Learning and Approximate Inference in Dynamic Hierarchical Models

Bart Bakker *

High Tech Campus, Prof. Holstlaan 4, 5656 AA Eindhoven, The Netherlands

Tom Heskes

Radboud University Nijmegen, Toernooiveld 1, Room A4026, 6525 ED Nijmegen, The Netherlands

Abstract

We present a variant of the dynamic hierarchical model (DHM) that describes a large number of parallel time series. The separate series, which may be interdependent, are modelled 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. We therefore propose two methods for fast, approximate inference: 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. We perform learning and forecasting on two data sets to 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. We further demonstrate that the use of approximate instead of exact inference does not lead to inferior results on our data sets.

Key words: time series, dynamic linear model, maximum likelihood estimation, variational approximation, expectation propagation

* Corresponding author.
Email addresses: bart.bakker@philips.com (Bart Bakker), t.heskes@cs.ru.nl (Tom Heskes).

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 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 dynamic linear model (see e.g. (Harrison and Stevens 1976; West and Harrison 1997)). At each time t, the probability of 'lower-level' states $\theta_{i,t}$, corresponding to the parallel time series, depends both on the previous state, $\theta_{i,t-1}$ and on the top-level state $\theta_{0,t}$ at the same time. The latter dependence includes the hierarchical model approach of e.g. (Lindley and Smith 1972) into the dynamic hierarchical model that we present in this article.

A similar combination has been proposed in (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. 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 to the exact model. The second approximation is closely related to a local optimization method introduced in (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 dynamic hierarchical model in Section 2. In Sections 3.1 and 3.2 we describe the aforementioned approximate inference methods. The extension of the model and its approximations to higher-level hierarchical models is discussed in Section 4 and a summary of related work is presented in Section 5. We evaluate the proposed techniques in Section 6. 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 in (Gamerman and Migon 1993), and thus we compare the various approaches. Section 7 concludes the article with a summary 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 $\mathbf{x}_{i,t}$ 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 $\epsilon_{i,t}$:

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

where we assume all noise terms $\epsilon_{i,t}$ 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:

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

where the noise $\boldsymbol{\xi}_{i,t}$ is assumed to be normally distributed around zero with variance Σ, 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

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

where the noise $\boldsymbol{\gamma}_t$ is also assumed to be normally distributed around zero, with variance Σ_M. Initial conditions are:

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

For computational reasons, 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.

For our particular extension of the standard dynamic hierarchical model we have chosen to use model parameters Λ (i.e. evolution matrices G and A, variances and initial values $\hat{\mathbf{M}}_1$ and $\hat{\boldsymbol{\theta}}_1$) that are: 1) independent of time, and 2) identical for each parallel DLM. Time-dependent versions of several parameters within Λ would be tedious but doable. This generalization is however outside the scope of the present article. Moreover, the much larger number of time-dependent model parameters may easily lead to an 'overfitting' on the training data (when the number of observations per model parameter becomes too small, the model may specialize too much on the observed data, which leads to poor generalization).

Parameters that differ between parallel DLMs would undermine the idea of learning parallel tasks. In this paper these shared model parameters strengthen the bond between different DLMs, whereas this connection would be lost when parameters become task-dependent.

The model is visualized in Figure 1. A *directed graphical model* like this represents each latent state by a separate node (ellipse). Lines between nodes represent conditional independences. Unconnected nodes are said to be conditionally independent. Take for example nodes $\boldsymbol{\theta}_{2,t-1}$, $\boldsymbol{\theta}_{2,t}$ and $y_{2,t}$. Node $y_{2,t}$ is connected to $\boldsymbol{\theta}_{2,t}$, but not to $\boldsymbol{\theta}_{2,t-1}$. Therefore we can say that $y_{2,t}$ is conditionally independent of $\boldsymbol{\theta}_{2,t-1}$ given $\boldsymbol{\theta}_{2,t}$, or $P(y_{2,t}|\boldsymbol{\theta}_{2,t-1}, \boldsymbol{\theta}_{2,t}) = P(y_{2,t}|\boldsymbol{\theta}_{2,t})$. More information on graphical models can be found e.g. in (Pearl 1988).

2.2 Maximum Likelihood Estimation

We implement an ML estimation of the model parameters, which are shared between the parallel DLMs. This requires the maximization of $\log P(Y_{1..T}|\Lambda, X_{1..T})$, the (log) probability of *all* observations $Y_{1..T} = [y_{1,1}, \ldots, y_{n,T}]$, given the model parameters

$$\Lambda = \{\Sigma, \Sigma_M, \sigma, A, G, \hat{\mathbf{M}}_1, \Sigma_{M_1}, \hat{\boldsymbol{\theta}}_1, \hat{\Sigma}_1\}\,, \tag{5}$$

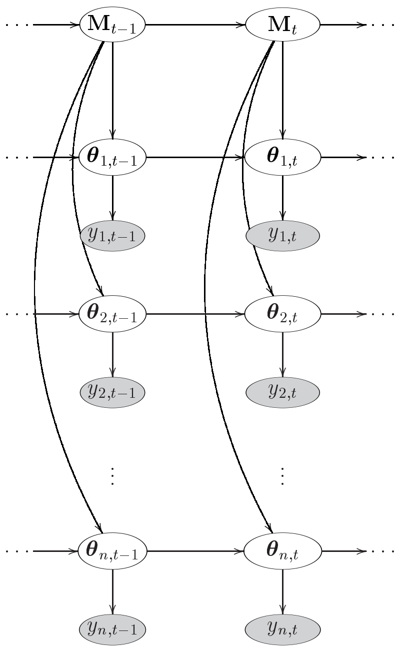


Figure 1. Graphical model. The shaded areas represent the observations $y_{i,t}$, the open ellipses represent the latent states. The top-level states (upper ellipses) are connected to all of the lower-level states (lower ellipses). Covariates $\mathbf{x}_{i,t}$ are left out for clarity.

and all covariates $X_{1..T} = [\mathbf{x}_{1,1}, \ldots, \mathbf{x}_{n,T}]$. We will assume from here on that $X_{1..T}$ is fixed and given, and omit it from our notation. Further, we define a super state

$$\mathbf{Z}_t = [\mathbf{M}_{t+1}, \boldsymbol{\theta}_{1t}, \ldots, \boldsymbol{\theta}_{nt}] \, , \tag{6}$$

and denote by $Z_{1..T}$ all latent states in the model up to time T.

A well-know method for ML estimation in such latent-variable models is the EM algorithm (see e.g. (Rubin 1991)). The EM algorithm is an iteration of two steps:

- E-step: we calculate $\langle \log P(Z_{1..T}, Y_{1..T}|\Lambda)\rangle_{P(Z_{1..T}|Y_{1..T},\Lambda')}$, the expectation value of the log probability of the latent states and the observations given Λ', the current best estimate for Λ. The set of model parameters Λ' is initialized (as Λ_0) at random at the start of the algorithm.
- M-step: we obtain a new estimate for the model parameters by choosing $\Lambda_{\text{new}} = \underset{\Lambda}{\text{argmax}} \langle \log P(Z_{1..T}, Y_{1..T}|\Lambda)\rangle_{P(Z_{1..T}|Y_{1..T},\Lambda')}$.

After each iteration Λ' is set to Λ_{new}, until the EM algorithm converges (when Λ_{new} is no longer significantly different from Λ'). The EM algorithm is guaranteed to converge to a (local) maximum of $\log P(Y_{1..T}|\Lambda)$. The expression that is maximized in the M-step is a quadratic expression in Λ, which makes this step relatively easy. The more involved part of the algorithm is the inference in the E-step. In 2.3 we give expressions for the required probability distributions, and in Sections 3.1 and 3.2 we describe approximations to perform this inference in linear time.

2.3 Inference

The ML estimation in Section 2.2 requires the calculation of the joint probability $P(Z_{1..T}, Y_{1..T}|\Lambda)$ and of the posterior probability of the latent states $P(Z_{1..T}|Y_{1..T}, \Lambda)$. The latter is directly related to the joint probability:

$$P(Z_{1..T}|Y_{1..T}, \Lambda) = \frac{P(Z_{1..T}, Y_{1..T}|\Lambda)}{\int dZ_{1..T} P(Z_{1..T}, Y_{1..T}|\Lambda)} \, . \tag{7}$$

The joint probability itself can be written as a product of 'two-slice potentials'

$$P(Z_{1..T}, Y_{1..T}|\Lambda) = \prod_t \Psi_t(\mathbf{Z}_{t-1}, \mathbf{Z}_t|\Lambda) \, , \tag{8}$$

with, for the hierarchical model of Figure 1,

$$\begin{aligned}\Psi_t(\mathbf{Z}_{t-1}, \mathbf{Z}_t|\Lambda) &= P(Y_t|\mathbf{Z}_t, \Lambda) P(\mathbf{Z}_t|\mathbf{Z}_{t-1}, \Lambda) \\ &= P(Y_t|\Theta_t, \Lambda) P(\Theta_t|\Theta_{t-1}, \mathbf{M}_t, \Lambda) P(\mathbf{M}_t|\mathbf{M}_{t-1}, \Lambda) \, ,\end{aligned} \tag{9}$$

where $\Theta_t = [\boldsymbol{\theta}_{1,t}, \ldots, \boldsymbol{\theta}_{n,t}]$ and we can further decompose

$$\begin{aligned}P(Y_t|\Theta_t, \Lambda) &= \prod_i P(y_{i,t}|\boldsymbol{\theta}_{i,t}, \Lambda) \quad \text{and} \\ P(\Theta_t|\Theta_{t-1}, \mathbf{M}_t, \Lambda) &= \prod_i P(\boldsymbol{\theta}_{i,t}|\boldsymbol{\theta}_{i,t-1}, \mathbf{M}_t, \Lambda) \, .\end{aligned} \tag{10}$$

The potential on the boundary $t = 1$ is slightly different:

$$\Psi_1(\mathbf{Z}_0, \mathbf{Z}_1, \Lambda) = \Psi_1(\mathbf{Z}_1, \Lambda) = P(Y_1|\Theta_1, \Lambda) P(\Theta_1, \Lambda) \, , \tag{11}$$

where $P(\Theta_1)$ is defined through Equation 4.

