Extensions of probabilistic PCA

Jaakko Luttinen, Alexander Ilin, Juha Karhunen, and Tapani Raiko

PCA of large-scale datasets with many missing values

Principal component analysis (PCA) is a classical data analysis technique. Some algorithms for PCA scale better than others to problems with high dimensionality. They also differ in the ability to handle missing values in the data. In our recent paper [1], a case is studied where the data are high-dimensional and a majority of the values are missing. In the case of very sparse data, overfitting becomes a severe problem even in simple linear models such as PCA. Regularization can be provided using the Bayesian approach by introducing prior for the model parameters. The PCA model can then be identified using, for example, maximum a posteriori estimates (MAPPCA) or variational Bayesian (VBPCA) learning.

In [1], we study different approaches to PCA for incomplete data. We show that faster convergence can be achieved using the following rule for the model parameters:

$$\theta_i \leftarrow \theta_i - \gamma \left(\frac{\partial^2 C}{\partial \theta_i^2} \right)^{-\alpha} \frac{\partial C}{\partial \theta_i},$$

where α is a control parameter that allows the learning algorithm to vary from the standard gradient descent ($\alpha = 0$) to the diagonal Newton's method ($\alpha = 1$). These learning rules can be used for standard PCA learning and extended to MAPPCA and VBPCA.

The algorithms were tested on the Netflix problem (`http://www.netflixprize.com/`), which is a task of predicting preferences (or producing personal recommendations) by using other people's preferences. The Netflix problem consists of movie ratings given by 480189 customers to 17770 movies. There are 100480507 ratings from 1 to 5 given, and the task is to predict 2817131 other ratings among the same group of customers and movies. 1408395 of the ratings are reserved for validation. Thus, 98.8% of the values are missing.

We used different variants of PCA in order to predict the test ratings in the Netflix data set. The obtained results are shown in Figure 1. The best accuracy was obtained using VB PCA with a simplified form of the posterior approximation (VBPCAd in Figure 1). That method was also able to provide reasonable estimates of the uncertainties of the predictions.

Robust PCA for incomplete data

Standard PCA is known to be sensitive to outliers in the data because it is based on minimisation of a quadratic criterion such as the mean-square representation error. Thus, corrupted or atypical observations may cause the failure of PCA, especially for data sets with missing values. A standard way to cope with this problem is replacing the quadratic cost function of PCA a function which grows more slowly.

In [2], we present a new robust PCA model based on the Student-t distribution and show how it can be identified for data sets with missing values. We make the assumption that the outliers can arise independently in each sensor (i.e. for each dimension of a data vector). This assumption is different to the previously introduced techniques [3] and it turns out to be important for

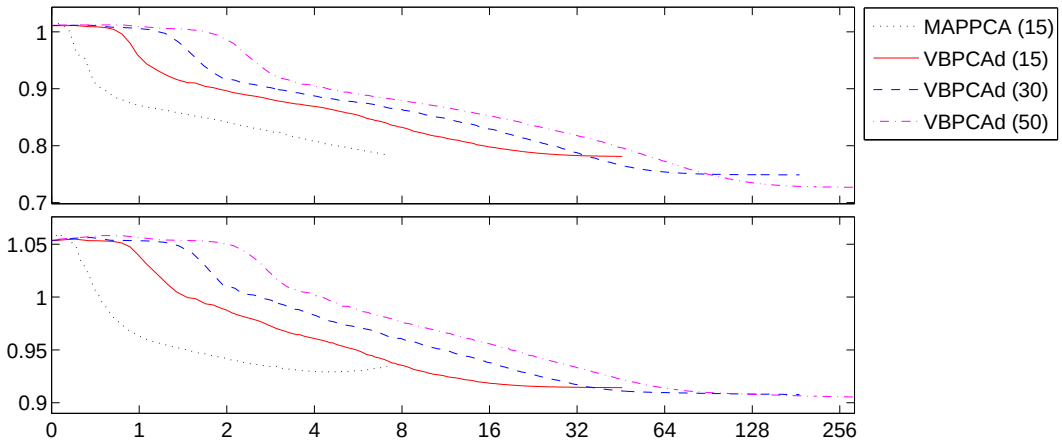

Figure 1: Root mean squared errors for the Netflix data (y-axis) plotted against the processor time in hours. The upper plot shows the training error while the lower plot shows the error for the probing data provided by Netflix. The time scale is linear from 0 to 1 and logarithmic above 1.

modeling incomplete data sets. The proposed model can improve the quality of the principal subspace estimation and provide better reconstructions of missing values. The model can also be used to remove outliers by estimating the true values of their corrupted components from the uncorrupted ones.

We tested the robust PCA model on the Helsinki Testbed data set which at the moment of our studies contained many atypical measurements and missing values. The model was used to estimate four principal components of the temperature measurements from 79 stations in Southern Finland. Figure 2 presents the reconstruction of the data using our robust PCA model for four different stations. The reconstructions look very reasonable with most of the outliers being removed.

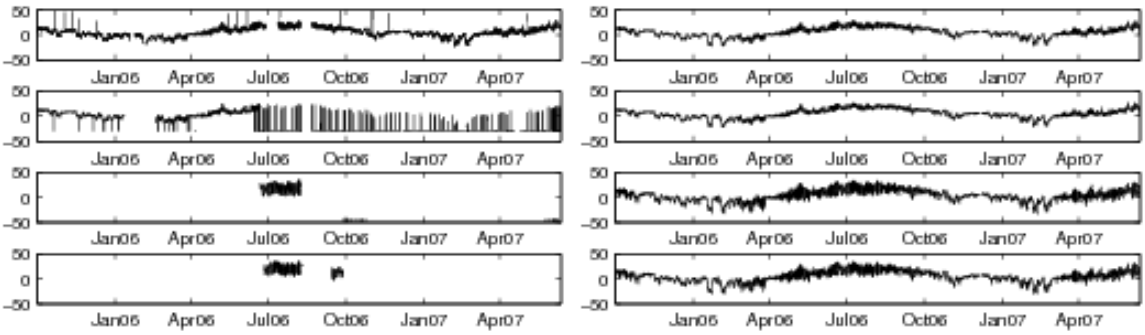

Figure 2: Four example signals from the Helsinki Testbed dataset and their reconstructions using the proposed robust PCA.

The results of this study are presented in more detail in the journal manuscript [4]. More traditional methods for robust PCA, also with missing values, have been studied in [5]. They are usually much easier to apply compared with Bayesian methods but less effective.

References

[1] A. Ilin and T. Raiko. Practical approaches to principal component analysis in the presence of missing values. Journal of Machine Learning Research, volume 11, pp. 1957–2000, July 2010.

[2] J. Luttinen, A. Ilin, and Juha Karhunen. Bayesian robust PCA for incomplete data. In *Proc. of the 8th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA 2009)*, pp. 66–73, Paraty, Brazil, March 2009.

[3] C. Archambeau, N. Delannay, M. Verleysen. Robust probabilistic projections. In *Proc. of the 23rd Int. Conf. on Machine Learning (ICML 2006)*, pp. 33-40, New York, NY, USA, 2006.

[4] J. Luttinen, A. Ilin, and Juha Karhunen. Bayesian robust PCA for incomplete data. To appear in *Neural Processing Letters*, 2012.

[5] J. Karhunen. Robust PCA methods for complete and missing data. *Neural Network World*, 21(5):357–392, 2011.