterion (BIC) [30] is an asymptotic approximation of it. In [16], a regularized BIC is introduced where the likelihood is evaluated at the maximum a posteriori estimator instead of the maximum likelihood estimator. This criterion is used in [24] to select the dimensionality in PPCA with covariates. We also refer to [5, 14, 26] for alternative approximations of the evidence. The underlying idea is that the likelihood is an increasing function of the complexity and thus of the dimensionality as well. This remark motivated the authors to use penalized likelihood criteria.

In this paper, a constrained version of PPCA, called isotropic PPCA, is considered. This model could appear as a restrictive model but it can be useful in specific situations. In particular, it has been proved to be efficient for classification problems in high dimension [7] where parsimonious models are desirable. This paper demonstrates the surprising result that the maximum likelihood criterion is asymptotically optimal in the case of the isotropic PPCA model, the complexity of the model being not an increasing function of the dimensionality. The ML criterion is compared in different situations on simulated and real data to two classical model selection criteria, AIC [1] and BIC [30], to the empirical scree-test of Cattell [12], and to the model-based methods [15], [21], and [23].

This paper is organized as follows. Section 2 introduces an isotropic version of probabilistic PCA and considers the estimation of its parameters. Section 3 focuses on the intrinsic dimension estimation and demonstrates that the maximum likelihood method can be used for this task in the context of the isotropic PPCA model. Section 4 illustrates on simulations and real datasets the behavior of the proposed approach in different situations and Section 5 gives some concluding remarks.

2. Isotropic Probabilistic PCA

In this section, after having recalled the Probabilistic PCA (PPCA) model, it is reformulated using an eigenvalue decomposition. An isotropic version of PPCA is then introduced and inference aspects are addressed.

2.1. Factor Analysis, Probabilistic PCA and Extreme Component Analysis

The Factor Analysis model [3, 4] links linearly a p-dimensional random vector y to a d-dimensional Gaussian vector x of latent variables:

$$y = Hx + \mu + \varepsilon.$$

The $p \times d$ factor matrix H relates the two random vectors and $\mu \in \mathbb{R}^p$ is a fixed location parameter. When $d < p$, the latent vector x provides a parsimonious representation of y. In this context, d is interpreted as the intrinsic dimension of y and is thus the parameter of interest in this study. Without loss of generality, it can be assumed that $x \sim \mathcal{N}(0, I_d)$. If, moreover, the noise ε is supposed to be Gaussian $\varepsilon \sim \mathcal{N}(0, \Psi)$, where Ψ is a $p \times p$ variance matrix, and independent from x, then we end up with a Gaussian distribution for the observations y, *i.e.* $y \sim \mathcal{N}(\mu, \Sigma)$ where:

$$\Sigma = HH^t + \Psi. \tag{1}$$

In such a case, the model parameters can be estimated by maximum likelihood even though an iterative procedure is involved. To overcome this practical difficulty, one can assume an isotropic noise $\Psi = bI_p$ with $b > 0$.

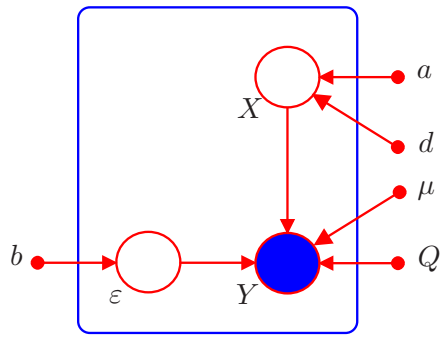


Figure 1: Graphical representation of the isotropic PPCA model.

This model is referred to as the Probabilistic PCA model [33] or to as the Sensible PCA model [27]. The variance matrix of y can be also simplified as:

$$\Sigma = HH^t + bI_p.$$

In contrast to the general Factor Analysis model, all parameters μ, b and H benefit from closed form estimators. It is assume, without loss of generality, that the columns $h_1, ..., h_d$ of H are orthogonal, *i.e.* H^tH is diagonal and $h_1, \ldots, h_d$ are eigenvectors of Σ associated to the eigenvalues $\|h_1\|^2 + b, \ldots, \|h_d\|^2 + b$. Consequently, the d eigenvalues associated to the latent subspace are always larger than the eigenvalue b (with multiplicity $p - d$) associated to the noise subspace. In contrast, in the Probabilistic Minor Component Analysis (PMCA) [37] method, the converse assumption is made. Finally, the two approaches are unified in the Extreme Component Analysis (XCA) method [36] where the noise ε is supposed to be orthogonal to the columns of H. This assumption yields $\Psi = b(I - H(H^tH)^{-1}H^t)$ in (1) and thus the eigenvalues of Σ are $\|h_1\|^2, \ldots, \|h_d\|^2$ and b. Since no assumption is made on their relative magnitudes, PPCA and PMCA may be interpreted as particular cases of XCA.