If we look at our dynamic hierarchical model in terms of the super state $\mathbf{Z}_t$ defined in Section 2.2, we can express the model as one large DLM and, in principle, we can apply the standard procedures for forecasting, filtering and smoothing. The evolution equation for $\mathbf{Z}_t$ reads

$$\begin{pmatrix} \mathbf{M}_{t+1} \\ \boldsymbol{\theta}_{1,t} \\ \boldsymbol{\theta}_{2,t} \\ \vdots \\ \boldsymbol{\theta}_{n,t} \end{pmatrix} = \begin{pmatrix} G & \mathbb{0} & \cdots\cdots & \mathbb{0} \\ \mathbb{1}-A & A & \mathbb{0} & \vdots \\ \mathbb{1}-A & \mathbb{0} & A \ \ddots & \vdots \\ \vdots & \vdots & \ddots \ \ddots & \mathbb{0} \\ \mathbb{1}-A & \mathbb{0} & \cdots \ \mathbb{0} & A \end{pmatrix} \begin{pmatrix} \mathbf{M}_t \\ \boldsymbol{\theta}_{1,t-1} \\ \boldsymbol{\theta}_{2,t-1} \\ \vdots \\ \boldsymbol{\theta}_{n,t-1} \end{pmatrix} + \Xi_t \tag{12}$$

where the noise covariance matrix Γ is of the form

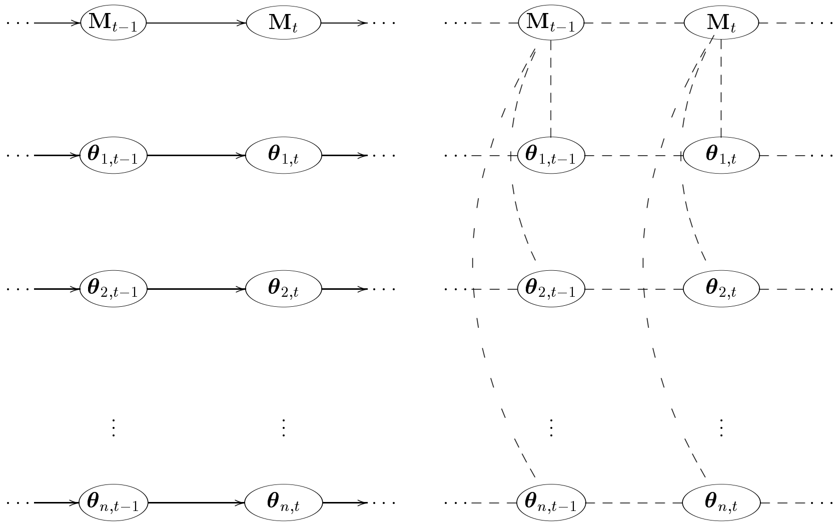
$$E[\Xi_t \Xi_t^T] = \begin{pmatrix} \Sigma_M & \mathbb{0} & \cdots & \mathbb{0} \\ \vdots & \Sigma & & \mathbb{0} \\ \mathbb{0} & & \ddots & \vdots \\ \mathbb{0} & \cdots & \mathbb{0} & \Sigma \end{pmatrix} . \tag{13}$$

Equations 12 and 13 describe a vector auto-regression VAR process (see e.g. (Chatfield 1989)). This process is stationary when the eigenvalues of the large evolution matrix in Equation 12 have absolute values that are smaller than 1. We do not explicitly constrain A and G to ensure stationarity; we do however maximize the likelihood of fixed observations $Y_{1..T}$ that have values within a fixed, finite range. Evolution matrices that lead to non-stable time series with extreme expectation values for $Z_{1..T}$ (and for $Y_{1..T}$) are therefore unlikely to occur.

It is possible to perform exact inference on Z, since all relationships are linear and all noise components are additive and normally distributed. However, the methods involved in this procedure require the inversion of matrices of size $(n+1) \cdot n_{\text{dim}} \times (n+1) \cdot n_{\text{dim}}$, with n the number of parallel DLMs and n_{dim} the dimension of the latent states. Such calculations become practically infeasible for large n: for many real-world problems an $n \gg 1$ would lead to unacceptable computation times. This is why in Sections 3.1 and 3.2 we will introduce two methods for approximate inference that scale linearly with the number of parallel series.

3 Approximations to the Dynamic Hierarchical Model

The combination of dependences between latent states within one DLM and between top- and lower-level DLMs in the model that is discussed in this article



(a) Graphical structure of the approximate model used in the variational approximation

(b) Graphical structure of the approximate model used in the factorial approximation

Figure 2. Graphical structure of the approximate models used in the variational approximation (left) and the factorial approximation (right). Ellipses represent latent states $\mathbf{M}_t$ for the top-level DLM and $\boldsymbol{\theta}_{i,t}$ for the lower-level time DLMs. Observations are left out for clarity. The dashed lines in the right graph indicate that although the approximate model is fully factorized, the connections between states are incorporated in each iteration step.

makes inference infeasible for large numbers of parallel time series. We therefore propose not to use the exact model to perform inference, but to use an approximation instead. We present two such approximations, a variational approximation and a factorial approximation. Both approximations feature only one, or even none of the aforementioned dependences explicitly. This makes it possible to perform inference within a reasonable time, even for larger numbers of parallel time series. Both approximations do however include both types of dependences *implicitly*, as will be shown in the remainder of this Section. In Section 6 we will show that the use of these approximate inference methods does not significantly affect performance.

3.1 A Variational Approximation

The variational approximation 'cuts' the dependences between the top-level states $\mathbf{M}_t$ and the lower-level states $\boldsymbol{\theta}_{i,t}$ (see Figure 2a). The resulting approximation $Q(Z_{1..T})$ of the posterior distribution reads

$$Q(Z_{1..T}) = Q_0(\mathbf{M}_{1..T}) \prod_i Q_i(\boldsymbol{\theta}_{i,1..T})$$
$$= \prod_{i=0}^{n} Q_i(\boldsymbol{\theta}_{i,1..T}) \; , \tag{14}$$

where for notational convenience we defined $\boldsymbol{\theta}_{0,t} = \mathbf{M}_t$. (See e.g. (Ghahramani and Jordan 1996; Jaakkola and Jordan 2000) for more details on this type of approximation.) This distribution uncouples the dynamics of the lower-level DLMs and the top-level DLM but does still take into account temporal dependences. The dependence between top- and lower-level states does however re-enter the equation when we minimize the KL-divergence between $Q(Z_{1..T})$ and the exact posterior:

$$\mathrm{KL}[Q, P] = \int dZ_{1..T} Q(Z_{1..T})[\log Q(Z_{1..T}) - \log P(Z_{1..T}|Y_{1..T}, \Lambda)] \; . \tag{15}$$

This minimization provides us with the closest (and thus presumably the best) approximation of the posterior, given the restrictions that we imposed on the functional form of $Q(Z_{1..T})$. Although the approximating model does not feature all dependences explicitly, the minimization of the KL-divergence does incorporate their *effect* in the approximation.

We minimize $\mathrm{KL}[Q, P]$ for (the parameters of) one function $Q_i(\boldsymbol{\theta}_{i,1..T})$ at a time, where we keep all other parameters fixed. The optimal $Q_i(\boldsymbol{\theta}_{i,1..T})$ can thus be expressed (see Appendix A) through a so-called 'mean field' equation:

$$Q_i(\boldsymbol{\theta}_{i,1..T}) \propto \exp\langle \log P(Z_{1..T}, Y_{1..T}) \rangle_{Q_{-i}} \; , \tag{16}$$

where $\langle .. \rangle_{Q_{-i}}$ denotes the average over the distribution

$$Q_{-i}(Z_{1..T}) = \prod_{j=0; j \neq i}^{n} Q_j(\boldsymbol{\theta}_{j,1..T}) \; . \tag{17}$$

One step of the approximation procedure includes the minimization of all functions $Q_i(\boldsymbol{\theta}_{i,1..T})$ (one by one). Since the optimal choice for $Q_i(\boldsymbol{\theta}_{i,1..T})$ depends on all other functions $Q_{-i}(Z_{1..T})$ through the expectation value in (16), we have to iterate this step until convergence, i.e. until there is no significant change in any of the functions $Q_i(\boldsymbol{\theta}_{i,1..T})$.

The mean field equations for the two-level hierarchical model take the following form. Let us first consider optimizing $Q_i(\boldsymbol{\theta}_{i,1..T})$ for one of the lower-level DLMs ($i \neq 0$). Substituting (9) and (10) in (16), we obtain

$$Q_i(\boldsymbol{\theta}_{i,1..T}) \propto \prod_t P(y_{i,t}|\boldsymbol{\theta}_{i,t}) \exp\langle \log P(\boldsymbol{\theta}_{i,t}|\boldsymbol{\theta}_{i,t-1}, \mathbf{M}_t) \rangle_{Q_0} \; . \tag{18}$$

This can be interpreted as the posterior of a standard DLM (as defined through Equations 1 and 2) with transformed states $\tilde{\boldsymbol{\theta}}_{i,t}$ and observations $\tilde{y}_{i,t}$. A detailed description is presented in Appendix A.

Similarly, with all $Q_i(\boldsymbol{\theta}_{i,1..T})$ fixed, the optimal $Q_0(\mathbf{M}_{1..T})$ can be found to obey

$$Q_0(\mathbf{M}_{1..T}) \propto \prod_t P(\mathbf{M}_t|\mathbf{M}_{t-1}) \prod_i \exp\langle \log P(\boldsymbol{\theta}_{i,t}|\boldsymbol{\theta}_{i,t-1}, \mathbf{M}_t)\rangle_{Q_i} . \tag{19}$$

This can also be interpreted as the posterior of a DLM if we can define an observation equation $P(\tilde{y}_t|\mathbf{M}_t)$ and observations $\tilde{y}_t$ such that

$$P(\tilde{y}_t|\mathbf{M}_t) \propto \prod_i \exp\langle \log P(\boldsymbol{\theta}_{i,t}|\boldsymbol{\theta}_{i,t-1}, \mathbf{M}_t)\rangle_{Q_i} . \tag{20}$$

$P(\tilde{y}_t|\mathbf{M}_t)$ and $\tilde{y}_t$ are defined in Appendix A. The standard form (a DLM) of the partial posterior distributions $Q_i(\boldsymbol{\theta}_{i,1..T})$ makes it relatively straightforward to calculate the expectation values $\langle\boldsymbol{\theta}_{i,t}\rangle_{Q_i(\boldsymbol{\theta}_{i,1..T})}$, which we will exploit in the next part of this Section.

A more detailed observation of Equations 18 and 19 shows that each $Q_i(\boldsymbol{\theta}_{i,1..T})$, $(i \neq 0)$ depends on $Q_0(\mathbf{M}_{1..T})$ only through the expectation values $\langle\mathbf{M}_t\rangle_{Q_0(\mathbf{M}_{1..T})}$, and vice versa: $Q_0(\mathbf{M}_{1..T})$ depends on the lower-level approximations only through the expectation values $\langle\boldsymbol{\theta}_{i,t}\rangle_{Q_i(\boldsymbol{\theta}_{i,1..T})}$. This, and the fact that each approximation can be written in the form of a DLM, facilitates the following iterative optimization procedure:

(1) Start with random values for $\langle\mathbf{M}_t\rangle_{Q_0(\mathbf{M}_{1..T})}$.
(2) Construct the posterior of the lower-level states given $\langle\mathbf{M}_t\rangle_{Q_0(\mathbf{M}_{1..T})}$.
(3) Infer $\langle\boldsymbol{\theta}_{i,t}\rangle_{Q_i(\boldsymbol{\theta}_{i,1..T})}$ from this posterior, for $i = 1..n$.
(4) Construct the posterior of the top-level states given $\langle\boldsymbol{\theta}_{i,t}\rangle_{Q_i(\boldsymbol{\theta}_{i,1..T})}$.
(5) Infer $\langle\mathbf{M}_t\rangle_{Q_0(\mathbf{M}_{1..T})}$ from this posterior.
(6) Repeat from step 2 until convergence.

