Variational Bayesian inference for forecasting hierarchical time series

Mijung Park MIJUNG@GATSBY.UCL.AC.UK
The Gatsby Computational Neuroscience Unit, University College London
Alexandra House, 17 Queen Square, London, WC1N 3AR, U.K.

Marcel Nassar MNASSAR@UTEXAS.EDU
Samsung Mobile Solutions Lab
4921 Directors Place 100, San Diego, CA 92121, U.S.A.

Abstract

In many real world data, time series are often hierarchically organized. Based on features such as products or geography, time series can be aggregated and disaggregated at several different levels. The so-called '*hierarchical time series*' are often forecast using simple top-town or bottom-up approaches. In this paper, we build a probabilistic model that involves dynamically evolving latent variables to capture the proportion changes in time series at each hierarchy. We derive the variational Bayesian expectation maximisation (VBEM) algorithm under the new model. In our algorithm, we implement the posterior inference in a sequential manner that significantly decreases computational overhead common in large hierarchical time series data. Furthermore, unlike the standard EM algorithm that provides point estimates of model parameters, our algorithm yields the distribution over the model parameters, which give us an insight to which subset of features yields the proportion changes of the time series. Simulation results show that our method significantly outperforms other methods in prediction.

1. Introduction

Time series in business and economics are often organized in a hierarchical structure based on dimensions such as products or regions. Forecasting the hierarchical time series is an important task in a number of industrial sectors (Sbrana & Silvestrini, 2013; Fischer et al., 2013; Kalchschmidt et al., 2006; Zotteri & Kalchschmidt, 2007; Fliedner, 1999). For instance, companies that offer a broad range of items or services to their customers need to plan their future supply process in order to minimize potential costs (e.g., inventory costs) (Kerkkaenen et al., 2009). A similar problem occurs in governmental budgeting that needs to be optimized throughout hierarchically organized departments. As an example, military budgeting in the context of hierarchical time series forecasting can be found in (Moon et al., 2013; 2012).

Accepted to the Divergence Methods in Probabilistic Inference (DMPI) workshop at International Conference on Machine Learning (ICML), Beijing, China, 2014.

1.1. Motivation

The main challenge in forecasting hierarchical time series is that various components at different levels of a hierarchy can interact in a complex manner: small changes at a given level of hierarchy can largely affect the time series at other levels. Furthermore, forecasting each times series individually can be time consuming and computationally intensive for large datasets. Consequently, there is a dire need for rapid, efficient, and automated forecasting methods that exploit the hierarchy in the time series to obtain better prediction performance.

1.2. Prior Work

Different approaches to forecasting hierarchical time series data are summarized in (Athanasopoulos et al., 2009; Hyndman et al., 2011) and can be categorized into three main categories: (1) top-down methods; (2) bottom-up methods; and (3) alternative statistical methods. In the following, we review each of these categories in detail.

1) Top-down (TD) approaches: distribute the top-level forecasts down the hierarchy using the historical proportions of the data (Gross & Sohl, 1990; Fliedner, 1999). As a result, only the top level forecast of the time series is needed. Examples of proportions are: (1) *average historical proportions* that are the average of the historical proportions of the bottom level series relative to the top level series over a certain period; and (2) *proportions of the his-*

torical averages that are average historical values of the bottom level series relative to the average values at the top level series. However, since these methods rely on historical and static proportion changes, they are unable to capture temporal dynamics in individual series. More recent work uses the so-called *forecasted proportion (FP)*, which is the proportions of the top-level forecasts relative to lower-level forecasts. Unfortunately, this approach produces biased revised forecasts even if base forecasts are unbiased (Athanasopoulos et al., 2009).

2) Bottom-up (BU) approaches: aggregate upper level time series from the bottom level time series using a summation operation based on the hierarchy of the data. These methods require forecasting the bottom level series only and do not lose any information due to aggregation. While these approaches are the most commonly used to hierarchical forecasting (Dangerfield & Morris, 1992; Zellner & Tobias, 2000), modeling the bottom level series is quite challenging due to the large amount of inherent noise in the individual time series.

3) Alternative statistical approaches: entail forecasting each series at an intermediate level of the hierarchy, then aggregating those at higher levels and disaggregating those at lower levels. This approach does not take into account inherent correlations of the hierarchy. For example, the "optimal combination approach" introduced in (Hyndman et al., 2011) entails individually forecasting all time series at all levels of the hierarchy, then revising the entire time series using a regression model that is fit by minimizing the variance among all revised forecasts. This method, which we denote by *Optimal*, finds a linear weight vector that optimally revises individual forecasts in the entire hierarchy (i.e. optimal in minimum variance sense). This method produces unbiased forecasts which are consistent throughout the entire hierarchy. However, it is computationally expensive compared to the other methods introduced so far because it requires individually forecasting all the time series at all levels of the hierarchy. Furthermore, the obtained linear weight vector can overfit the data as it happens frequently in the linear regression setting.