2.2. Isotropic Probabilistic PCA

Similarly, it may be of interest in specific contexts, such as high-dimensional classification, to consider an isotropic factor matrix. In this case, the matrix H can be rewritten as $H = \sqrt{a-b}V$ with $a > b$ and where V is a $p \times d$ matrix such that $V^tV = I_d$. Thus, the variance matrix of the observation y is given by:

$$\Sigma = (a - b)VV^t + bI_p.$$

Let U be a $p \times (p - d)$ matrix such that $Q := [V, U]$ is an orthogonal $p \times p$ matrix containing p eigenvectors of Σ. Introducing $\Delta = Q^t\Delta Q$ the diagonal

matrix of eigenvalues, an alternative, and more intuitive, parametrization of Σ is

$$\Sigma = Q\Delta Q^t.$$

Moreover, the matrix Δ associated with the isotropic PPCA model has the following form:

$$\Delta = \begin{array}{c} \left(\begin{array}{c|c} \begin{matrix} a & & 0 \\ & \ddots & \\ 0 & & a \end{matrix} & \mathbf{0} \\ \hline \mathbf{0} & \begin{matrix} b & & 0 \\ & \ddots & \\ & & \ddots \\ 0 & & b \end{matrix} \end{array}\right) \end{array} \begin{array}{l} \left.\vphantom{\begin{matrix} a \\ \ddots \\ a \end{matrix}}\right\} d \\ \left.\vphantom{\begin{matrix} b \\ \ddots \\ \ddots \\ b \end{matrix}}\right\} (p-d) \end{array}$$

with $a > b$. Let us emphasize that, since H is supposed to have only two different eigenvalues, the assumption $a > b$ is made without loss of generality and thus this model can also be interpreted as an isotropic XCA model.

The isotropic PPCA model is parametrized by μ, Q, a, b and d. A graphical representation of the isotropic PPCA model is given by Figure 1. As it can be observed on Figure 2 which illustrates the model in a 3-dimensional space, such a model assumes that the distribution is spherical and modelled by a within the d-dimensional latent subspace where the data actually live. The d-dimensional latent subspace is spanned by the d first columns of Q which control the orientation of the subspace whereas μ locates the subspace in the original space. The isotropic PPCA model supposes as well that the variance of noise can be modelled outside the latent subspace with a unique parameter b. Finally, it should also be noted that the mixture model introduced in [8] is a mixture of isotropic PPCA applied to discriminant analysis, *i.e.* each class is modelled by a specific isotropic PPCA model.

2.3. Inference for isotropic PPCA

Before focusing on the estimation of the intrinsic dimension d, the inference on model parameters for the isotropic PPCA model is considered. In the case of the isotropic PPCA model, the parameters to be estimated are μ, a, b, U and V. As in the classical Gaussian framework, the maximum likelihood strategy is retained for parameter estimation. Denoting by n the number of observations, the log-likelihood associated with the isotropic PPCA model is:

$$-\frac{2}{n}\log(L) = d\log(a) + (p-d)\log(b) + \frac{1}{a}\sum_{j=1}^{d} v_j^t W v_j + \frac{1}{b}\sum_{j=1}^{p-d} u_j^t W u_j, \quad (2)$$

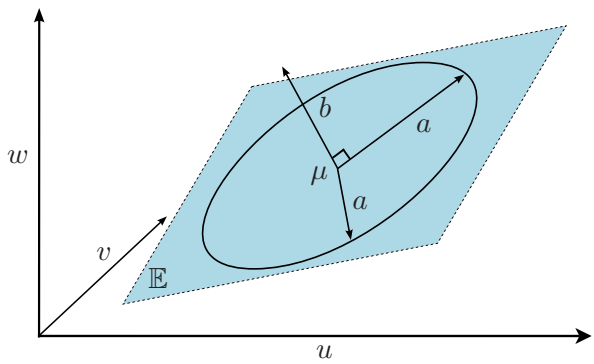


Figure 2: The isotropic PPCA model: a controls the variance in the latent subspace $\mathbb{E}$ spanned by the d first columns of Q, μ locates the subspace in the original space and b controls the variance outside $\mathbb{E}$.

where W is the empirical variance matrix:

$$W = \frac{1}{n}\sum_{\ell=1}^{n}(x_\ell - \hat{\mu})(x_\ell - \hat{\mu})^t,\ \hat{\mu} = \frac{1}{n}\sum_{\ell=1}^{n} x_\ell.$$

The estimation of the matrices U and V is similar to the estimation of H in the context of the actual PPCA model (see [32] for further details). For a given value of d, the ML estimator of the transformation matrix V is the matrix containing the eigenvectors associated with the d largest eigenvalues of the empirical variance matrix W. Similarly, the ML estimator of U is the matrix containing the eigenvectors associated with the $p-d$ smallest eigenvalues of W. Using this eigenvalue decomposition of W in (2), we obtain

$$-\frac{2}{n}\log(L) = d\log(a) + (p-d)\log(b) + \frac{1}{a}\sum_{j=1}^{d}\lambda_j + \frac{1}{b}\sum_{j=d+1}^{p}\lambda_j, \qquad (3)$$

where λ_j is the jth eigenvalue of W. It follows that

$$\hat{a} = \frac{1}{d}\sum_{j=1}^{d}\lambda_j \text{ and } \hat{b} = \frac{1}{(p-d)}\sum_{j=d+1}^{p}\lambda_j.$$

As one can observe, ML estimates of parameters a and b are respectively the means of the largest and smallest eigenvalues of the empirical variance matrix. From a numerical point of view, such estimates should be more stable than eigenvalue estimates when the number of observations is small compared to the data dimension p. Furthermore, it is not necessary in practice to compute the $(p-d)$ smallest eigenvalues of W since $\hat{b}$ can be computed as $\hat{b} = (\mathrm{tr}(W) - da)/(p-d)$.

3. Estimation of the intrinsic dimension by maximum likelihood

In this section, we focus on the estimation of the intrinsic dimension d^*. First the following proposition is proved.