In each iteration step we perform inference on $n+1$ independent DLMs: once for the top-level states and n times for the lower-level states. The approximate distribution obtained in this way is a first-order Markov model, and is visualized in Figure 2a. Inference is performed through a method called 'Kalman smoothing' (see e.g. (Haykin 2001)). This method is standard for DLMs, and it is linear in T.

It can be shown (see Appendix B) that the marginal means $\langle\boldsymbol{\theta}_{i,t}\rangle_{Q_i(\boldsymbol{\theta}_{i,1..T})}$, $\langle\mathbf{M}_t\rangle_{Q_0(\mathbf{M}_{1..T})}$ under the (converged) approximating distribution are identical to the marginal means under the exact distribution. Any discrepancy between approximate and exact inference is therefore due to the error made in the estimation of the variances.

3.2 A Factorial Approach

As an alternative to the variational approximation, we consider a method known as 'expectation propagation' (see (Minka 2001)), which is an iterative variant of a local optimization algorithm introduced by (Boyen and Koller 1998). We approximate the posterior $P(Z_{1..T}|Y_{1..T})$ with a distribution of the form

$$Q(Z_{1..T}) = \prod_t Q_t(\mathbf{Z}_t) = \prod_t Q_{0,t}(\mathbf{M}_t) \prod_i Q_{i,t}(\boldsymbol{\theta}_{i,t}) \, . \tag{21}$$

Note that this distribution is fully factorized, i.e. does not contain any links between consecutive states. We further decompose Q into

$$Q(Z_{1..T}) = \prod_t \lambda_t(\mathbf{Z}_t)\mu_t(\mathbf{Z}_t) \tag{22}$$

with

$$\lambda_t(\mathbf{Z}_t) = \lambda_{0,t}(\mathbf{M}_t) \prod_i \lambda_{i,t}(\boldsymbol{\theta}_{i,t}) \, ,$$

and similarly for $\mu_t(\mathbf{Z}_t)$.

At first sight, this factorized distribution seems to be less powerful than the approximation in (14) for the variational approach since no dependences whatsoever remain. This fully factorized approximation will be fitted to the exact dynamic hierarchical model again and, like in the variational approximation, the various dependences will be brought back implicitly into the approximating model.

The factorial approach does not (explicitly) minimize a distance measure between the approximate and the exact posterior distribution, but instead *equates* parts of the approximation to parts of the exact distribution. Consider the approximate posterior as described in Equations 21 and 22, and the exact posterior as described in (7) and (8). In each step of the factorial approximation process we equate part of the approximating distribution to the exact potential $\Psi_t(\mathbf{Z}_{t-1}, \mathbf{Z}_t)$, yielding:

$$Q(Z_{1..T}) \propto \prod_{t'<t-1} Q_{t'}(\mathbf{Z}_{t'})\lambda_{t-1}(\mathbf{Z}_{t-1})\Psi_t(\mathbf{Z}_{t-1}, \mathbf{Z}_t)\mu_t(\mathbf{Z}_t) \prod_{t'>t} Q_{t'}(\mathbf{Z}_{t'}) \, ,$$

where the approximating functions $\lambda_t(\mathbf{Z}_t)$ and $\mu_{t-1}(\mathbf{Z}_{t-1})$ have been replaced by the exact potential $\Psi_t(\mathbf{Z}_{t-1}, \mathbf{Z}_t)$. The optimal estimates for $\lambda_t(\mathbf{Z}_t)$ and $\mu_{t-1}(\mathbf{Z}_{t-1})$ are found through a number of marginalization steps. First, we calculate the joint distribution of $\mathbf{Z}_{t-1}$ and $\mathbf{Z}_t$:

$$Q_{t-1:t}(\mathbf{Z}_{t-1}, \mathbf{Z}_t) = \frac{1}{c_t} \lambda_{t-1}(\mathbf{Z}_{t-1}) \Psi_t(\mathbf{Z}_{t-1}, \mathbf{Z}_t) \mu_t(\mathbf{Z}_t) \ , \tag{23}$$

where c_t is a normalization constant. Further marginalization straightforwardly leads to expressions for each function $\lambda_{i,t}(\boldsymbol{\theta}_{i,t})$ and $\mu_{i,t-1}(\boldsymbol{\theta}_{i,t-1})$. For example, integration of $Q_{t-1:t}(\mathbf{Z}_{t-1}, \mathbf{Z}_t)$ over $\mathbf{Z}_{t-1}$ and $\mathbf{Z}_{-i,t}$ (all elements of $\mathbf{Z}_t$ except $\boldsymbol{\theta}_{i,t}$) yields the update formula for $\lambda_{i,t}$:

$$\lambda_{0,t}(\mathbf{M}_t) = \frac{Q_{0,t}(\mathbf{M}_t)}{\mu_{0,t}(\mathbf{M}_t)} \quad \text{and} \quad \lambda_{i,t}(\boldsymbol{\theta}_{i,t}) = \frac{Q_{i,t}(\boldsymbol{\theta}_{i,t})}{\mu_{i,t}(\boldsymbol{\theta}_{i,t})} \ , \tag{24}$$

where $Q_{0,t}$ and $Q_{i,t}$ are the marginal distributions for $\mathbf{M}_t$ and $\boldsymbol{\theta}_{i,t}$, respectively. Similarly, integration over $\mathbf{Z}_{-i,t-1}$ and $\mathbf{Z}_t$ yields

$$\mu_{0,t-1}(\mathbf{M}_{t-1}) = \frac{Q_{0,t-1}(\mathbf{M}_{t-1})}{\lambda_{0,t-1}(\mathbf{M}_{t-1})} \quad \text{and} \quad \mu_{i,t-1}(\boldsymbol{\theta}_{i,t-1}) = \frac{Q_{i,t-1}(\boldsymbol{\theta}_{i,t-1})}{\lambda_{i,t-1}(\boldsymbol{\theta}_{i,t-1})} \ . \tag{25}$$

In principle the choice of the ordering of the updates is free. However, it makes sense to iterate going forward and backward in time: first compute all λ's from $t = 1$ to T, then all μ's from $t = T$ to 1, then update all λ's from $t = 1$ to T, and so forth until convergence.

At convergence, we can improve upon the fully factorized approximation by computing the two-slice distributions

$$Q_{0,t-1:t}(\mathbf{M}_{t-1}, \mathbf{M}_t) = \int d\boldsymbol{\theta}_{1,t-1} \dots d\boldsymbol{\theta}_{n,t-1} d\boldsymbol{\theta}_{1,t} \dots d\boldsymbol{\theta}_{n,t} Q_{t-1:t}(\mathbf{Z}_{t-1}, \mathbf{Z}_t) \ , \tag{26}$$

with $Q_{t-1:t}(\mathbf{Z}_{t-1}, \mathbf{Z}_t)$ as defined in Equation 23, and similarly for $Q_{i,t-1:t}(\boldsymbol{\theta}_{i,t-1}, \boldsymbol{\theta}_{i,t})$ and $Q_{0,i,t}(\mathbf{M}_t, \boldsymbol{\theta}_{i,t})$. By construction, the single-slice marginals $Q_{0,t-1}(\mathbf{M}_{t-1}), Q_{0,t}(\mathbf{M}_t), Q_{i,t-1}(\boldsymbol{\theta}_{i,t-1})$ and $Q_{i,t}(\boldsymbol{\theta}_{i,t})$ are consistent with these two-slice marginals.

The factorial approach can be called 'greedy', since it iterates over local estimation steps instead of performing one global estimation. Each iteration step involves an update, where (part of) the factorial distribution is fitted to Equation 8. The factorial approximation has very few functional constraints: each marginal probability $Q_{i,t}(\boldsymbol{\theta}_{i,t})$ can in principle be fitted independently from all the others. This, in combination with the local, greedy optimization algorithm, is expected to yield very close approximations to the exact marginal probabilities $P(\mathbf{M}_t|Y_{1..T},\Lambda)$, $P(\boldsymbol{\theta}_{i,t}|Y_{1..T},\Lambda)$, $P(\mathbf{M}_{t-1},\mathbf{M}_t|Y_{1..T},\Lambda)$, $P(\boldsymbol{\theta}_{i,t-1},\boldsymbol{\theta}_{i,t}|Y_{1..T},\Lambda)$ and $P(\mathbf{M}_t,\boldsymbol{\theta}_{i,t}|Y_{1..T},\Lambda)$, which are used in the learning algorithm that is described in Section 2.2. We will show (in Appendix C) that the means in the factorial approximation coincide with the means of the exact posterior. Another way to show that this approximation features exact means is presented in (Rusmevichientong and Van Roy 2001). In Section 6 we further show that (for our example) the factorial approach makes a closer approximation of the variance than the variational approximation does.

4 Generalization to More than Two Levels

It is straightforward to generalize the exact hierarchical method and the two approximations to more than two levels. We outline the extension to a three-level model (see also Figure 3). Addition of subsequent higher levels proceeds in the same way.

Parameters at the lowest level, which refer to the lower-level DLMs and couple directly to the observations, are denoted $\boldsymbol{\phi}_{i,j,t}$. These parameters represent latent states that, as before, depend on the previous states $\boldsymbol{\phi}_{i,j,t-1}$ and on a mid-level state, denoted $\boldsymbol{\theta}_{i,t}$. These mid-level states $\boldsymbol{\theta}_{i,t}$ are part of an ensemble of 'mid-level DLMs', that constitute the second level. Each state on this level is connected to a selection from the lowest-level states (all with the same index i), to a top-level state $\mathbf{M}_t$ and of course to its predecessor $\boldsymbol{\theta}_{i,t-1}$.

In the variational approximation we now iterate over three levels. The approximations for the two outer levels are expressed as before, for the two-level model. The solution for the middle level contains averages over both lowest- and mid-level dynamics:

$$Q_i(\boldsymbol{\theta}_{i,1..T}) \propto \prod_t \exp\langle \log P(\boldsymbol{\phi}_{i,j,t}|\boldsymbol{\phi}_{i,j,t-1},\boldsymbol{\theta}_{i,t})\rangle_{Q_{i,j}} \times \\ \exp\langle \log P(\boldsymbol{\theta}_{i,t}|\boldsymbol{\theta}_{i,t-1},\mathbf{M}_t)\rangle_{Q_0} \,. \tag{27}$$

The obvious iteration scheme is lower-middle-top-middle-lower-middle-top and so on.

