Pattern Recognition Letters 32 (2011) 1706–1713

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

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*

ARTICLE INFO

Article history:
Received 26 October 2010
Available online 22 August 2011
Communicated by M.A. Girolami

Keywords:
Probabilistic PCA
Isotropic model
Dimension reduction
Intrinsic dimension
Maximum likelihood
Asymptotic consistency

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.

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) (Jolliffe, 1986) and its probabilistic version (PPCA) (Tipping and Bishop, 1999a,b) are two popular techniques. A central issue in dimension reduction is choosing a sensible number of dimensions to be retained. We refer to Camastra (2003) 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 (Fukunaga and Olsen, 1971) consists of estimating the rank of the variance matrix computed locally on a Voronoi tessellation. In (Bruske and Sommer, 1998), the Voronoi tessellation is computed thanks to a topology representing network. The algorithms proposed by Pettis et al. (1979) and Verveer and Duin (1995) 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 (Jolliffe, 1986) sometimes associated with Cattell's scree test (Cattell, 1966) and its non-linear extensions based either on auto-associative models (Karhunen and Joutsensalo, 1994; Chalmond and Girard, 1999) or Mercer kernels (Schölkopf et al., 1998). Multidimensional scaling type algorithms aim at finding the projection which (locally) preserve the distances among data. Recent methods include LLE (Roweis and Saul, 2000) and ISOMAP (Tenenbaum et al., 2000).
- *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, Kegl (2002) addresses this problem through the estimation of the box-counting dimension and some heuristic methods are introduced in (Camastra and Vinciarelli, 2002). 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 (Levina and Bickel, 2005), the number of points in a small sphere is modeled by a Poisson process. We also refer to MacKay and Ghahramani (2005) for

[*] Corresponding author.
E-mail address: charles.bouveyron@univ-paris1.fr (C. Bouveyron).

0167-8655/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2011.07.017

a bias correction of the previous ML estimator. In a similar spirit, Fan et al. (2009) uses a polynomial regression based on a uniformity assumption. Several methods are based on a Bayesian approach: Minka (2000) proposes a direct calculation of the Laplace approximation of the marginal likelihood while the Bayesian Information Criterion (BIC) (Schwarz, 1978) is an asymptotic approximation of it. In (Fraley and Raftery, 2007), 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 (Nyamundanda et al., 2010) to select the dimensionality in PPCA with covariates. We also refer to Bishop (1999), Everson and Roberts (2000) and Rajan and Rayner (1997) 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 (Bouveyron and Girard, 2009) 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 (Akaike, 1974) and BIC (Schwarz, 1978), to the empirical scree-test of Cattell (1966), and to the model-based methods (Fan et al., 2009; Levina and Bickel, 2005; Minka, 2000).

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 (Bartholomew, 1987; Basilevsky, 1994) 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 $b > 0$. This model is referred to as the Probabilistic PCA model (Tipping and Bishop, 1999b) or to as the Sensible PCA model (Roweis, 1998). 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 $h_1, \ldots, h_d$ 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) (Williams and Agakov, 2002) method, the converse assumption is made. Finally, the two approaches are unified in the Extreme Component Analysis (XCA) method (Welling et al., 2003) 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:

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

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 Fig. 1. As it can be observed on Fig. 2 which illustrates the model in a 3-dimensional space, such a model assumes that the distribution is spherical and modeled by a within the d-dimensional latent