1.3. Contribution

In this paper, we introduce a hierarchical Bayesian *dynamic proportions* model (DPM) to hierarchical forecasting. Our method entails forecasting the least *noisy* top-level time series; and then sequentially disaggregating the predicted time series from the top to the bottom, i.e., the top level forecast is disaggregated into a second level forecast, which in turn is used to disaggregate the third level, *etc.*, based on the latent dynamical systems that control proportion changes in time series.

We demonstrate that DPM captures many of the properties exhibited by hierarchical time series; however, its non-Gaussian observation model leads to an analytically intractable inference. As a result, we derive a computationally efficient inference algorithm for DPM using the variational Bayesian expectation maximisation (VBEM) framework (Bishop, 2006). In the VBE-step, we compute the forward/backward messages at each time step to obtain the posterior over the latent states in a computationally efficient way. Further, we impose the automatic relevance determination (ARD) prior on the dynamics matrix in the latent dynamical system and infer the "effective" dimensions from the posterior over the dynamics matrix in the VBM-step. We demostrate the effectiveness of our technique on simulated data.

The paper is organized as follows. In Sec. 2, we first introduce our hierarchical Bayesian dynamic proportions model. In Sec. 3 and 4, we derive the VBEM algorithm for the proposed model. In Sec. 5, we present simulation results and finally conclude in Sec. 6.

2. Hierarchical Bayesian dynamic proportions model

2.1. Observation model

Suppose an observation of the top level at time t denoted by n_t is disaggregated into k different categories. We model the observations at the subsequent level as multinomial random variables, $\mathbf{y}_t = [y_t^1, \cdots, y_t^k]^T$ such that $n_t = \sum_{j=1}^k \mathbf{y}_t^j$. Conditioned on latent states $\mathbf{z}_t \in \mathbb{R}^k$, the likelihood of the observed data is given by

$$\begin{aligned} p(\mathbf{y}_t|\mathbf{z}_t, n_t) &= \text{Mu}(\pi(\mathbf{z}_t)|n_t), \qquad (1) \\ &= \frac{n_t!}{y_t^1! \cdots y_t^k!} \prod_{j=1}^{k} [\pi^j(\mathbf{z}_t)]^{y_t^j}, \qquad (2) \end{aligned}$$

where each proportion is given by the *softmax function* as follows:

$$\pi^j(\mathbf{z}_t) = \frac{\exp(z_t^j)}{\sum_{j=1}^k \exp(z_t^j)}. \qquad (3)$$

2.2. Latent states

The latent states in the HB-DP model evolve linearly with time:

$$p(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(A\mathbf{z}_{t-1}, \Sigma), \qquad (4)$$

where $A \in \mathbb{R}^{k\times k}$ is the dynamics matrix. The evolution noise covariance is denoted by $\Sigma \in \mathbb{R}^{k\times k}$ Therefore the parameters in our model are $\theta = \{A, \Sigma\}$.

2.3. Priors on parameters

Typically, it is expected that future proportion of a given category would depend on the current propotions of few related categories. To capture this, we impose the ARD prior on each row $\mathbf{a}_j$ of the dynamic matrix A :

$$p(\mathbf{a}_j|\boldsymbol{\alpha}) = \mathcal{N}(\mathbf{a}_j|0, \mathrm{diag}(\alpha_1, \cdots, \alpha_k)^{-1}). \quad (5)$$

This prior induces many zeros (i.e., sparse) in the estimate of A, which tells us which elements in A contribute dynamical proportion changes. If we share the hyperparameters $\boldsymbol{\alpha}$ across rows the resulting prior on the dynamic matrix A is given by

$$p(A|\boldsymbol{\alpha}) = \prod_{j=1}^{k} \mathcal{N}(\mathbf{a}_j|0, \mathrm{diag}(\boldsymbol{\alpha})^{-1}). \quad (6)$$

This prior on A could be less flexible than an independent ARD prior on each row of A using different precisions for each row, i.e., $\mathbf{a}_j \sim \mathcal{N}(0, \mathrm{diag}(\boldsymbol{\alpha}_j)^{-1})$. Our choice for the shared precisions across rows of A is based on that: (1) having too many hyperparameters can be harmful due to over-fitting; (2) our interest is not finding maximally sparse A, but finding which dimension in the latent variables contributes proportion changes in the time series.