The factorial approach proceeds in much the same way as before. The two-slice potential $\Psi_t(\mathbf{Z}_{t-1},\mathbf{Z}_t)$ now contains terms from three levels. These terms include connections within each level as well as connections between lowest- and mid-level states, and between mid- and top-level states. For the three-level model of Figure 3 this potential reads

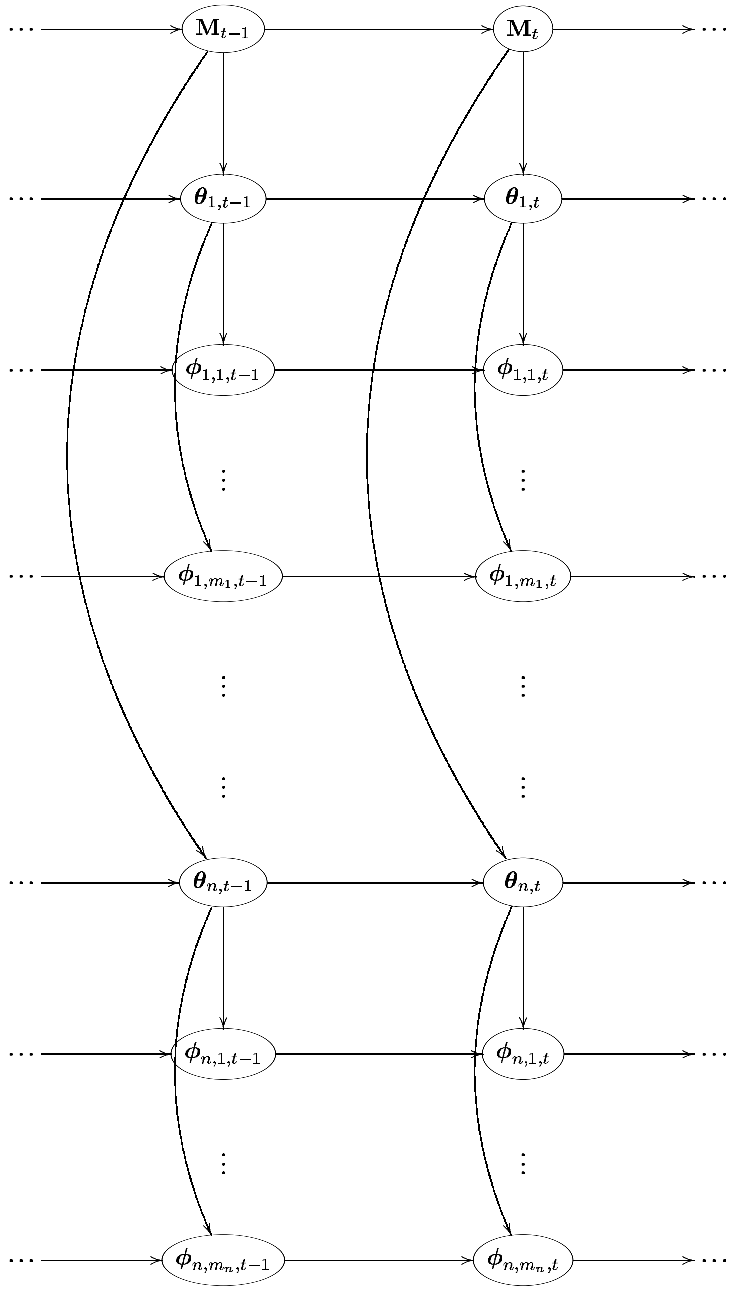


Figure 3. Three-level model. The top ellipses represent the latent states of the top-level states $\mathbf{M}_t$. The latent states $\boldsymbol{\theta}_{i,t}$, which are connected to $\mathbf{M}_t$, are on the second level. Each of these states is connected to a set of latent states $\boldsymbol{\phi}_{i,j,t}$ on the lowest level. Observations (left out for clarity) are connected to the lowest-level states.

$$\begin{aligned}\Psi_t(\mathbf{Z}_{t-1}, \mathbf{Z}_t) &= P(Y_t|\mathbf{Z}_t)P(\mathbf{Z}_t|\mathbf{Z}_{t-1}) \\ &= P(Y_t|\Phi_t)P(\Phi_t|\Phi_{t-1}, \Theta_t)P(\Theta_t|\Theta_{t-1}, \mathbf{M}_t)P(\mathbf{M}_t|\mathbf{M}_{t-1})\ ,\end{aligned} \tag{28}$$

where we can further decompose

$$
\begin{aligned}
P(Y_t|\Phi_t) &= \prod_i \prod_{j=1}^{n_i} P(y_{i,j,t}|\boldsymbol{\phi}_{i,j,t}), \\
P(\Phi_t|\Phi_{t-1},\Theta_t) &= \prod_i \prod_{j=1}^{n_i} P(\boldsymbol{\phi}_{i,j,t}|\boldsymbol{\phi}_{i,j,t-1},\boldsymbol{\theta}_{i,t}) \quad \text{and} \\
P(\Theta_t|\Theta_{t-1},\mathbf{M}_t) &= \prod_i P(\boldsymbol{\theta}_{i,t}|\boldsymbol{\theta}_{i,t-1},\mathbf{M}_t)\ ,
\end{aligned}
$$

where n_i is the number of lower-level DLMs that is connected to the mid-level DLM $\boldsymbol{\theta}_{i,t}$. Forward and backward passes are performed in the same way as before, where marginalization now yields factorized distributions $Q_{0,t}(\mathbf{M}_t)$, $Q_{i,j,t}(\boldsymbol{\phi}_{i,j,t})$ and $Q_{i,t}(\boldsymbol{\theta}_{i,t})$.

5 Related Work

We have already mentioned the work of Gamerman and Migon (Gamerman and Migon 1993), which describes a top-level DLM with latent states $\mathbf{M}_t$, and lower-level latent states $\boldsymbol{\theta}_{i,t}$ that are inferred from the top-level states $\mathbf{M}_t$ (via a hierarchical model). This model does feature no direct dependences between different states $\boldsymbol{\theta}_{i,t}$. The subject of hierarchical dynamic models has of course been treated in a wider context. Cargnoni, Müller and West (1997) presented a model with the same hierarchical structure as (Gamerman and Migon 1993), but for non-normal multivariate time series. The non-normality of this model prevents exact inference, and sampling methods are proposed to simulate the posterior.

An alternative method to modeling parallel time series has been presented by Zhang, Lin, Raz and Sowers (1998). Each parallel time series is modeled as the sum of a common (for all time series) smoothing spline function, a regression term (or fixed effect) $\mathbf{x}_{i,t}^T\beta$ on the covariates $\mathbf{x}_{i,t}$ and a series-dependent random effects term $\mathbf{z}_{i,t}^T\mathbf{b}_i$, where $\mathbf{z}$ is the subset of the covariates that corresponds to the random effect.

More work on time series modeling through the use of random effects has been done by Aguilar and West (1998). Here, the fixed effect is modeled through a DLM structure, whereas the random effects are not linked through time.

Camargo and Gamerman (2000) have combined a hierarchical model structure with a DLM structure within parallel time series through a mixture model. The mixture elements are the hierarchical model of (Gamerman and Migon 1993), and independent DLMs for each parallel series. The probability of the mixture components is a time-dependent Bernoulli distribution.

An example of a more traditional approach can be found in the work of Bunn and Vassilopoulos (1999), who implement a combination of individual and group seasonal estimation for short-term sales forecasting of multiple retail products.

The work presented in this article is to our knowledge the first implementation of a graphical model that features dependences both between lower-level states at subsequent times and between top-level and lower-level states.

6 Results

Description of the data. We tested our model on two databases, one created artificially, the other containing real-world data. The covariates $\mathbf{x}_{i,t}$ and the initial state means $\boldsymbol{\theta}_1, \mathbf{M}_1$ in the artificial database were drawn from a normal distribution with zero mean and unit variance. Propagation matrices for the lower-level DLMs and the top-level DLM and covariance matrices for the three sources of noise (on the top-level state predictions, on the lower-level state predictions and on the observations) were also generated randomly. The latent states were 3-dimensional. Given these parameters, observations $y_{i,t}$ were generated according to Equations 1, 2 and 3.

The second database contains sales figures for single copy newspaper sales in the Netherlands. The observations represent the numbers of newspapers sold on 156 consecutive Saturdays, at 343 separate outlets throughout the Netherlands. The covariates contain information about the weather, season, short-term previous sales (four to six weeks ago) and long-term previous sales (54 to 56 weeks ago). Results from previous studies (see e.g. (Bakker and Heskes 2003)) were used to implement a lower-dimensional (3-dimensional) representation of the input covariates. Both covariates and observations are scaled per outlet to have zero mean and unit variance.

Application of the extended model. Figure 4 shows the means over time for the first dimension of the three-dimensional latent state vectors for four parallel newspaper outlets and the top-level DLM. Plots are drawn for the model we propose here and for the DHM as decribed in (Gamerman and Migon 1993), i.e. without direct dependences between lower-level latent states. It can be seen that under the latter model the dynamics for the parallel DLMs are less smooth than for our approach. This is due to the fact that lower-level states in the standard DHM are not linked through time. The new dependences also translate to better predictions, as will be shown later in this section.

Quality of the approximations. For both the top-level DLM and the lower-level DLMs we inferred the single state marginals $P(\boldsymbol{\theta}_{i,t}|Y_{1..T}, \Lambda)$ using the exact posterior, the variational approach or the factorial approximation. We

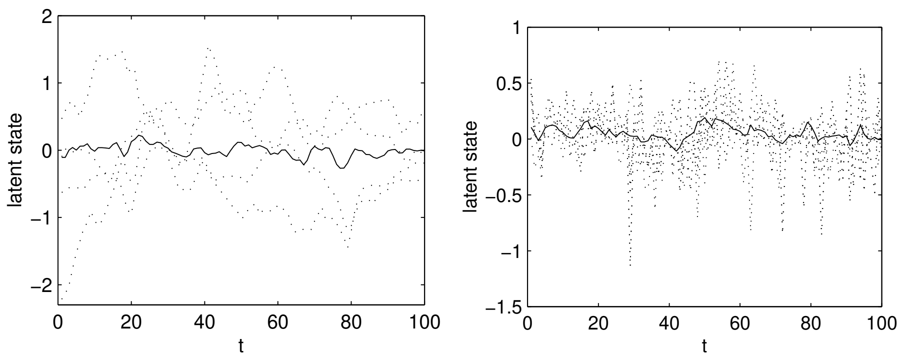


Figure 4. Means for the first elements of the latent states over time for 4 outlets from the newspaper data set, and the corresponding top-level DLM. Dotted lines correspond to lower-level states, solid lines represent top-level states. The left panel plots the exact means for the hierarchical model presented in this article, means inferred from the standard hierarchical model are plotted in the right panel.

rated the quality of both approximations through the KL-divergences between the approximated and the exact marginals. Inference was performed on random selections of series, taken from the artificial data set. For each random draw we selected n (varying from 10 to 50) different series of 80 consecutive observations and covariates. The parameters Λ for the generating model were chosen randomly for each draw, and for each n we performed inference on 10 independent draws. The mean-dependent parts of the KL-distances for both approximations were always zero after convergence, as expected (see Appendix B and C). The variance-dependent parts of the KL-distances (averaged over the 10 independent draws) are shown in Figure 5, as a function of n. The factorial approximation is clearly more accurate than the variational approximation for any number of tasks, although the difference does get smaller for increasing n. An example of the difference between the estimated variances for the two approximations is given in the lower panels of the same figure.

