Kernel principal component analysis from a data modelling point of view

Ivan Markovsky

Abstract

I show an example of how the principal component analysis method can be used for solving a specific computer vision problem—the one of fitting an ellipse to a set of points. In the example, the feature map is *given* (postulated) as part of the problem specification. More generally, the feature map in the kernel principal component analysis has the interpretation of the model class in data modelling problems. This interpretation relates the kernel selection problem in machine learning to the model selection problem in system identification.

Data fitting by a second order model

Consider the second order model

$$\mathscr{B}(A,b,c) := \{\, d \in \mathbb{R}^2 \mid d^\top A d + b^\top d + c = 0 \,\} \subset \mathbb{R}^2, \qquad \text{with} \quad A = A^\top \tag{1}$$

and define a parameter vector

$$\theta := \operatorname{col}(a_{11}, a_{12}, a_{22}, b_1, b_2, c) \in \mathbb{R}^6.$$

Any $\theta \in \mathbb{R}^6$ corresponds to a second order model in $\mathbb{R}^2$, defined by (1), and vice verse, to any second order model in $\mathbb{R}^2$ there is a nonunique parameter vector $\theta \in \mathbb{R}^6$, such that the model is (1).

It can be verified that

$$\mathscr{B}(A,b,c) = \{\, d \in \mathbb{R}^2 \mid \theta^\top \Phi(d) = 0 \,\} =: \mathscr{B}(\theta),$$

where

$$\Phi(d) := \operatorname{col}(d_1 d_1, 2 d_1 d_2, d_2 d_2, d_1, d_2, 1). \tag{2}$$

Note that $\mathscr{B}(\theta) = \mathscr{B}(\alpha\theta)$ for any $\alpha \neq 0$, so that the parameter θ is unique only up to a multiplication by a nonzero constant.

The considered data fitting problem is to find a second order model that best matches a set of given points

$$\mathscr{D} = \{d^{(1)}, \dots, d^{(N)}\} \subset \mathbb{R}^{\mathtt{d}}$$

in the sense of minimization of the fitting criterion

$$\sum_{i=1}^{N} \big(\theta^\top \Phi(d^{(i)})\big)^2. \tag{3}$$

If the data point $d^{(i)}$ is on the surface defined by θ, i.e., if $d^{(i)} \in \mathscr{B}(\theta)$, then $\theta^\top \Phi(d^{(i)}) = 0$ and therefore this point does not have a contribution to the total fitting error. Conversely, if the point is not on the surface, i.e., if $d^{(i)} \notin \mathscr{B}(\theta)$, then $\theta^\top \Phi(d^{(i)}) \neq 0$ and the point has a contribution to the fitting error. Therefore, the fitting error (3) indeed penalizes the deviation of the points from the model $\mathscr{B}(\theta)$.

The data fitting problem for the second order model (1) combined with the fitting criterion (3) gives the following optimization problem

$$\text{minimize} \quad \text{over } \theta \in \mathbb{R}^6 \quad \theta^\top \begin{bmatrix} \Phi(d^{(1)}) & \cdots & \Phi(d^{(N)}) \end{bmatrix} \begin{bmatrix} \Phi(d^{(1)}) & \cdots & \Phi(d^{(N)}) \end{bmatrix}^\top \theta \quad \text{subject to} \quad \|\theta\| = 1. \tag{4}$$

The constraint $\|\theta\| = 1$ is imposed in order to be avoided the trivial solution $\theta = 0$ and involves no loss of generality because θ is nonunique and can be scaled arbitrarily. The solution of (4) is given by the left singular vector of the matrix $\begin{bmatrix} \Phi(d^{(1)}) & \cdots & \Phi(d^{(N)}) \end{bmatrix}$, corresponding to the smallest singular value.

Comments

1. The fact that the solution of the data fitting problem (4) is given by the singular value decomposition of the matrix

$$\begin{bmatrix}\Phi(d^{(1)}) & \cdots & \Phi(d^{(N)})\end{bmatrix}$$

makes a link with the principal component analysis (PCA) of the transformed data points

$$\Phi(d^{(1)}),\ldots,\Phi(d^{(N)}).$$

The optimal solution θ^* turns out to be the *last* principal component of the transformed data. In terms of the original data $d^{(1)},\ldots,d^{(N)}$, the solution of the data fitting problem is given by the kernel PCA with a feature mapping (2).

2. The described data fitting method (or application of the kernel PCA) is known in the computer vision literature as the *algebraic ellipsoid fitting method* and is known to give poor fits. An improved method is described in [4].

3. In the data fitting problem, the feature map Φ is derived from the model (1), so that it is *given* in the problem formulation. In other words, the choice of the model class (in the example, the set of all second order models) corresponds to the choice of the feature map in the kernel PCA.

4. With the interpretation of the feature map as a model, the choice of the feature map (or the kernel) can be related to the model selection problem in system identification [7, 3]. The model selection problem is an important problem with extensive literature, see, e.g., [8]. Perhaps the most popular methods used for model selection in system identification are regularization [2], cross validation [6], information criterion [1], and the minimum description length [5]. These methods are therefore applicable to the kernel selection problem.

5. Comment 4 opens many specific research questions. The high level goal of research in this direction would be to clarify:

   - To what models in system identification correspond the popular choices of kernels in machine learning and conversely, what are the kernels corresponding to the popular model classes in system identification?
   - What does system identification have to offer to the machine learning community in terms of model selection methods, and conversely, are there machine learning specific kernel selection techniques that can provide new solutions to model selection problems in system identification?

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Automat. Control*, 6:716–723, 1974.

[2] P.-C. Hansen. *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*. SIAM, 1997.

[3] L. Ljung. *System Identification: Theory for the User*. Prentice-Hall, Upper Saddle River, NJ, 1999.

[4] I. Markovsky, A. Kukush, and S. Van Huffel. Consistent least squares fitting of ellipsoids. *Numerische Mathematik*, 98(1):177–194, 2004.

[5] J. Rissanen. Modeling by the shortest data description. *Automatica*, 14:465–471, 1978.

[6] D. Sima. *Regularization techniques in model fitting and parameter estimation*. PhD thesis, Faculty of Engineering, K.U.Leuven, 2006.

[7] T. Söderström and P. Stoica. *System Identification*. Prentice Hall, 1989.

[8] P. Stoica and Y. Selén. Model-order selection: A review of information criterion rules. *IEEE Signal Processing Magazine*, 21:36–47, 2004.