We impose the Gaussian prior on the initial latent states:

$$p(\mathbf{z}_0) = \mathcal{N}(\mathbf{z}_0|\boldsymbol{\mu}_0, \Sigma_0). \quad (7)$$

In total, the hyperparameters are $\phi = \{\boldsymbol{\alpha}, \boldsymbol{\mu}_0, \Sigma_0\}$.

2.4. Approximate posterior

We assume the approximate posterior over the parameter θ and latent variables, $q(\theta, \mathbf{z}_{0:T})$, is factorized in the following way:

$$p(\theta, \mathbf{z}_{0:T}|\mathbf{y}_{1:T}) \approx q(\theta, \mathbf{z}_{0:T}) = q_\theta(\theta) q_{\mathbf{x}}(\mathbf{z}_{0:T}). \quad (8)$$

where the approximate posterior over the parameters is further factorized as

$$\begin{aligned} q_\theta(\theta) &= q_{A|\Sigma}(A|\Sigma) q_\Sigma(\Sigma), & (9)\\ &= q_{A|\Sigma}(A|\Sigma)\delta_\Sigma(\Sigma - \Sigma_{ML}) & (10)\end{aligned}$$

We approximate the posterior where Σ coincides the ML estimate of Σ for simplicity.

2.5. Variational lower bound

Using the approximate posterior, we can lower bound the marginal likelihood of the observations by the KL divergence between the approximate posterior $q(\theta, \mathbf{z}_{0:T})$ and the true posterior $p(\theta, \mathbf{z}_{0:T}, \mathbf{y}_{1:T})$:

$$\log p(\mathbf{y}_{1:T}) \geq \int d\theta \; d\mathbf{z}_{0:T} \; q(\theta, \mathbf{z}_{0:T}) \; \log \frac{p(\theta, \mathbf{z}_{0:T}, \mathbf{y}_{1:T})}{q(\theta, \mathbf{z}_{0:T})}.$$

We maximize the lower bound by iterating the variational Bayesian expectation maximization (VBEM) algorithm (Beal, 2003), which consists of : (1) variational Bayesian expectation (VBE) step for computing $q_{\mathbf{z}}(\mathbf{z}_{0:T})$:

$$q_{\mathbf{z}}(\mathbf{z}_{0:T}) \propto \exp\left[\int d\theta q_\theta(\theta) \log p(\mathbf{z}_{0:T}, \mathbf{y}_{1:T}|\theta)\right], \quad (11)$$

and (2) variational Bayesian maximization (VBM) step for computing $q_\theta(\theta)$:

$$q_\theta(\theta) \propto p(\theta) \exp\left[\int d\mathbf{z}_{0:T} q_{\mathbf{z}_{0:T}}(\mathbf{z}_{0:T}) \log p(\mathbf{z}_{0:T}, \mathbf{y}_{1:T}|\theta)\right]. \quad (12)$$

In each iteration, we also update the hyperparameters by computing the derivatives of the lower bound given $q(\theta, \mathbf{z}_{0:T})$ with respect to each hyperparameter.

3. Variational Bayesian EM

3.1. VBE step

In VBE step, we compute

$$\log q_{\mathbf{z}}(\mathbf{z}_{0:T}) = \mathbb{E}_{q_\theta(\theta)} \log p(\mathbf{z}_{0:T}, \mathbf{y}_{1:T}|\theta) + const, \quad (13)$$

where the integrand in eq. 13, the so-called *complete-data log likelihood*, is written by

$$\sum_{t=1}^{T} \{\log p(\mathbf{y}_t|\mathbf{z}_t) + \log p(\mathbf{z}_t|\mathbf{z}_{t-1}, \theta)\}, \quad (14)$$

which tells us that the log posterior over latent variables is quadratic in each $\mathbf{z}_t$. This enables us to use the sequential forward/backward message passing algorithm (see Fig. 1) to compute the posterior over latent variables in the following.