Forecasting. For both databases we applied the EM algorithm to find ML model parameters for different numbers of parallel tasks (n). For the inference step in this algorithm we used either exact inference, or one of the approximating methods. For the newspaper database we also applied inference based on the DHM *without* connections between subsequent lower-level states. For each number of tasks, and for each inference method, we performed 10 independent ML estimations, where each time we took a random selection of n tasks to be used for ML estimation and evaluation. The first 80 observation/covariate pairs of each task were used for estimation of the model parameters, one-step-ahead forecasting was performed on the subsequent 20 data samples ($\mathbf{x}_{i,t}, y_{i,t}$ for $t = 81 \ldots 100$). That is, we obtained ML parameters Λ_{ML}, and used them to calculate

$$\langle y_{i,t} \rangle_{P(y_{i,t}|\mathbf{x}_{i,t}, Y_{1..t-1}, \Lambda_{\mathrm{ML}})} \,, \tag{29}$$

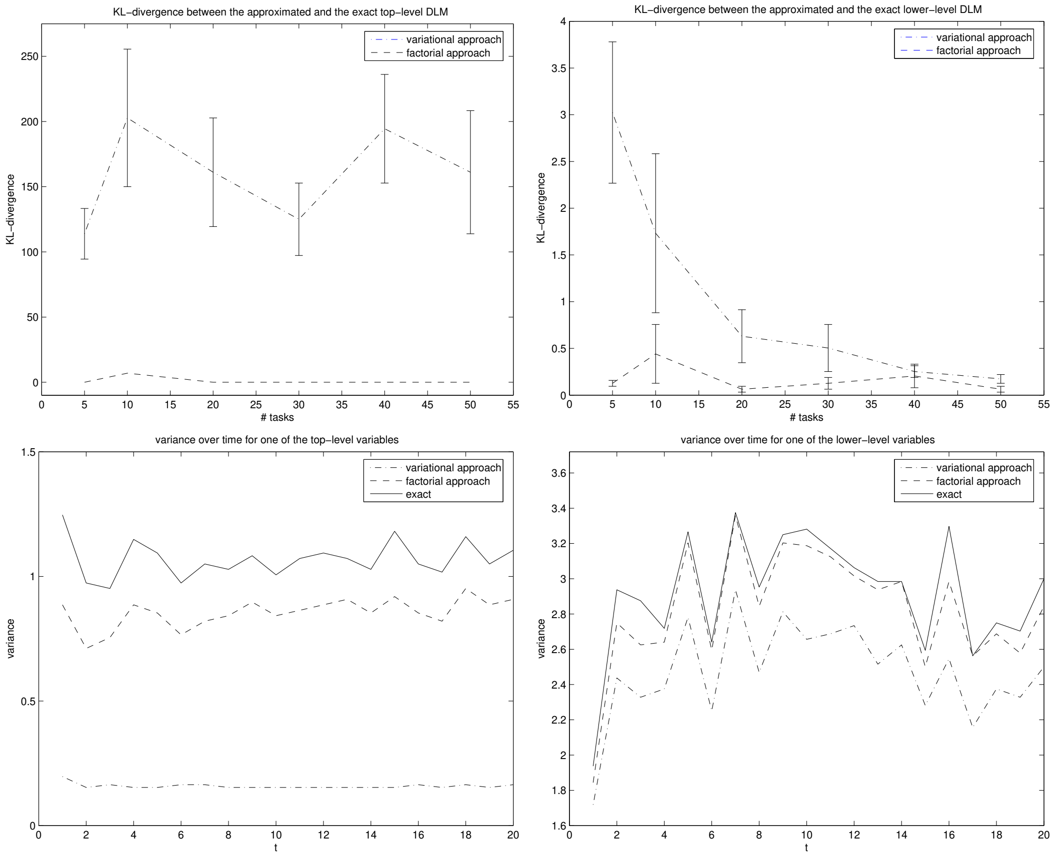


Figure 5. Top: Average variance-dependent parts of the KL-divergences between the approximated marginals and the exact marginals. The left panel plots the average divergences between marginals of the top-level DLM, the right panel plots the average divergences between the marginals of the lower-level DLMs. The dash-dotted line plots the divergence between the variational approach and the exact model, the dashed line the divergence for the factorial approach. Bottom: the variance in the first dimension of the top-level state (left) and one from a set of 10 lower-level states (right) over time. Solid lines correspond to the exact model, dash-dotted lines to the variational approximation and dashed lines to the factorial approximation.

for $t = 81 \ldots 100$, starting with $t = 81$. After prediction of each $y_{i,t}$ we used the true value for $y_{i,t}$ to update the posterior, where we kept the old values for Λ_{ML}. In each trial we used the same inference method (exact or approximate) both for ML estimation and for posterior updates. Each of the trials was rated through the average squared error

$$E = [20 \cdot n]^{-1} \sum_{i=1}^{n} \sum_{t=81}^{100} \left(\langle y_{i,t} \rangle - y_{i,t} \right)^2 , \tag{30}$$

and computation time. Since the outputs $y_{i,t}$ for both data sets have zero mean and unit variance, an error of $E = 1$ corresponds to a model that always predicts $y_{i,t} = 0$.

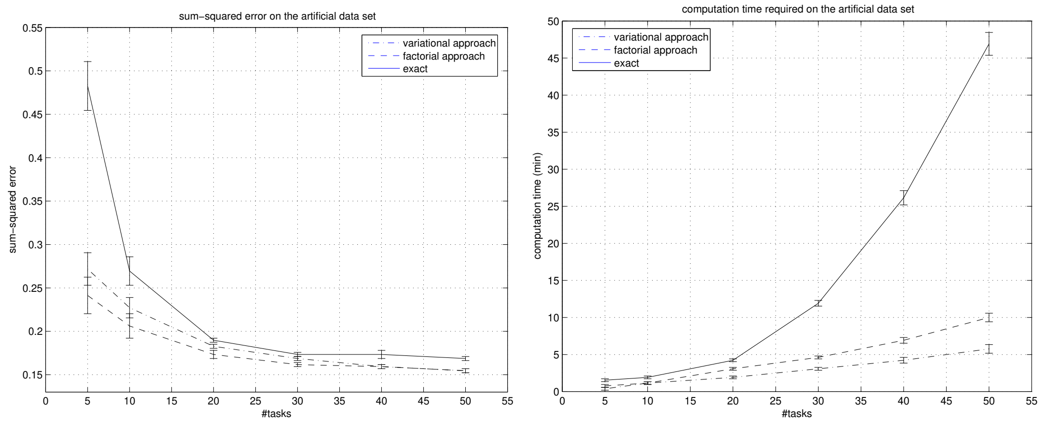


Figure 6. Average squared error (left) and computation time (right) as a function of the number of parallel tasks for the variational approximation (dash-dotted line), the factorial approximation (dashed line) and the exact model (solid line).

Figure 6 shows the average squared error and required computation time as a function of the number of parallel DLMs for the artificial data set. Results are displayed for the two approximation methods and for exact inference. It can be seen that, whereas the computation time for exact inference grows strongly with the number of parallel tasks, for either form of approximate inference it grows only linearly. This gain of speed infers but a small loss of accuracy: the average squared error of the approximate inference methods is not significantly higher than the error incurred through exact inference. Similar results were obtained from the newspaper data. Note that, since the means in both approximate inference methods are identical to the exact means (given identical model parameters Λ), the expectation values for the responses $y_{i,t}$ are exact as well: an approximation of the *optimized* exact model (i.e., approximate inference on a model with parameters Λ that were obtained using exact inference) would feature the exact same responses, and have the same error. The difference in performance between the exact model and the approximations is due to the fact that the approximations are made *during* ML estimation (in the E-step of the EM algorithm), and not afterwards.

For very small numbers of tasks, the results for exact inference are actually inferior to those for approximate inference. This is due to a form of 'overfitting': although the observations that were used for ML estimation have a high likelihood under the model with parameters Λ_{ML}, the model generalizes poorly for new observations. This effect appears to be weaker in both approximate methods. Note however that parallel time series modeling is aimed at larger numbers of parallel tasks, where this problem is no longer an issue.

Performance on the newspaper data is presented in Figure 7. We consider the average performance of four different forecasting methods on 10 randomly chosen sets of parallel tasks (i.e. sales outlets). Forecasting performance, which is extremely poor ($E \approx 1$) when we use only a single task, improves strongly

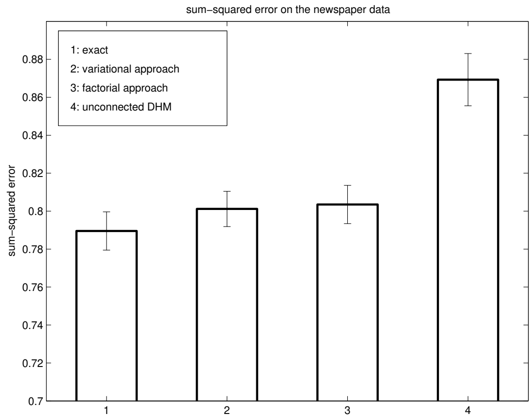


Figure 7. The average squared error on the newspaper data for the variational approximation (bar 2), the factorial approximation (bar 3) and the standard DHM (bar 4). The performance of the extended model with exact inference is represented by bar 1. Each error is an average over 16 parallel tasks; a larger number of parallel tasks did not further decrease the sum-squared error.

with increasing numbers of tasks n, until $n = 16$. None of the methods that we considered showed significant further improvement in performance for larger numbers of parallel tasks. The first three methods include the dynamic hierarchical model that is described in this article, once with exact inference in the expectation step of the EM-algorithm, once using the variational approach and once using the factorial approximation. The fourth method involves the DHMwithout dependences between the lower-level latent states. The ML estimation process appears not to be hindered by the use of approximate inference instead of exact inference. The standard DHM however, incurs a clearly higher error than all of the methods presented in this article.

7 Discussion

In this article we have presented a dynamic hierarchical model with dependences between the lower-level states in each time series. We have showed that such dependences are beneficial on a database of real-world data: predictions based on the new model were more accurate than those based on a similar model without these dependences.

Exact inference in the connected model is not practically feasible for large numbers of parallel time series. We therefore presented two methods for approximate inference and showed that both methods, although they are approximations, do feature exact means for the latent states. Comparison of the approximations to exact inference on two different databases confirmed that whereas the required computation time for exact inference is cubic in the number of parallel time series, for the approximating methods it is only linear. The performance after approximate inference was shown to be competitive with the performance after exact inference.

The two approximating methods were compared two each other, both in terms of KL-divergence between approximate and exact marginals and in terms of forecasting and required computation time. Inference through the slightly slower factorial approximation was shown to be closer to exact inference than inference through the variational approximation. This superiority with respect to inference did however not cash out in the form of better predictions (since both approximations did not perform significantly worse than exact inference anyway). Nevertheless, other, more demanding datasets may still benefit from a closer approximation.