Proposition: *The maximum likelihood of the actual intrinsic dimension d^* is asymptotically unique and consistent in the case of the isotropic PPCA model.*

Proof: Since d is an integer parameter, it is possible to compute the likelihood for each value of $d = 1, ..., p-1$ and to select the value associated to the largest likelihood. From Equation (3), the maximized log-likelihood of the isotropic PPCA model can be written at the optimum $\hat{\theta} = (\hat{\mu}, \hat{a}, \hat{b}, \hat{U}, \hat{V})$ as

$$-\frac{2}{n}\log(L(\hat{\theta}, d)) = d\log(\hat{a}) + (p-d)\log(\hat{b}) + \frac{\mathrm{tr}(W)}{\hat{b}} + \left(\frac{1}{\hat{a}} - \frac{1}{\hat{b}}\right)\sum_{j=1}^{d}\lambda_j.$$

Since $\hat{a} = \frac{1}{d}\sum_{j=1}^{d}\lambda_j$, $\hat{b} = \frac{1}{(p-d)}\sum_{j=d+1}^{p}\lambda_j$ and $\mathrm{tr}(W) = \sum_{j=1}^{p}\lambda_j$, the log-likelihood reduces to:

$$-\frac{2}{n}\log(L(\hat{\theta}, d)) = d\log(\hat{a}) + (p-d)\log(\hat{b}) + p.$$

Consequently, the maximization of the likelihood is equivalent to the minimization of $\phi_n(d) = d\log(\hat{a}) + (p-d)\log(\hat{b})$. Asymptotically, as the number of (independent) observations n tends to infinity, W converges almost surely (a.s.) to Σ. As a consequence of Lemma 2.1 [34], which holds for all symmetric matrix, $\hat{\lambda}_j$ converges a.s. to a if $j \leq d^*$ and $\hat{\lambda}_j$ converges a.s. to b if $j > d^*$. Two cases can arise.

Situation $d \leq d^*$. In this case, $\hat{a} \to a$ and $\hat{b} \to \frac{1}{p-d}\left[(d^*-d)a + (p-d^*)b\right]$ a.s. when $n \to \infty$. Consequently, $\phi_n(d) \to \phi(d)$ a.s. where

$$\phi(d) = d\log(a) + (p-d)\log\left(\frac{(d^*-d)}{(p-d)}a + \frac{(p-d^*)}{(p-d)}b\right),$$

or equivalently

$$\frac{\phi(d) - p\log(a)}{p-d^*} = \frac{(p-d)}{(p-d^*)}\log\left(1 + \frac{(p-d^*)}{(p-d)}\left(\frac{b}{a} - 1\right)\right) = \delta\log\left(1 + \frac{\gamma}{\delta}\right),$$

with $\delta = \frac{(p-d)}{(p-d^*)}$ and $\gamma = \frac{b}{a} - 1$. Thus, the study of $\phi(d)$ reduces to the study of $\psi(\delta) = \delta\log\left(1 + \frac{\gamma}{\delta}\right)$ where $\delta \geq 1$ and $\gamma \leq 0$. Since ψ is a strictly increasing function on $[1, +\infty)$ for all $\gamma \leq 0$, its minimum is reached for $\delta = 1$ and therefore the minimum of ϕ on $[1, d^*]$ is reached for $d = d^*$.

Situation $d \geq d^*$. Here, $\hat{a} \to \frac{1}{d}(d^*a + (d-d^*)b)$ and $\hat{b} \to b$ a.s. when $n \to \infty$. It leads to $\phi_n(d) \to \phi(d)$ a.s. where

$$\phi(d) = d \log\left(\frac{d^*}{d}a + \frac{d-d^*}{d}b\right) + (p-d)\log(b).$$

Similarly to the first situation, we can write

$$\frac{\phi(d) - p\log(b)}{d^*} = \frac{d}{d^*}\log\left(1 + \frac{d^*}{d}\left(\frac{a}{b} - 1\right)\right) = \delta \log\left(1 + \frac{\gamma}{\delta}\right),$$

with $\delta = \frac{d}{d^*}$ and $\gamma = \frac{a}{b} - 1$. Again, the study of $\phi(d)$ reduces to the study of $\psi(\delta) = \delta \log\left(1 + \frac{\gamma}{\delta}\right)$ where $\delta \geq 1$ and $\gamma \geq 0$. Remarking that ψ is a strictly increasing function on $[1, +\infty)$ for all $\gamma \geq 0$, its minimum is reached for $\delta = 1$ and therefore the minimum of ϕ on $[d^*, p]$ is reached for $d = d^*$. As a conclusion, we have proved that the likelihood associated with the model has asymptotically a unique maximum for the actual intrinsic dimension d^* of the data. □

From this proposition, it is deduced that the maximum likelihood criterion can be used to estimate d^* in the context of the isotropic PPCA model. Usually, as for instance for the general PPCA model, model selection criteria using the maximum likelihood need an additional penalty term because the maximum likelihood of a model is asymptotically a non decreasing function of the number of model parameters. However, for isotropic PPCA, the proposition states that the likelihood is asymptotically maximum for the intrinsic dimension d^* of the data. Therefore ML criterion is a good candidate to estimate the intrinsic dimension of a dataset in the isotropic PPCA framework. The reason why ML can be used to estimate d^* for the isotropic PPCA model is the perfect duality between the subspace spanned by the eigenvectors associated with the d largest eigenvalues of W and the supplementary noise subspace with dimension $(p-d)$. Because of the symmetry between a and b occurring in the isotropic PPCA model, the number of parameters to be estimated is the number of model parameters, for fixed d, is $\nu(d) = p + 2 + \min\{d(p-(d+1)/2), (p-d)(p-(p-d+1)/2)\}$. Consequently, the complexity of the isotropic PPCA model does not increase strictly with d: it increases between 1 and the actual intrinsic dimension d^* and decreases between d^* and $(p-1)$. Other criteria of the form ML + pen(n) where pen(n) is a penalty such that pen(n)/n tends to 0 as n tends to infinity are other consistent criteria to estimate this intrinsic dimension. In the following such criteria as AIC and BIC are compared with ML criterion in a numerical experiment.