Forward message (filtering)
The forward message at each time is given by

$$\begin{aligned} \alpha(\mathbf{z}_t) &\triangleq p(\mathbf{z}_t|\mathbf{y}_{1:t}),\\ &= \mu_{f_t->\mathbf{z}_t}(\mathbf{z}_t),\\ &= \int f_t(\mathbf{z}_{t-1}, \mathbf{z}_t)\alpha(\mathbf{z}_{t-1}) d\mathbf{z}_{t-1},\\ &\propto p(\mathbf{y}_t|\mathbf{z}_t) \times\\ &\quad \int \exp\left(\mathbb{E}_{q_\theta(\theta)} \log p(\mathbf{z}_t|\mathbf{z}_{t-1})\right) \alpha(\mathbf{z}_{t-1}) d\mathbf{z}_{t-1}, \end{aligned}$$

where we approximate $\alpha(\mathbf{z}_{t-1})$ to a Gaussian:

$$\alpha(\mathbf{z}_{t-1}) \approx \mathcal{N}(\boldsymbol{\mu}_{t-1}, V_{t-1}), \quad (15)$$

and assume $\alpha(\mathbf{z}_0) = \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$. After some algebraic manipulation, the forward message is given by

$$\alpha(\mathbf{z}_t) \approx p(\mathbf{y}_t|\mathbf{z}_t)\mathcal{N}(\mathbf{z}_t|\tilde{\boldsymbol{\mu}}_t, \tilde{V}_t), \quad (16)$$

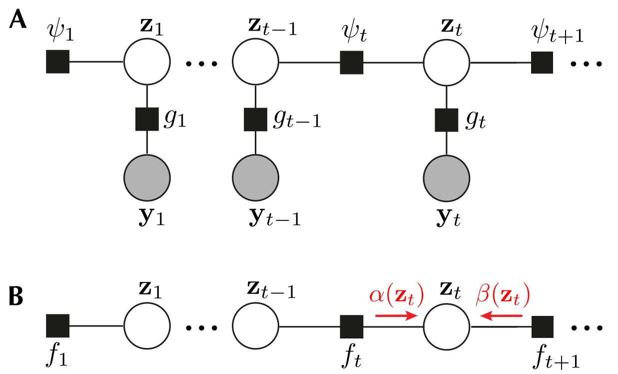

Figure 1. **Factor graph representation of dynamic proportion model**. **A**: A fragment of the factor graph showing both latent and observed variables. **B**: A simplified factor graph by absorbing the emission probabilities into the transition probability factors.

where the mean and covariance are given by

$$\tilde{V}_t^{-1} = \Sigma + \langle A \rangle V_{t-1} \langle A \rangle^\top, \tag{17}$$
$$\tilde{\boldsymbol{\mu}}_t = \langle A \rangle \boldsymbol{\mu}_{t-1}, \tag{18}$$

where the mean of A w.r.t. $q_\theta(\theta)$ is denoted by $\langle A \rangle$.

Due to the multinomial likelihood term, eq. 16 is not Gaussian in $\mathbf{z}_t$. We approximate $\alpha(\mathbf{z}_t)$ as a Gaussian by finding the first two moments of eq. 16. The first derivative expression gives us the mean update

$$\boldsymbol{\mu}_t = \tilde{\boldsymbol{\mu}}_t + \tilde{V}_t[\mathbf{y}_t - n_t \pi(\boldsymbol{\mu}_t)], \tag{19}$$

where $\boldsymbol{\mu}_t$ appears in both sides. Iterative methods, e.g., Newton's method, are typically used to solve this equation. (See appendix A for details). The second derivative expression gives us the covariance update

$$V_t^{-1} = W_t + \tilde{V}_t^{-1}, \tag{20}$$

where $W_t = n_t[\mathrm{diag}(\pi(\boldsymbol{\mu}_t)) - \pi(\boldsymbol{\mu}_t)\pi(\boldsymbol{\mu}_t)^T]$. The two updates above yield $\alpha(\mathbf{z}_t) \approx \mathcal{N}(\boldsymbol{\mu}_t, V_t)$.

Backward message (smoothing)
The backward message at each time is given by

$$\begin{aligned}
\beta(\mathbf{z}_t) &\triangleq p(\mathbf{y}_{t+1:T}|\mathbf{z}_t), \\
&= \mu_{f_{t+1} -> \mathbf{z}_t}(\mathbf{z}_t), \\
&= \int f_{t+1}(\mathbf{z}_t, \mathbf{z}_{t+1}) \beta(\mathbf{z}_{t+1}) d\mathbf{z}_{t+1}, \\
&\propto \int p(\mathbf{y}_{t+1}|\mathbf{z}_{t+1}) \exp\left(\mathbb{E}_{q_\theta(\theta)} \log p(\mathbf{z}_{t+1}|\mathbf{z}_t)\right) \\
&\qquad \beta(\mathbf{z}_{t+1}) d\mathbf{z}_{t+1},
\end{aligned} \tag{21}$$