Interesting work on the *clustering* of groups of parallel time series has been done by Smyth (1997) and Gaffney and Smyth (1999). They present a method to detect similarities between parallel time series, which are used to group these series into meaningful clusters. This idea of 'time series clustering', which is primarily concerned with finding a meaningful structure inside the parallel series, may be successfully combined with multitask learning, which is aimed at making better predictions. Multitask learning could be more effective when the full set of parallel tasks is divided into smaller subsets (or clusters), and different sets of hyperparameters are used for different clusters.

The current model features only continuous latent states, and all distributions are Gaussian. Interesting work may be done in the implementation of switching Kalman filters, discrete state variables and non-normal distributions.

References

[Aguilar and West 1998] Aguilar, O. and West, M., 1998. Analysis of hospital quality monitors using hierarchical time series models. In: Case Studies Bayesian Statistics in Science and Technology: Case Studies 4, New York. Springer-Verlag.

[Bakker and Heskes 2003] Bakker, B. and Heskes, T., 2003. Task clustering and gating for Bayesian multitask learning. Journal of Machine Learning Research, 4:83–99.

[Boyen and Koller 1998] Boyen, X. and Koller, D., 1998. Tractable inference for complex stochastic processes. Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, 33–42.

[Bunn and Vassilopoulos 1999] Bunn, D. and Vassilopoulos, A., 1999. Comparison of seasonal estimation methods in multi-item short-term forecasting. International Journal of Forecasting, 15:431–443.

[Camargo and Gamerman 2000] Camargo, E. and Gamerman, D., 2000. Discrete mixture alternatives to dynamic hierarchical models. Estadstica, 52:39–77.

[Cargnoni, Müller and West 1997] Cargnoni, C., Müller, P., and West, M., 1997. Bayesian forecasting of multinomial time series through conditionally Gaussian dynamic models. Journal of the American Statistical Association, 92:640–647.

[Chatfield 1989] Chatfield, C., 1989. The Analysis of Time Series: an Introduction. Chapman & Hall, London, fourth edition.

[Gaffney and Smyth 1999] Gaffney, S. and Smyth, P., 1999. Trajectory clustering using mixtures of regression models. In: Proceedings of the ACM SIGKDD Conference, pages 63–72.

[Gamerman and Migon 1993] Gamerman, D. and Migon, H., 1993. Dynamic hierarchical models. Journal of the Royal Statistical Society, Series B, 55:629–642.

[Ghahramani and Jordan 1996] Ghahramani, Z. and Jordan, M., 1996. Factorial hidden Markov models. In: Touretzky, D., Mozer, M., and Hasselmo, M.: Advances in Neural Information Processing Systems 8, pages 472–478. MIT Press.

[Harrison and Stevens 1976] Harrison, P. and Stevens, C., 1976. Bayesian forecasting. Journal of the Royal Statistical Society B, 38:205–227.

[Haykin 2001] Haykin, S., editor, 2001. Kalman Filtering and Neural Networks. Wiley, Canada.

[Jaakkola and Jordan 2000] Jaakkola, T. and Jordan, M., 2000. Bayesian parameter estimation via variational methods. Statistics and Computing, 10:25–37.

[Jazwinski 1970] Jazwinski, A., 1970. Stochastic Processes and Filtering Theory. Academic Press, New York.

[Kim 1994] Kim, C., 1994. Dynamic linear models with Markov-switching. Journal of Econometrics, 60:1–22.

[Lindley and Smith 1972] Lindley, D. and Smith, A., 1972. Bayes estimates for the linear model. Journal of the Royal Statistical Society B, 34:1–41.

[Minka 2001] Minka, T., 2001. Expectation propagation for approximate Bayesian inference. In: Uncertainty in Artificial Intelligence: Proceedings of the Seventeenth Conference (UAI-2001), pages 362–369, San Francisco, CA. Morgan Kaufmann Publishers.

[Pearl 1988] Pearl, J., 1988. Probabilistic Reasoning in Intelligent systems: Networks of Plausible Inference. Morgan Kaufmann, San Francisco, CA.

[Rubin 1991] Rubin, D. B., 1991. EM and beyond. Psychometrika, 56(2):241–254.

[Rusmevichientong and Van Roy 2001] Rusmevichientong, P. and Van Roy, B., 2001. An analysis of belief propagation on the turbo decoding graph with gaussian densities. IEEE Transactions on Information Theory, 47:745–765.

[Smyth 1997] Smyth, P., 1997. Clustering sequences with hidden Markov models. In: Mozer, M., Jordan, M., and Petsche, T., editors: Advances in Neural Information Processing Systems, volume 9, pages 648–654. The MIT Press.

[West and Harrison 1997] West, M. and Harrison, J., editors, 1997. Bayesian Forecasting and Dynamic Models. Springer, New York.

[Zhang, Lin, Raz and Sowers 1998] Zhang, D., Lin, X., Raz, J., and Sowers, M., 1998. Semiparametric stochastic mixed models for longtitudinal data. Journal of the American Statistical Association, 93:710–719.

APPENDIX A: THE VARIATIONAL APPROACH

The variational approach features an approximation of the posterior $P(Z_{1..T}|Y_{1..T},\Lambda)$ in the form of $Q(Z_{1..T}) = \prod_{i=0}^{n} Q_i(\boldsymbol{\theta}_{i,1..T})$. We find those parameters for $Q(.)$ that minimize

$$\mathrm{KL}[Q,P] = \int dZ_{1..T} Q(Z_{1..T})[\log Q(Z_{1..T}) - \log P(Z_{1..T}|Y_{1..T},\Lambda)]\,. \tag{31}$$

This minimization is performed iteratively with respect to (the parameters of) each $Q_i(\boldsymbol{\theta}_{i,1..T})$ in turn. The part of the KL-divergence that depends on $Q_i(\boldsymbol{\theta}_{i,1..T})$ reads

$$\begin{aligned}\mathrm{KL}[Q_i,P] &= \int dZ_{1..T} Q(Z_{1..T})[\log Q_i(\boldsymbol{\theta}_{i,1..T}) - \\ &\qquad\qquad \log P(Z_{1..T}|Y_{1..T},\Lambda)] \\ &\propto \int d\boldsymbol{\theta}_{i,1..T} Q_i(\boldsymbol{\theta}_{i,1..T})[\log Q_i(\boldsymbol{\theta}_{i,1..T}) - \\ &\quad \int dZ_{-i,1..T} Q_{-i}(Z_{-i,1..T}) \log P(Z_{1..T}|Y_{1..T},\Lambda)]\,,\end{aligned} \tag{32}$$

where $Q_{-i}(Z_{-i,1..T})$ is the product over all $Q_j(\boldsymbol{\theta}_{j,1..T})$ except for $j = i$. The last line is in fact the KL-divergence between $Q_i(\boldsymbol{\theta}_{i,1..T})$ and

$$\exp[\int dZ_{-i,1..T} Q_{-i}(Z_{-i,1..T}) \log P(Z_{1..T}|Y_{1..T},\Lambda)]\,. \tag{33}$$

Minimization of the KL-divergence with respect to $Q_i(\boldsymbol{\theta}_{i,1..T})$ therefore implies

$$Q_i(\boldsymbol{\theta}_{i,1..T}) = \exp\langle \log P(Z_{1..T}|Y_{1..T},\Lambda)\rangle_{Q_{-i}}\,. \tag{34}$$

Each step of the iterative ML estimation process minimizes the KL-divergence between the approximating and the exact distribution with respect to one of the distributions $Q_i(\boldsymbol{\theta}_{i,1..T})$. The optimum for each $Q_i(\boldsymbol{\theta}_{i,1..T})$ can be interpreted as the posterior of a standard Kalman filter model under the proper variable transformations. In the following we show what transformations are required both for $Q_i(\boldsymbol{\theta}_{i,1..T})$, $i > 0$ and for $Q_0(\boldsymbol{\theta}_{0,1..T})$, the top-level DLM.

The variational approximation for the lower-level DLMs with index i,

$$Q_i(\boldsymbol{\theta}_{i,1..T}) \propto \prod_t P(y_{i,t}|\boldsymbol{\theta}_{i,t}, \Lambda) \exp\langle \log P(\boldsymbol{\theta}_{i,t}|\boldsymbol{\theta}_{i,t-1}, \mathbf{M}_t, \Lambda)\rangle_{Q_0} , \tag{35}$$

can be written as the posterior of a standard DLM with

$$\tilde{\boldsymbol{\theta}}_{i,t} = \boldsymbol{\theta}_{i,t} - \boldsymbol{\alpha}_t , \tag{36}$$

where

$$\boldsymbol{\alpha}_1 = \mathbf{0} \quad \text{(a vector of zeros)} \quad , \tag{37}$$

$$\boldsymbol{\alpha}_t = A\boldsymbol{\alpha}_{t-1} + (\mathbb{1} - A)\langle \mathbf{M}_t \rangle \text{ for } t > 1 , \tag{38}$$

and

$$\tilde{y}_{i,t} = y_{i,t} - \mathbf{x}_{i,t}^T \boldsymbol{\alpha}_t . \tag{39}$$

Carefully writing out the terms in Equation (35) and substituting $\tilde{\boldsymbol{\theta}}_{i,t} + \boldsymbol{\alpha}_t$ for $\boldsymbol{\theta}_{i,t}$ yields

$$Q_i(\boldsymbol{\theta}_{i,1..T}) \propto P(\tilde{\boldsymbol{\theta}}_{i,1}|\Lambda) \prod_{t=2}^{T} P(\tilde{\boldsymbol{\theta}}_{i,t}|\tilde{\boldsymbol{\theta}}_{i,t-1}, \Lambda) \prod_{t=1}^{T} P(\tilde{y}_{i,t}|\mathbf{x}_{i,t}, \tilde{\boldsymbol{\theta}}_{i,t}, \Lambda) , \tag{40}$$

a DLM with states $\tilde{\boldsymbol{\theta}}_{i,t}$ and observations $\tilde{y}_{i,t}$.