4. Numerical Experiments

This section presents numerical experiments on simulated and real datasets in order to highlight the main features of different intrinsic dimension estimation methods in the context of the isotropic PPCA model. The maximum

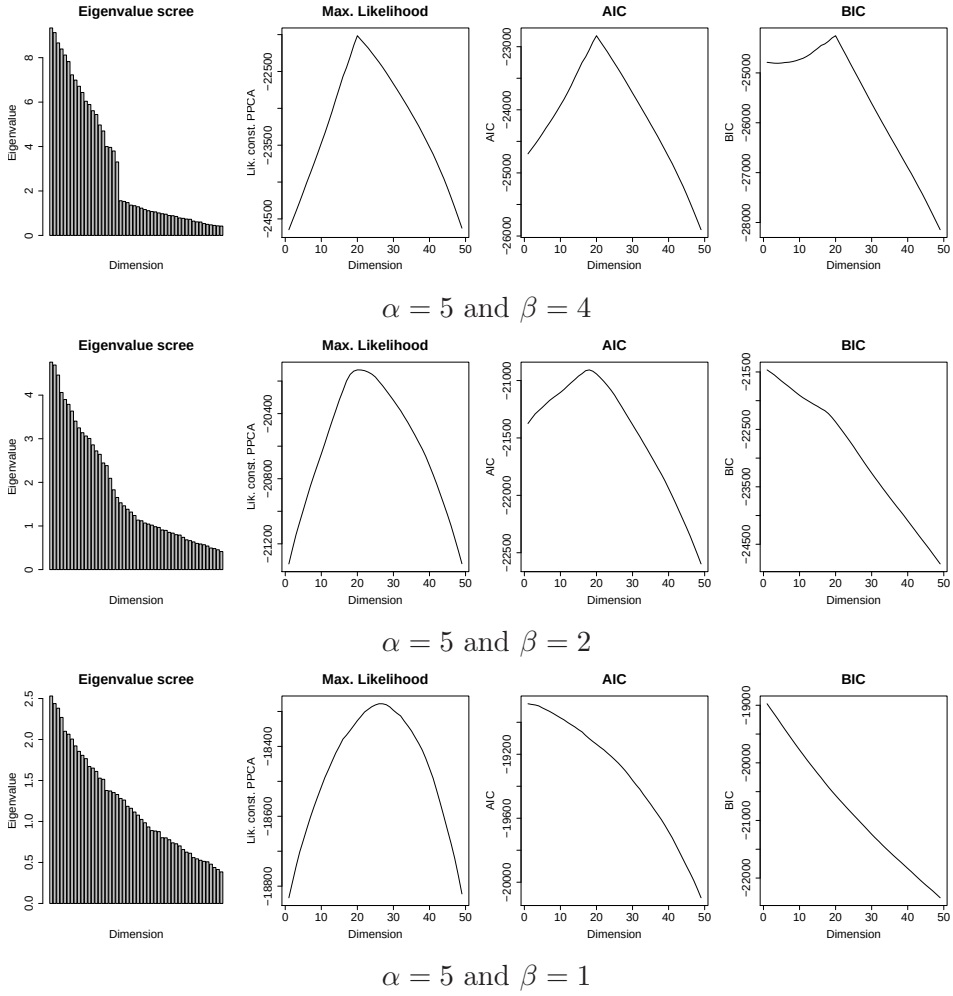


Figure 3: Dimension selection on data simulated according to the isotropic PPCA model with ML, AIC and BIC. The data were simulated with $p = 50$, $b = 1$ and the actual intrinsic dimension is $d^* = 20$.

likelihood criterion, for which we have demonstrated the asymptotic consistency, is compared in the following to two penalized likelihood criteria (AIC and BIC), an empirical criterion (Cattell's scree-test), the MleDim method of [21], the Laplace approach of [23] and the cross-validated likelihood. For the sake of simplicity, the following experiments will be set up according to two parameters: $\alpha = n/p$ and $\beta = d^* a/[(p-d^*)b]$. The parameter α controls the estimation conditions through the ratio between the number of observations and the dimension of the observation space. The second parameter, β, controls the signal to noise ratio through the condition number a/b of the variance matrix. We recall that AIC and BIC respectively penalize the log-likelihood by the quantities $\nu(\mathcal{M})$ and $\nu(\mathcal{M})\log(n)/2$ where $\nu(\mathcal{M})$ is the

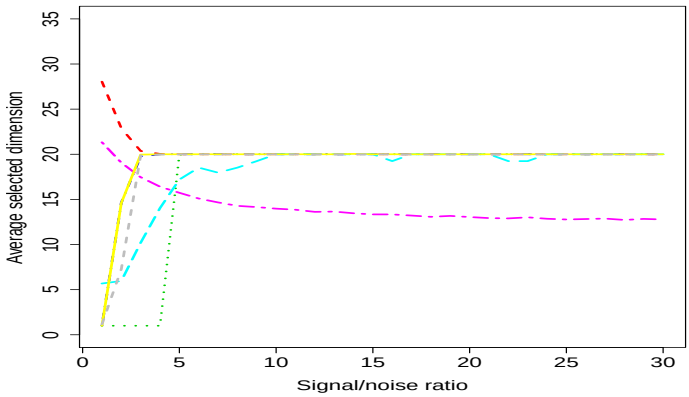
number of independent parameters (complexity) of the used model $\mathcal{M}$. The scree-test of Cattell is an empirical method which compares the differences between consecutive eigenvalues with a fixed threshold for finding a breakdown point in the eigenvalue scree. We refer respectively to [21] and [23] for details on the MleDim and Laplace approaches. Finally, for all the following experiments, the parameters p, b and d^* will remain fixed to the values $p = 50$, $b = 1$ and $d^* = 20$.

