3.3 Reconstruction of historical climate data by Gaussian-process factor analysis

Alexander Ilin and Jaakko Luttinen

Studying natural variability of climate is a topic of intensive research in climatology. In our earlier research, we have extended the classical technique of rotated Principal Components, or Empirical Orthogonal Functions, by introducing the concept of "interesting structure" for massive sets of spatio-temporal climate measurements. In our case, the goal of exploratory analysis is to find signals with some specific structures of interest. They may for example manifest themselves mostly in specific variables, which exhibit prominent variability in a specific timescale etc. An example of such analysis can be extracting clear trends or quasi-oscillations from climate records. The procedure for obtaining suitable rotations of EOFs can be based on the general algorithmic structure of denoising source separation (DSS) [1].

However, understanding long-term variability of climate faces the problem of the scarcity of climate observations in the past. Thus, reconstruction of historical climate becomes an important problem.

The standard methods of statistical reconstruction are ad hoc adjustments of PCA for incomplete data making such additional assumptions as temporal and spatial smoothness of the observed climate variables. These assumptions were used, for example, in [2] to reconstruct the global sea surface temperatures (SST) in the 1856–1991 period from the MOHSST5 data set (which is largely based on the measurements made from merchant ships). The method presented there uses additional information about the quality of the data and this uncertainty information is derived from the number of different sources which were used to compute each data sample.

In our recent papers [3, 4], we use the Bayesian framework to perform statistical reconstructions of spatio-temporal data. In [3], we adopt the basic variational Bayesian PCA model and use additional uncertainty information to improve the reconstruction performance.

In [4], we present a more advanced probabilistic model called *Gaussian-process factor analysis (GPFA)*. The method is based on standard matrix factorization:

$$\mathbf{Y} = \mathbf{W}\mathbf{X} + \text{noise} = \sum_{d=1}^{D} \mathbf{w}_{:d}\mathbf{x}_{d:}^{\mathrm{T}} + \text{noise},$$

where $\mathbf{Y}$ is a data matrix in which each row contains measurements in one spatial location and each column corresponds to one time instance. Each $\mathbf{x}_{d:}^{T}$ is a row vector representing the time series of one of the D factors, whereas $\mathbf{w}_{:d}$ is a column vector of loadings which are spatially distributed. Matrix $\mathbf{Y}$ may contain missing values and the samples can be unevenly distributed in space and time.

We assume that both factors $\mathbf{x}_{d:}$ and corresponding loadings $\mathbf{w}_{:d}$ have prominent structures that we model using the tool of Gaussian processes [5]. The model is identified in the framework of variational Bayesian learning and high computational cost of GP modeling is reduced by using sparse approximations derived in the variational methodology.

In the experiments reported in [4], we show that GPFA can provide better reconstructions of global SST set compared to variational Bayesian PCA. Figure 3.2 shows the spatial and temporal patterns of the four most dominant principal components found by GPFA from the MOHSST5 data set. The obtained test reconstruction errors were 0.5714 for GPFA and 0.6180 for VBPCA, which can be seen as a significant improvement.

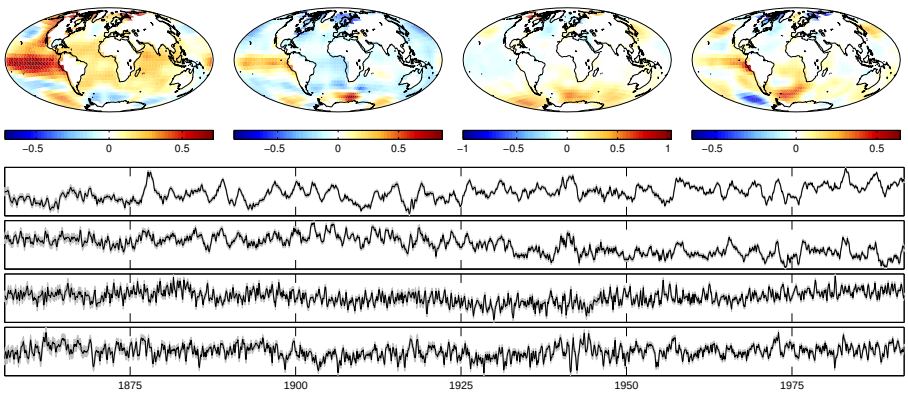

Figure 3.2: The spatial and temporal patterns of the four most dominating principal components estimated by GPFA from the MOHSST5 dataset. The solid lines and gray color in the time series show the mean and two standard deviations of the posterior distribution.

References

[1] J. Särelä and H. Valpola. Denoising source separation. *Journal of Machine Learning Research*, 6:233–272, 2005.

[2] A. Kaplan, M. Cane, Y. Kushnir, M. Blumenthal, B. Rajagopalan. Analysis of global sea surface temperatures 1856–1991. *Journal of Geophysical Research*, 103:18567–18589, 1998.

[3] A. Ilin and A. Kaplan. Bayesian PCA for reconstruction of historical sea surface temperatures. In *Proc. of the IEEE International Joint Conference on Neural Networks (IJCNN 2009)*, pp. 1322–1327, Atlanta, USA, June 2009.

[4] J. Luttinen and A. Ilin. Variational Gaussian-process factor analysis for modeling spatio-temporal data. In *Advances in Neural Information Processing Systems (NIPS) 22*, Vancouver, Canada, Dec. 2009.

[5] C. E. Rasmussen, C. K. I. Williams. *Gaussian processes for machine learning*. MIT Press, 2006.