The approximation for the top-level DLM $Q_0(\mathbf{M}_{1..T})$ can be interpreted as the posterior of a DLM with evolution equation

$$\mathbf{M}_t = G\mathbf{M}_{t-1} + \boldsymbol{\gamma}_t$$

and observation equation $P(\tilde{\mathbf{y}}_t|\mathbf{M}_t, \Lambda)$. The observations $\tilde{\mathbf{y}}_t$ and the observation covariance matrix are defined as

$$\tilde{\mathbf{y}}_t = n^{-1} \sum_i (\mathbb{1} - A)^{-1} (\langle \boldsymbol{\theta}_{i,t} \rangle - A\langle \boldsymbol{\theta}_{i,t-1} \rangle) \text{ for } t > 1 \tag{41}$$

$$\tilde{\mathbf{y}}_1 = \mathbf{0} \tag{42}$$

and

$$\tilde{\Sigma}_y^{-1} = n(\mathbb{1} - A)^T \Sigma^{-1} (\mathbb{1} - A) \,. \tag{43}$$

Since $\tilde{\mathbf{y}}_t$ is a vector, we define covariate matrices $\tilde{C}_t$ (instead of covariate vectors $\mathbf{x}_t$). These matrices contain only zeros for $t = 1$ and are defined as the unity matrix otherwise. Inserting these transformed parameters, we obtain

$$\begin{aligned} P(\tilde{\mathbf{y}}_t|\mathbf{M}_t, \Lambda) &= \prod_i \exp\langle \log P(\boldsymbol{\theta}_{i,t}|\boldsymbol{\theta}_{i,t-1}, \mathbf{M}_t, \Lambda)\rangle_{Q_i} \\ &\propto \exp(-\tfrac{1}{2}\tilde{\mathbf{y}}_t^T \tilde{\Sigma}_y^{-1} \tilde{\mathbf{y}}_t + \tilde{\mathbf{y}}_t^T \tilde{\Sigma}_y^{-1} \tilde{C}_t \mathbf{M}_t \\ &\qquad\qquad - \tfrac{1}{2}\mathbf{M}_t^T \tilde{C}_t^T \tilde{\Sigma}_y^{-1} \tilde{C}_t \mathbf{M}_t) \\ &\propto P(\tilde{\mathbf{y}}_t|\boldsymbol{\theta}_{i,t}, \tilde{C}_t, \Lambda) \,, \end{aligned} \tag{44}$$

an observation equation of the form 1 with predictions

$$\tilde{\mathbf{y}}_t = \tilde{C}_t \mathbf{M}_t + \Theta_t, \quad E(\Theta_t \Theta_t^T) = \tilde{\Sigma}_y \,. \tag{45}$$

Note that in the second line we dropped all terms that do not depend on $\mathbf{M}_t$, and added a term $\tilde{\mathbf{y}}_t^T \tilde{\Sigma}_y^{-1} \tilde{\mathbf{y}}_t$, which does not depend on $\mathbf{M}_t$ either.

APPENDIX B: EXACT MEANS IN THE VARIATIONAL APPROACH

It can be proven that the means that are inferred from the variational approximation coincide with the exact means. In the variational approximation the KL-divergence between the approximate posterior and the exact posterior is minimized. Both distributions can be rewritten as a normal distribution over one large state that contains all latent states present in the model. The means under the two posterior distributions will be shown to coincide when the KL-divergence between the distributions is minimized.

Both the exact and the approximate posterior can be written as one large normal distribution of $\mathbf{Z}_{1..T}$, which is a vector of length $L = n_{\text{inputs}} \cdot (n+1) \cdot T$ and strings together all states at all times:

$$\mathbf{Z}_{1..T} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_T] \ , \tag{46}$$

where $\mathbf{Z}_t$ is the super state for time t, defined in Section 2.2.

The exact posterior reads:

$$P(\mathbf{Z}_{1..T}|Y_{1..T}, \Lambda) = \mathcal{N}(\mathbf{m}, \Sigma_Z) \ , \tag{47}$$

where Σ_Z is a block matrix of dimension $L \times L$, defined through:

$$\Sigma_Z^{-1} = \begin{pmatrix} \Gamma_{0,0}^{-1} & \Gamma_{0,1}^{-1} & \ldots & \Gamma_{0,n}^{-1} \\ \Gamma_{1,0}^{-1} & \Gamma_{1,1}^{-1} & \ldots & \Gamma_{1,n}^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_{n,0}^{-1} & \Gamma_{n,1}^{-1} & \ldots & \Gamma_{n,n}^{-1} \end{pmatrix} \ , \tag{48}$$

and the blocks $\Gamma_{i,j}^{-1}$ are $n_{\text{inputs}} \cdot T \times n_{\text{inputs}} \cdot T$ matrices corresponding to the inverse covariance of $\Theta_i = [\boldsymbol{\theta}_{i,1}, \ldots, \boldsymbol{\theta}_{i,T}]$ and $\Theta_j = [\boldsymbol{\theta}_{j,1}, \ldots, \boldsymbol{\theta}_{j,T}]$. The elements of this matrices follow from Equations 1 through 4 and Equations 7 through 11. The same equations define the means $\mathbf{m}$.

For the variational approximation we can define a similar distribution. The approximate posterior reads

$$\hat{P}(\mathbf{Z}_{1..T}|Y_{1..T}, \Lambda) = \mathcal{N}(\hat{\mathbf{m}}, \hat{\Sigma}_Z) \ , \tag{49}$$

with

$$
\hat{\Sigma}_Z^{-1} = \begin{pmatrix} \hat{\Gamma}_{0,0}^{-1} & \mathbb{0} & \dots & \mathbb{0} \\ \mathbb{0} & \hat{\Gamma}_{1,1}^{-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbb{0} \\ \mathbb{0} & \dots & \mathbb{0} & \hat{\Gamma}_{n,n}^{-1} \end{pmatrix} , \tag{50}
$$

where the blocks $\hat{\Gamma}_{i,i}^{-1}$ and the elements of $\hat{\mathbf{m}}$ are free to choose (provided that $\hat{\Gamma}_{i,i}$ is positive definite).

Note that for both the exact and the approximating distribution, it does not matter whether we write it in the form of a (series of) DLM(s), or as one large normal distribution. The inferred means, for example, are the same for both ways of writing.

The KL-divergence between Q and P has a mean-dependent term that reads

$$
\mathrm{KL}_{\mathrm{mean}} = \tfrac{1}{2}(\mathbf{m} - \hat{\mathbf{m}})^T \Sigma_Z (\mathbf{m} - \hat{\mathbf{m}}) , \tag{51}
$$

and a term that depends only on the variances. Since there are no constraints on what values $\hat{\mathbf{m}}$ may take, it is clear that minimization of the KL-divergence, and therefore minimization of the above mean-dependent term with respect to $\hat{\mathbf{m}}$, infers that the approximating means $\hat{\mathbf{m}}$ are identical to the exact means $\mathbf{m}$.

The variance-dependent term reads

$$
\mathrm{KL}_{\mathrm{variance}} = \tfrac{1}{2}\mathrm{Tr}\ \Sigma_Z^{-\frac{1}{2}} \hat{\Sigma}_Z \Sigma_Z^{-\frac{1}{2}} - \tfrac{1}{2}\log|\hat{\Sigma}_Z| , \tag{52}
$$

where $\mathrm{Tr}\ M$ and $|M|$ represent the trace of M and the determinant of M, respectively. Minimization of this term with respect to $\hat{\Sigma}_Z$ would yield the exact variance if the elements of $\hat{\Sigma}_Z$ were completely free to choose. This is however not the case: the variational approximation excludes dependences between states $\boldsymbol{\theta}_{i,t}$ and $\boldsymbol{\theta}_{j,t'}$ for $i \neq j$, and the corresponding elements of $\hat{\Sigma}_Z$ must be zero. The approximate total variance $\hat{\Sigma}_Z$ can therefore not be identical to the exact variance Σ_Z, and the marginal variances $\hat{\Sigma}_{Z,it}$ and $\Sigma_{Z,it}$ do not coincide.

APPENDIX C: EXACT MEANS IN THE FACTORIAL APPROACH

The means inferred through the factorial approximation are identical to the exact means as well. To show this, we first note that fixed points of expectation propagation as performed in Section 3.2 correspond to fixed points of the 'Bethe free energy' (see (Minka 2001)), which reads

$$
\begin{aligned}
F(\Psi, Q) = & -\sum_{t=1}^{T-1} \int d\mathbf{Z}_t Q_t(\mathbf{Z}_t) \log Q_t(\mathbf{Z}_t) \\
& + \sum_{t=1}^{T} \int dZ_{t-1,t} Q_{t-1:t}(Z_{t-1,t}) \log[\frac{Q_{t-1:t}(Z_{t-1,t})}{\Psi_t(Z_{t-1,t})}] \,,
\end{aligned} \tag{53}
$$

where

$$
Q_t(\mathbf{Z}_t) = \prod_{i=0}^{n} \lambda_{i,t}(\boldsymbol{\theta}_{i,t}) \mu_{i,t}(\boldsymbol{\theta}_{i,t}) \tag{54}
$$

and

$$
Q_{t-1:t}(Z_{t-1,t}) = \prod_{i=0}^{n} \lambda_{i,t-1}(\boldsymbol{\theta}_{i,t}) \Psi_t(Z_{t-1,t}) \mu_{i,t}(\boldsymbol{\theta}_{i,t}) \,, \tag{55}
$$

as in Section 3.2, and Ψ_t is the potential defined in (9).

The integral $\int d\mathbf{Z} Q(\mathbf{Z}) \log Q(\mathbf{Z})$, where $Q(.)$ is a multivariate normal distribution, is in fact independent of the parameter values of $Q(.)$. The larger part of the above expression for $F(\Psi, Q)$ therefore has no influence on the fixed points and can be ignored, leaving the $Q(.)$ dependent part

$$
-\sum_{t=1}^{T} \int dZ_{t-1,t} Q_{t-1:t}(Z_{t-1,t}) \log \Psi_t(Z_{t-1,t}) \,.
$$

Equation 53 can be re-expressed further as

$$
\begin{aligned}
F(\Psi, Q) = & -\mathrm{KL}(Q(Z_{1..T}) | P(Z_{1..T})) \\
& +\mathrm{KL}(Q_{\mathrm{even}}(Z_{1..T}) | P(Z_{1..T})) \\
& +\mathrm{KL}(Q_{\mathrm{odd}}(Z_{1..T}) | P(Z_{1..T})) + C \,,
\end{aligned} \tag{56}
$$

the sum of three KL-divergences, each between an approximating distribution and the exact posterior distribution for the extended DHM. The distributions in Equation 56 are defined through:

$$
\begin{aligned}
Q(Z_{1..T}) &= \prod_t Q_t(\mathbf{Z}_t) \,, \\
Q_{\text{even}}(Z_{1..T}) &= \prod_{t=\text{even}} Q_{t-1:t}(Z_{t-1,t}) \,, \\
Q_{\text{odd}}(Z_{1..T}) &= \prod_{t=\text{odd}} Q_{t-1:t}(Z_{t-1,t}) \,,
\end{aligned}
$$

and

$$
P(Z_{1..T}) \propto \prod_{t=1}^{T} \Psi_t(Z_{t-1,t}) \,.
$$

Minimization of (the mean-dependent part of) this sum of KL-divergences will be shown to imply exact means for the factorial approximation.