4.1. An introductory example

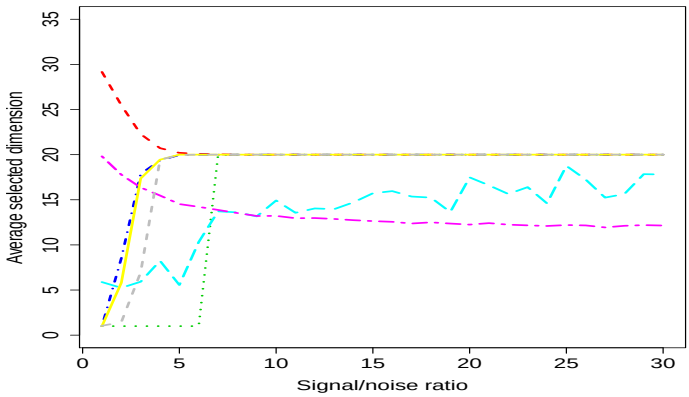
The first experiment aims to show the behavior of the three likelihood-based criteria (ML, AIC and BIC) according to the signal to noise ratio β for a fixed value of α. The simulated model for this experiment is the isotropic PPCA model. The parameter α has been set to 5 which means that the estimation conditions are favorable. Figure 3 shows the eigenvalue scree (left panels) and the behavior of the three likelihood-based criteria (from left to right, ML, AIC and BIC) for different values of β. The first row of Figure 3 considers an easy situation where the eigenvalue scree has a clear breakdown point between relevant and irrelevant dimensions and all criteria succeed in finding the correct intrinsic dimension $d^* = 20$. The second row presents a slightly more difficult situation for which ML and AIC still succeed in determining d^* whereas BIC penalizes too much the likelihood and fails in determining d^*. Finally, the last row focuses on a difficult situation where there is no elbow in the eigenvalue scree. In this case, AIC and BIC fail in estimating d^* by proposing $\hat{d} = 1$. Conversely, ML slightly overestimates the actual value of d^* by proposing $\hat{d} = 28$. It should be noticed that, in the dimension reduction framework, slightly overestimating the intrinsic dimension is preferable to underestimating it because the probability of discarding relevant dimensions is lower.

4.2. Influence of the signal to noise ratio

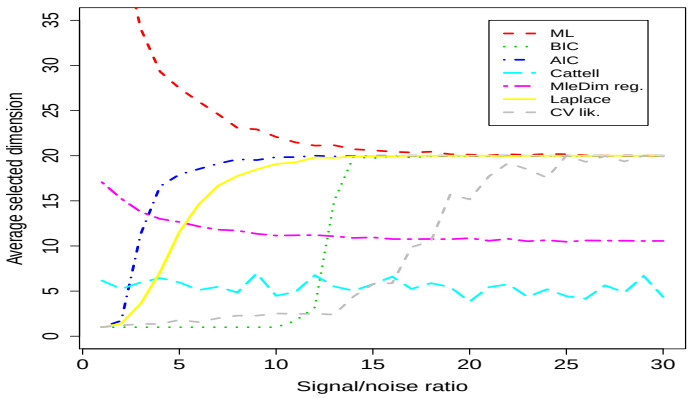
The second experiment focuses on the influence of the signal to noise ratio (parameter β) on the intrinsic dimension estimation with the seven studied dimension selection methods and this for different values of $\alpha = n/p$. In order not to favour the likelihood-based methods, ML, AIC and BIC, the simulated model used in this numerical experiment will not be the isotropic PPCA model but a uniform model with d^* dimensions of variance $a/12$ and $(p - d^*)$ dimensions of variance $b/12$ is used. The results have been averaged from 50 independent simulated datasets. Figure 4 (a)–(c) aims to compare the behavior of criteria based on the isotropic PPCA model with state of the art criteria according to β for three values of $\alpha = 1, 2, 3$. Among the criteria based on the isotropic PPCA, ML, AIC, BIC and the 10 fold cross-validated likelihood (CV Lik.) are studied. The scree-test of Cattell, the MleDim criteria [21] and the Laplace approximation of the integrated likelihood [23] stand for the state of the art criteria. From this figure, it appears that, for



(a) $\alpha = 3$ and $\beta = 1, ..., 30$



(b) $\alpha = 2$ and $\beta = 1, ..., 30$



(c) $\alpha = 1$ and $\beta = 1, ..., 30$

Figure 4: Average selected dimension according to the signal to noise ratio β for fixed values of α. The data were simulated according to a uniform distribution (see text for details) with $p = 50$ and the actual intrinsic dimension is $d^* = 20$.

