Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics & Data Analysis 52 (2007) 821–839

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Learning and approximate inference in dynamic hierarchical models

Bart Bakker[a,*], Tom Heskes[b]

[a] High Tech Campus 11, Prof. Holstlaan 4, 5656 AE Eindhoven, The Netherlands
[b] Radboud University Nijmegen, Toernooiveld 1, Room A4026, 6525 ED Nijmegen, The Netherlands

Available online 12 January 2007

Abstract

A new variant of the dynamic hierarchical model (DHM) that describes a large number of parallel time series is presented. The separate series, which may be interdependent, are modeled through dynamic linear models (DLMs). This interdependence is included in the model through the definition of a 'top-level' or 'average' DLM. The model features explicit dependences between the latent states of the parallel DLMs and the states of the average model, and thus the many parallel time series are linked to each other. The combination of dependences *within* each time series and dependences *between* the different DLMs makes the computation time that is required for exact inference cubic in the number of parallel time series, however, which is unacceptable for practical tasks that involve large numbers of parallel time series. Therefore, two methods for fast, approximate inference are proposed: a variational approximation and a factorial approach. Under these approximations, inference can be performed in linear time, and it still features exact means. Learning is implemented through a maximum likelihood (ML) estimation of the model parameters. This estimation is realized through an expectation maximization (EM) algorithm with approximate inference in the E-step. Examples of learning and forecasting on two data sets show that the addition of direct dependences has a 'smoothing' effect on the evolution of the states of the individual time series, and leads to better prediction results. The use of approximate instead of exact inference is further shown not to lead to inferior results on either data set.

Keywords: Time series; Dynamic linear model; Maximum likelihood estimation; Variational approximation; Expectation propagation

1. Introduction

Many real-world tasks can be viewed as parallel time series. Consider for example weather prediction for various parts of the same country, stock price prediction for a portfolio of stocks traded on the same stock exchange, or sales figures for a number of different items sold in the same store. Models that describe such tasks can make use both of the fact that the data has the form of a time series, and therefore may have a specific behavior through time, and of the fact that these tasks are similar to each other, and thus may have a (hidden) inter-dependence.

In this article we propose a new combination of the hierarchical models of Lindley and Smith (1972) and the dynamic linear models (DLMs) of Harrison and Stevens (1976). The proposed model features both a 'top-level' (average) time series and a set of parallel or 'lower-level' time series. Each series is modeled through a DLM (see e.g. Harrison and Stevens, 1976; West and Harrison, 1997). At each time t, the probability of lower-level states $\boldsymbol{\theta}_{i,t}$, corresponding to the

[*] Corresponding author. Tel.: +31 40 2747 857; fax: +31 40 2744 906.
E-mail addresses: bart.bakker@philips.com (B. Bakker), t.heskes@cs.ru.nl (T. Heskes).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.01.001

parallel time series, depends both on the previous state, $\boldsymbol{\theta}_{i,t-1}$ and on the top-level state $\mathbf{M}_{t-1}$. The latter dependence includes the hierarchical model approach of e.g. Lindley and Smith (1972) into the dynamic hierarchical model (DHM) that we present in this article.

A similar combination has been proposed by Gamerman and Migon (1993). This combination models the top-level states through a DLM and the lower-level states are inferred from the top-level states. The latent states of the lower-level time series in this model feature no direct inter-dependences. In this article we add these dependences, which will be shown to have a smoothing effect on the dynamics of the lower-level DLMs, and lead to better predictions of future observations.

The addition of these dependences makes inference infeasible for larger numbers of parallel DLMs. In fact, the computation time that is required for exact inference increases cubically with the number of parallel tasks, as we will show in Sections 2.3 and 4. We therefore present two approximating methods to perform inference on the proposed model. The first approximation makes use of the so-called variational approach (see also Jaakkola and Jordan, 2000): we construct an approximating model which consists of a single, independent DLM for each parallel time series, and one independent top-level DLM. This approximating model is optimized through minimization of its Kullback–Leibler (KL) divergence from the exact model. The second approximation is closely related to a local optimization method introduced by Boyen and Koller (1998), where the approximating model consists of independent probability distributions for each individual state, including the top-level states.

We present our version of the DHM in Section 2. In Sections 3.1 and 3.2 we describe the aforementioned approximate inference methods. We evaluate the proposed techniques in Section 4. This evaluation is based on forecasting results on two databases, one with artificially generated data, and one that concerns single copy newspaper sales. In this evaluation we use an expectation-maximization (EM)-algorithm to estimate the parameters of the forecasting models. The expectation step of the EM-algorithm uses either exact inference, one of the two approximating methods or a DHM as described by Gamerman and Migon (1993), and thus we compare the various approaches. Section 5 concludes the article with an overview of related work, a summary of the contribution described in this article and an outlook on future work.

2. A hierarchical time series model

2.1. The extended model

We consider a collection of n parallel time series indexed by i, each characterized by T combinations of a covariate vector $\mathbf{x}_{i,t}$ (of dimension d) and a response $y_{i,t}$. The response is modeled as a linear function of corresponding covariates $\mathbf{x}_{i,t}$ with additional Gaussian noise $\varepsilon_{i,t}$:

$$y_{i,t} = \boldsymbol{\theta}_{i,t}^T \mathbf{x}_{i,t} + \varepsilon_{i,t}, \tag{1}$$

where we assume all noise terms $\varepsilon_{i,t}$ to be independent of each other and normally distributed around zero, with variance σ^2. $\boldsymbol{\theta}_{i,t}$ is the (time-dependent) regression parameter, which plays the role of a dynamic latent variable. The prediction for each new state is a weighted average of the old state, propagated through the evolution matrix A, and the corresponding top-level state:

$$\boldsymbol{\theta}_{i,t} = A\boldsymbol{\theta}_{i,t-1} + B\mathbf{M}_{t-1} + \boldsymbol{\xi}_{i,t}, \tag{2}$$

where the noise $\boldsymbol{\xi}_{i,t}$ is assumed to be normally distributed around zero with variance Σ^2, and B is the weight (matrix) for the top-level state. The top-level state thus couples the dynamics of the lower-level DLMs. The evolution of the top-level states obeys

$$\mathbf{M}_t = G\mathbf{M}_{t-1} + \gamma_t, \tag{3}$$

where the noise γ_t is also assumed to be normally distributed around zero, with variance Σ_M^2. Initial conditions are

$$\mathbf{M}_1 \sim \mathcal{N}(\hat{\mathbf{M}}_1, \Sigma_{M_1}^2) \quad \text{and} \quad \boldsymbol{\theta}_{i,1} \sim \mathcal{N}(\hat{\boldsymbol{\theta}}_1, \hat{\Sigma}_1^2). \tag{4}$$

We choose $B = \mathbb{1} - A$ for the implementations in this article. Other values for B are possible, but they make the approximate inference methods that are described in this article more complex, and lead to increased computation times. We have observed no improvement in performance when B is left free to choose.