The equality (56) follows directly when we write out the expressions for the three KL-divergences:

$$
\begin{aligned}
&\mathrm{KL}(Q(Z_{1..T})|P(Z_{1..T})) = \\
&\quad \int dZ_{1..T} \prod_{t'=1}^{T} Q_{t'}(\mathbf{Z}_{t'}) \log \frac{\prod_{t=1}^{T} Q_t(\mathbf{Z}_t)}{\prod_{t=1}^{T} \Psi_t(Z_{t-1,t})} \\
&= \sum_{t=1}^{T} \int d\mathbf{Z}_t Q_t(\mathbf{Z}_t) \log Q_t(\mathbf{Z}_t) - \\
&\quad \sum_{t=1}^{T} \int dZ_{t-1,t} Q_{t-1}(\mathbf{Z}_{t-1}) Q_t(\mathbf{Z}_t) \log \Psi_t(Z_{t-1,t}) \\
&= C_1 - \sum_{t=1}^{T} \int dZ_{t-1,t} Q_{t-1}(\mathbf{Z}_{t-1}) Q_t(\mathbf{Z}_t) \log \Psi_t(Z_{t-1,t}) \,,
\end{aligned}
\tag{57}
$$

where C_1 is a constant.

The KL-divergence between the exact posterior and the 'even' approximation reads

$$
\begin{aligned}
&\mathrm{KL}(Q_{\text{even}}(Z_{1..T})|P(Z_{1..T})) = \\
&\int dZ_{1..T} \prod_{t'=\text{even}} Q_{t'-1:t'}(Z_{t'-1,t'}) \log \frac{\prod_{t=\text{even}} Q_{t-1:t}(Z_{t-1,t})}{\prod_{t=1}^{T} \Psi_t(Z_{t-1,t})} \\
&\quad = \sum_{t=\text{even}} \int dZ_{t-1,t} Q_{t-1:t}(Z_{t-1,t}) \log Q_{t-1:t}(Z_{t-1,t}) \\
&\quad\quad - \sum_{t=\text{even}} \int dZ_{t-1,t} Q_{t-1:t}(Z_{t-1,t}) \log \Psi_t(Z_{t-1,t}) \\
&\quad\quad - \sum_{t=\text{odd}} \int dZ_{t-2,..,t+1} Q_{t-2:t-1}(Z_{t-2,t-1}) \times \\
&\quad\quad\quad\quad \log \Psi_t(Z_{t-1,t}) Q_{t:t+1}(Z_{t,t+1}) \,.
\end{aligned}
\tag{58}
$$

The expression $\log \Psi_t(Z_{t-1,t})$ contains terms that are linear or quadratic in $\mathbf{Z}_{t-1}$ and $\mathbf{Z}_t$ (including cross terms between the two). Since convergence of the factorial approach implies that

$$\langle \mathbf{Z}_t \rangle_{Q_{t-1:t}} = \langle \mathbf{Z}_t \rangle_{Q_{t:t+1}} = \langle \mathbf{Z}_t \rangle_{Q_t} \; ,$$

$$\langle \mathbf{Z}_t^T \mathbf{Z}_t \rangle_{Q_{t-1:t}} = \langle \mathbf{Z}_t^T \mathbf{Z}_t \rangle_{Q_{t:t+1}} = \langle \mathbf{Z}_t^T \mathbf{Z}_t \rangle_{Q_t}$$

and

$$\langle \langle \mathbf{Z}_{t-1}^T \mathbf{Z}_t \rangle_{Q_{t-2:t-1}} \rangle_{Q_{t:t+1}} = \langle \mathbf{Z}_{t-1} \rangle_{Q_{t-1}}^T \langle \mathbf{Z}_t \rangle_{Q_t} \; ,$$

we can simplify Equation 58 to

$$\begin{aligned} \mathrm{KL}(Q_{\text{even}}(Z_{1..T})|P(Z_{1..T})) = {} & C_2 \\ & - \sum_{t=\text{even}} \int dZ_{t-1,t} Q_{t-1:t}(Z_{t-1,t}) \log \Psi_t(Z_{t-1,t}) \\ & - \sum_{t=\text{odd}} \int dZ_{t-1,t} Q_{t-1}(\mathbf{Z}_{t-1}) Q_t(\mathbf{Z}_t) \log \Psi_t(Z_{t-1,t}) \; . \end{aligned} \tag{59}$$

Similarly, we find that

$$\begin{aligned} \mathrm{KL}(Q_{\text{odd}}(Z_{1..T})|P(Z_{1..T})) = {} & C_3 \\ & - \sum_{t=\text{odd}} \int dZ_{t-1,t} Q_{t-1:t}(Z_{t-1,t}) \log \Psi_t(Z_{t-1,t}) \\ & - \sum_{t=\text{even}} \int dZ_{t-1,t} Q_{t-1}(\mathbf{Z}_{t-1}) Q_t(\mathbf{Z}_t) \log \Psi_t(Z_{t-1,t}) \; . \end{aligned} \tag{60}$$

In the sum of KL-divergences in Equation 56 the first lines in Equations 59 and 60 combine to form the $Q(.)$ dependent part of Equation 53, whereas the second lines are exactly canceled by the term from $\mathrm{KL}(Q(Z_{1..T})|P(Z_{1..T}))$.

As in Appendix B we can write out the mean-dependent terms of each of the three KL-divergences. These terms are in fact identical, since each KL-divergence features the same exact mean $\mathbf{m}$ and, due to the convergence properties of the factorial approach, have identical approximate means $\hat{\mathbf{m}}$ as well. Therefore,

$$\mathrm{KL}_{\text{mean}} = \tfrac{1}{2}(\mathbf{m} - \hat{\mathbf{m}})^T \Sigma_Z (\mathbf{m} - \hat{\mathbf{m}}) \; , \tag{61}$$

the exact same term as in 51, where now $\hat{\mathbf{m}}$ is the mean of $Q(Z_{1..T})$ (and of $Q_{\text{even}}(Z_{1..T})$ and $Q_{\text{odd}}(Z_{1..T})$). Minimization of (61) infers exact means in the factorial approach.

Similarly, we can express the variance-dependent term as

$$
\begin{aligned}
\mathrm{KL}_{\text{variance}} = {} & \tfrac{1}{2}\mathrm{Tr}\ \Sigma_Z^{-\frac{1}{2}}(\hat{\Sigma}_{Z,\text{even}} + \hat{\Sigma}_{Z,\text{odd}} - \hat{\Sigma}_Z)\Sigma_Z^{-\frac{1}{2}} - \\
& \tfrac{1}{2}\log|\hat{\Sigma}_{Z,\text{even}}| - \tfrac{1}{2}\log|\hat{\Sigma}_{Z,\text{odd}}| + \tfrac{1}{2}\log|\hat{\Sigma}_Z|\,,
\end{aligned}
$$

where $\hat{\Sigma}^{-1}_{Z,\text{odd}}$, $\hat{\Sigma}^{-1}_{Z,\text{even}}$ and $\hat{\Sigma}^{-1}_Z$ read

$$
\begin{pmatrix}
\hat{\Gamma}^{-1}_{1,1} & \mathbb{0} & \mathbb{0} & \cdots\cdots & \cdots & \cdots & \mathbb{0} \\
\mathbb{0} & \hat{\Gamma}^{-1}_{2,2} & \hat{\Gamma}^{-1}_{2,3} & & & & \vdots \\
\mathbb{0} & \hat{\Gamma}^{-1}_{3,2} & \hat{\Gamma}^{-1}_{3,3} & & \mathbb{0} & & \vdots \\
\vdots & & & \ddots & & & \vdots \\
\vdots & & & & \ddots & & \vdots \\
\vdots & & \mathbb{0} & & \hat{\Gamma}^{-1}_{T-2,T-2} & \hat{\Gamma}^{-1}_{T-2,T-1} & \mathbb{0} \\
\vdots & & & & \hat{\Gamma}^{-1}_{T-1,T-2} & \hat{\Gamma}^{-1}_{T-1,T-1} & \mathbb{0} \\
\mathbb{0} & \cdots & \cdots & \cdots\cdots & \mathbb{0} & \mathbb{0} & \hat{\Gamma}^{-1}_{T,T}
\end{pmatrix},
$$

$$
\begin{pmatrix}
\hat{\Gamma}^{-1}_{1,1} & \hat{\Gamma}^{-1}_{1,2} & \mathbb{0} & \mathbb{0} & \cdots & \cdots & \mathbb{0} & \mathbb{0} \\
\hat{\Gamma}^{-1}_{2,1} & \hat{\Gamma}^{-1}_{2,2} & \mathbb{0} & \mathbb{0} & \cdots & \cdots & \mathbb{0} & \mathbb{0} \\
\mathbb{0} & \mathbb{0} & \hat{\Gamma}^{-1}_{3,3} & \hat{\Gamma}^{-1}_{3,4} & \ddots & & \vdots & \vdots \\
\mathbb{0} & \mathbb{0} & \hat{\Gamma}^{-1}_{4,3} & \hat{\Gamma}^{-1}_{4,4} & & \ddots & \vdots & \vdots \\
\vdots & \vdots & \ddots & & \ddots & & \mathbb{0} & \mathbb{0} \\
\vdots & \vdots & & \ddots & & \ddots & \mathbb{0} & \mathbb{0} \\
\mathbb{0} & \mathbb{0} & \cdots & \cdots & \mathbb{0} & \mathbb{0} & \hat{\Gamma}^{-1}_{T-1,T-1} & \hat{\Gamma}^{-1}_{T-1,T} \\
\mathbb{0} & \mathbb{0} & \cdots & \cdots & \mathbb{0} & \mathbb{0} & \hat{\Gamma}^{-1}_{T,T-1} & \hat{\Gamma}^{-1}_{T,T}
\end{pmatrix}
$$

and

$$
\begin{pmatrix}
\hat{\Gamma}_{1,1}^{-1} & \mathbb{0} & \cdots & \cdots & \cdots & \cdots & \cdots & \mathbb{0} \\
\mathbb{0} & \hat{\Gamma}_{2,2}^{-1} & \ddots & & & & & \mathbb{0} \\
\vdots & \ddots & \hat{\Gamma}_{3,3}^{-1} & \ddots & & & & \vdots \\
\vdots & & \ddots & \hat{\Gamma}_{4,4}^{-1} & & \ddots & & \vdots \\
\vdots & \vdots & \ddots & & \ddots & & & \vdots \\
\vdots & & & & \ddots & \ddots & \ddots & \vdots \\
\vdots & & & & & \ddots & \hat{\Gamma}_{T-1,T-1}^{-1} & \mathbb{0} \\
\mathbb{0} & \cdots & \cdots & \cdots & \cdots & \cdots & \mathbb{0} & \hat{\Gamma}_{T,T}^{-1}
\end{pmatrix}
$$

respectively, where the blocks $\hat{\Gamma}_{t,t'}^{-1}$ are $n_{\text{inputs}} \cdot n \times n_{\text{inputs}} \cdot n$ matrices corresponding to the approximate inverse covariance of $\Theta_t = [\boldsymbol{\theta}_{1,t}, \ldots, \boldsymbol{\theta}_{n,t}]$ and $\Theta_{t'} = [\boldsymbol{\theta}_{1,t'}, \ldots, \boldsymbol{\theta}_{n,t'}]$. Although the approximated variance in the factorial approach does contain non-block-diagonal terms (as opposed to the variational approximation), the variance dependent term $\text{KL}_{\text{variance}}$ will generally still not be zero.