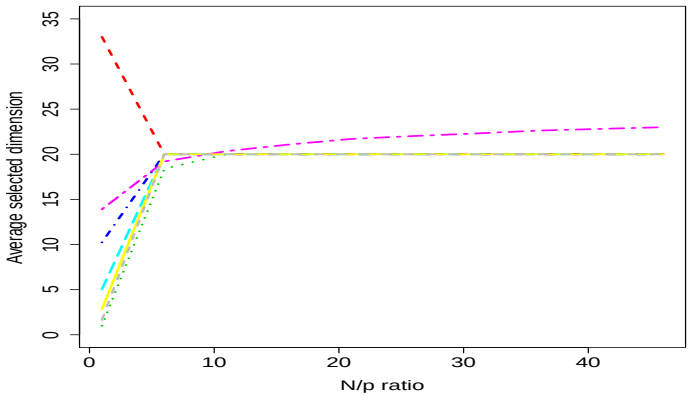
large values of α and β, all the methods perform rather well except MleDim. When α decreases to 1, AIC, CV Lik., Laplace and ML are clearly more efficient than the other criteria. ML could be recommended since it slightly overestimate d^* while the other ones underestimate it. We also experimented the adaption of the Laplace criterion to the isotropic PPCA model. As expected, it behaves slightly better than BIC but remains outperformed by AIC and ML.

4.3. Influence of the n/p ratio

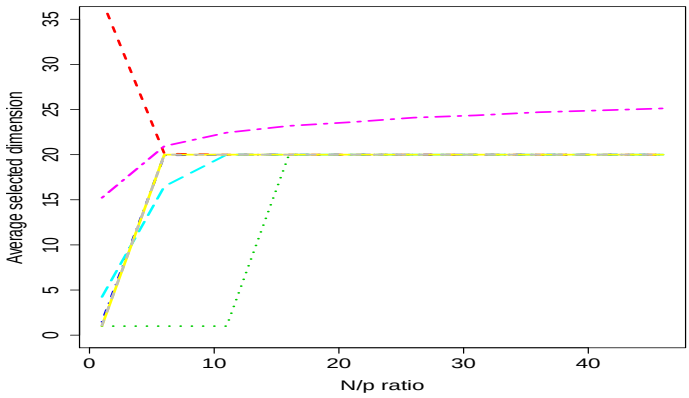
We now focus on the influence of the n/p ratio (parameter α) on the intrinsic dimension estimation with the seven studied criteria for different values of the signal to noise parameter β. Again, the results have been averaged from 50 independent simulated datasets. Figure 5 shows the behavior of the criteria ML, AIC, BIC, Cattell, MleDim, Laplace and CV Lik. according to α for three values of $\beta = 1, 2, 3$. When there is a clear breakdown point in the eigenvalue scree (Figure 5 (a)–(b)), all criteria are efficient for a large range of α values except, as previously, MleDim. It should be however noticed that the scree-test of Cattell and BIC appear again to be less efficient than the other criteria. When there is no clear breakdown point in the eigenvalue scree (Figure 5 (c)), BIC and MleDim fail in estimating d^* whatever the α value is. Cattell, AIC, Laplace and CV Lik. tend to underestimate d^* whereas ML only slightly overestimates it. This study demonstrates as well that the task of estimating the intrinsic dimension of a dataset is extremely difficult when α and β are both close to 1.

4.4. Application to supervised classification

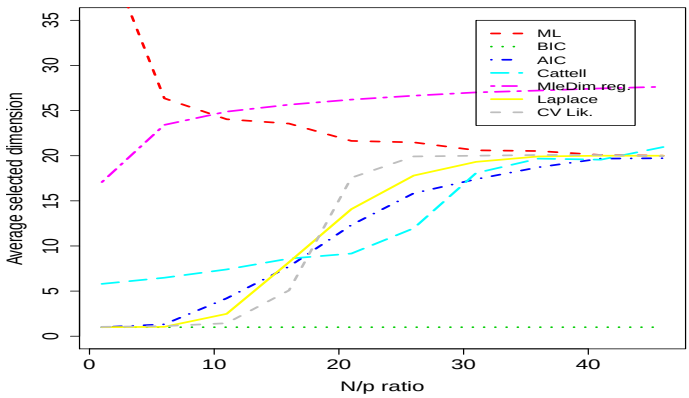
Supervised classification offers the ability to numerically evaluate the performance of the studied methods on real data (for which the actual intrinsic dimension is unknown) through the correct classification rate. We selected six datasets on the UCI Machine Learning Repository (*http://archive.ics.uci.edu/ml/*): Abalone, Glass, Satellite, Sonar, USPS and Wine datasets. We also considered the Maldi dataset [2], coming from mass-spectrometry, which has less observations than measured variables. Indeed, this kind of situations has become a recurrent and challenging scenario in several scientific fields such as, for instance, Biology or Medecine. The USPS dataset has been modified to focus on discriminating the three most difficult classes to be classify, namely the classes of the digits 3, 5 and 8. This dataset has been called USPS 358. The second and the third columns of Table 1 give respectively the number of observations and the number of dimensions of the six datasets. Each dataset was randomly split into a learning set of 90% of the observations and a validation set made of the remaining observations for simulating a difficult classification situation. The intrinsic dimension d was then selected with each dimension selection method before designing the classifier using quadratic discriminant analysis (QDA) on the d first principal



(a) $\beta = 3$ and $\alpha = 1, ..., 50$



(b) $\beta = 2$ and $\alpha = 1, ..., 50$



(c) $\beta = 1$ and $\alpha = 1, ..., 50$

Figure 5: Average selected dimension according to the n/p ratio α for fixed values of β. The data were simulated according to a uniform distribution (see text for details) with $p = 50$ and the actual intrinsic dimension is $d^* = 20$.

Dataset	obs.	dim.	ML	BIC	AIC	IsoLaplace	Laplace
Abalone	4177	8	54.1±1.9 (2)	54.1±1.9 (2)	54.1±1.9 (2)	54.1±1.9 (2)	53.0±2.0 (7)
Glass	214	9	53.2±10.8 (6)	53.2±10.8 (6)	53.2±10.8 (6)	53.2±10.8 (6)	54.1±11.1 (8)
Wine	178	13	98.1±3.1 (5)	96.5±4.0 (3)	96.8±3.9 (4)	96.5±3.9 (3)	98.6±2.6 (9)
Satellite	6435	36	85.1±1.3 (8)	85.1±1.4 (8)	85.1±1.4 (8)	85.1±1.4 (8)	85.7±1.2 (33)
Sonar	208	60	79.0±8.0 (25)	78.6±8.2 (20)	78.9±8.1 (21)	78.5±8.2 (20)	77.8±9.0 (56)
USPS 358	2248	256	98.1±1.0 (81)	98.2±1.0 (57)	98.1±1.0 (73)	98.1± 1.0 (63)	95.6±1.6 (255)
Maldi	112	3268	95.1±6.1 (32)	90.9±10.0 (2)	93.6±8.8 (8)	90.1±10.1 (1)	90.1±10.1 (1)

Table 1: Classification results on real-world datasets: reported values are average correct classification rates computed on validation sets. Standard deviations are also provided as well as average selected dimensions which are given into parentheses.

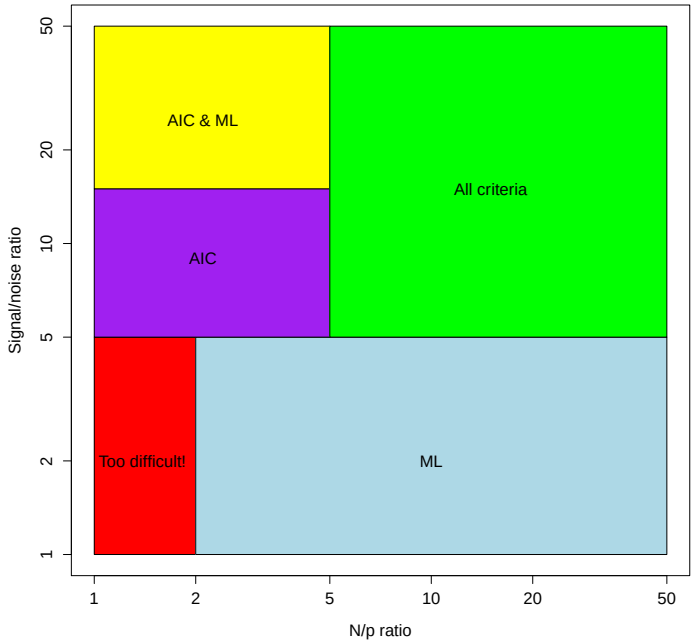


Figure 6: Recommended criteria for intrinsic dimension selection according to the n/p and signal to noise ratios in the context of the isotropic PPCA model.

components (classical PCA). The correct classification rate was computed afterward on the validation set. The results have been averaged from 50 repetitions of the experimental setup. This experimental setup has been applied to the criteria based on the isotropic PPCA model (ML, AIC, BIC and Iso-Laplace) and to the Laplace criterion based on the standard PPCA model. Table 1 reports the average correct classification rate in percentage of each dimension selection method for the seven datasets. The standard deviations are also provided and the average selected dimensions are given into parentheses. The main point is that isotropic criteria always select dramatically less dimensions than the non isotropic model while providing similar or better classification results. Among the criteria based on the isotropic PPCA model, ML is always similarly or more efficient than the others and provides, in addition, more stable results since its variance is usually less than the one of the other criteria. In particular, ML performs better than other criteria on the Maldi dataset for which the N/p ratio is lower than 1.

5. Discussion

The present work focused on the estimation of the intrinsic dimension d which controls in PPCA the number of parameters to be estimated. This

problem can be regarded as a model selection problem. From this point of view, it can be thought of as surprising to propose the maximum likelihood (ML) as a model selection criterion since in most situations this criterion could be expected to increase with the model complexity. Because of the duality between the two subspaces considered with our model, it is not surprising that the ML criterion is able to find the actual intrinsic dimension of the data without requiring an additional complexity penalty.

The theoretical result of Section 3 ensures that the ML criterion is consistent to estimate the actual intrinsic dimension of the isotropic PPCA model. In practice, the sample size n is finite and can be small in regard to p. Thus, it could happen that the sample variability leads to a ML criterion whose maximum is attained for a larger dimension than the actual intrinsic dimension d^*. In such cases and especially for small n, a slight penalty term, as the AIC penalty term, could be desirable to select a proper intrinsic dimension. Figure 6 displays a summary of the recommendations that could be given from our experience. As it can be seen on this figure, it appear that AIC can outperform ML criterion when $n/p < 5$ for a moderate signal to noise ratio. Finally, when $n/p < 2$, no method performs well in selecting d and, in this case, it could be recommended to compare the dimensions selected by AIC and ML and to choose d on an empirical ground when the AIC and ML selected dimensions differ. And, in that purpose, the recommendations provided by Figure 6 could be helpful.

The theoretical result exhibited in this work should have interesting applications in methods related to or based on the isotropic probabilistic PCA model. On the one hand, the intrinsic dimension selection approach proposed in this work could be used to approximately determine the intrinsic dimension in PPCA, PMCA and XCA. Indeed, although the maximum likelihood estimate is not asymptotically consistent for the PPCA, PMCA and XCA models, it could provide a first approximation of the intrinsic dimension for those models. On the other hand, the isotropic PPCA model has been used in [8], in a supervised classification framework, for modelling and classifying the data of K classes in different subspaces with specific intrinsic dimensions d_k^*, $k = 1, ..., K$, estimated through an empirical strategy. In such a context, the BEC criterion [6] could be also used since it is a penalized-likelihood criterion taking into account the classification goal. Alternatively, the asymptotic optimality of the ML criterion for the isotropic PPCA model should allow this classification method to efficiently determine the intrinsic dimension d_k^* of each class using the ML criterion avoiding numerical problems when α or/and β go close to 1. Furthermore, the low and insensitive computational cost of the likelihood-based methods is a great advantage over the other dimension selection methods while incorporated in iterative processes such as the EM algorithm.

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] T. Alexandrov, J. Decker, B. Mertens, A.M. Deelder, R.A. Tollenaar, P. Maass, and H. Thiele. Biomarker discovery in MALDI-TOF serum protein profiles using discrete wavelet transformation. *Bioinformatics*, 25(5):643–649, 2009.

[3] D. J. Bartholomew. *Latent Variable Models and Factor Analysis*. Charles Griffin & Co. Ltd, London, 1987.

[4] A. Basilevsky. *Statistical Factor Analysis and Related Methods*. Wiley, New York, 1994.

[5] C. Bishop. Bayesian PCA. In *11th Annual Conference on Neural Information Processing Systems*, 1999.

[6] G. Bouchard and G. Celeux. Selection of generative models in classification. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 28(4):544–554, 2006.

[7] C. Bouveyron and S. Girard. Robust supervised classification with mixture models: Learning from data with uncertain labels. *Pattern Recognition*, 42(11):2649–2658, 2009.

[8] C. Bouveyron, S. Girard, and C. Schmid. High Dimensional Discriminant Analysis. *Communications in Statistics: Theory and Methods*, 36(14):2607–2623, 2007.

[9] J. Bruske and G. Sommer. Intrinsic dimensionality estimation with optimally topology preserving maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5):572–575, 1998.

[10] F. Camastra. Data dimensionality estimation methods: a survey. *Pattern Recognition*, 36:2945–2954, 2003.

[11] F. Camastra and A. Vinciarelli. Estimating the intrinsic dimension of data with a fractal-based method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(10):1404–1407, 2002.

[12] R. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276, 1966.

[13] B. Chalmond and S. Girard. Nonlinear modeling of scattered multivariate data and its application to shape change. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5):422–432, 1999.

[14] R. Everson and S. Roberts. Inferring the eigenvalues of covariance matrices from limited, noisy data. *IEEE transactions on signal processing*, 48(7):2083–2091, 2000.

[15] M. Fan, H. Qiao, and B. Zhang. Intrinsic dimension estimation of manifolds by incising balls. *Pattern Recognition*, 42(5):780–787, 2009.

[16] C. Fraley and A. Raftery. Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*, 24:155–181, 2007.

[17] K. Fukunaga and D.R. Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on Computers*, 20(2):176–183, 1971.

[18] I. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

[19] J. Karhunen and J. Joutsensalo. Representation and separation of signals using nonlinear PCA type learning. *Neural Networks*, 7(1):113–127, 1994.

[20] B. Kegl. Intrinsic Dimension Estimation Using Packing Numbers. In *15th Annual Conference on Neural Information Processing Systems*, 2002.

[21] E. Levina and P. Bickel. Maximum Likelihood Estimation of Intrinsic Dimension. In *17th Annual Conference on Neural Information Processing Systems*, 2005.

[22] D. MacKay and Z. Ghahramani. Comments on 'maximum likelihood estimation of intrinsic dimension' by E. Levina and P. Bickel, 2005. Technical report. `inference.phy.cam.ac.uk/mackay/dimension`.

[23] T. Minka. Automatic choice of dimensionality for PCA. In *13th Annual Conference on Neural Information Processing Systems*, 2000.

[24] G. Nyamundanda, L. Brennan, and I.C. Gormley. Probabilistic principal component analysis for metabolomic data. *BMC Bioinformatics*, 11:571–581, 2010.

[25] E. Pettis, T. Bailey, A. Jain, and R. Dubes. An intrinsic dimensionality estimator from nearest-neighbor information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1:25–37, 1979.

[26] J.J. Rajan and P.J.W. Rayner. Model order selection for the singular value decomposition and the discrete Karhunen-Loève transform using a Bayesian approach. *IEE proceedings Vision, image and signal processing*, 144(2):116–123, 1997.

[27] S. Roweis. EM algorithms for PCA and SPCA. In *10th Annual Conference on Neural Information Processing Systems*, 1998.

[28] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[29] B. Schölkopf, A. Smola, and K-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[30] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[31] J. Tenenbaum, V. De Silva, and J. Langford. A global geometric framework for non linear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[32] M. Tipping and C. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11(2):443–482, 1999.

[33] M. Tipping and C. Bishop. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society, Series B*, 3(61):611–6222, 1999.

[34] D. Tyler. Asymptotic Inference for Eigenvectors. *Annals of Statistics*, 9(4):725–736, 1981.

[35] P. Verveer and R. Duin. An evaluation of intrinsic dimensionality estimators. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(1):81–86, 1995.

[36] M. Welling, F. Agakov, and C. Williams. Extreme Components Analysis. In *16th Annual Conference on Neural Information Processing Systems*, 2003.

[37] C. Williams and F. Agakov. Products of Gaussians and Probabilistic Minor Component Analysis. *Neural Computation*, 14(5):1169–1182, 2002.