where we assume $\beta(\mathbf{z}_{t+1}) \approx \mathcal{N}(\boldsymbol{\eta}_{t+1}, \Gamma_{t+1})$. The initial values for $(\boldsymbol{\eta}_T, \Gamma_T)$ are the mean and covariance $(\boldsymbol{\mu}_T, V_T)$ of the last forward message. Unfortunately, the integral in eq. 21 is analytically intractable due to the multinomial likelihood term. Here, we approximate the integrand of the above equation as a joint Gaussian in $[\mathbf{z}_t \ \mathbf{z}_{t+1}]^T$, and extract those parts that correspond to $\mathbf{z}_t$ to approximately compute the integral.

The first derivative of the logarithm of the integrand denoted by $\Phi(\mathbf{z}_t, \mathbf{z}_{t+1})$ is given by

$$\begin{aligned}
\frac{\partial \Phi(\mathbf{z}_t, \mathbf{z}_{t+1})}{\partial [\mathbf{z}_t \ \mathbf{z}_{t+1}]} &= \begin{bmatrix} -\langle A \rangle^\top \Sigma^{-1} \langle A \rangle & \langle A \rangle^\top \Sigma^{-1} \\ \Sigma^{-1} \langle A \rangle & -(\Sigma^{-1} + \Gamma_{t+1}^{-1}) \end{bmatrix} \begin{bmatrix} \mathbf{z}_t \\ \mathbf{z}_{t+1} \end{bmatrix} \\
&+ \begin{bmatrix} \mathbf{0} \\ \mathbf{y}_{t+1} - n_{t+1}\pi(\mathbf{z}_{t+1}) + \Gamma_{t+1}^{-1} \boldsymbol{\eta}_{t+1} \end{bmatrix},
\end{aligned}$$

which we use to compute the joint mode of $[\mathbf{z}_t \ \mathbf{z}_{t+1}]^T$.

The second derivative of $\Phi(\mathbf{z}_t, \mathbf{z}_{t+1})$ is given by

$$\frac{\partial^2 \Phi(\mathbf{z}_t, \mathbf{z}_{t+1})}{\partial [\mathbf{z}_t \ \mathbf{z}_{t+1}]^2} = -\begin{bmatrix} \langle A \rangle^\top \Sigma^{-1} \langle A \rangle & -\langle A \rangle^\top \Sigma^{-1} \\ -\Sigma^{-1} \langle A \rangle & \Sigma^{-1} + W_{t+1} + \Gamma_{t+1}^{-1} \end{bmatrix}. \tag{22}$$

where we denote $W(\hat{\mathbf{z}}_{t+1})$ by W_{t+1}. Using Schur complement, we obtain the covariance Γ_t by

$$\Gamma_t^{-1} = \langle A \rangle^\top \Sigma^{-1} \langle A \rangle - \langle A \rangle^\top \Sigma^{-1} \Gamma_{t+1}^{*} \Sigma^{-1} \langle A \rangle.$$

where $\Gamma_{t+1}^{*-1} = \Sigma^{-1} + \Gamma_{t+1}^{-1} + W_{t+1}$.

Computing posterior marginals

Using the forward and backward messages, we can compute the posterior marginals for the latent variables. First, we define

$$\begin{aligned}
\gamma(\mathbf{z}_t) &\triangleq p(\mathbf{z}_t|\mathbf{y}_{1:T}), \\
&\propto p(\mathbf{z}_t|\mathbf{y}_{1:t}) p(\mathbf{y}_{t+1:T}|\mathbf{z}_t) = \alpha(\mathbf{z}_t)\beta(\mathbf{z}_t), \\
&\propto \mathcal{N}(\mathbf{z}_t|\hat{\boldsymbol{\mu}}_t, \hat{V}_t),
\end{aligned} \tag{23}$$

where the mean and covariance are given by

$$\hat{\boldsymbol{\mu}}_t = \hat{V}_t (V_t^{-1} \boldsymbol{\mu}_t + \Gamma_t^{-1} \boldsymbol{\eta}_t), \tag{24}$$
$$\hat{V}_t = (V_t^{-1} + \Gamma_t^{-1})^{-1}. \tag{25}$$

Second, we also define the joint posterior between neighboring (in time) latent variables

$$\begin{aligned}
\xi(\mathbf{z}_{t-1}, \mathbf{z}_t) &\triangleq p(\mathbf{z}_{t-1}, \mathbf{z}_t|\mathbf{y}_{1:T}), \\
&\propto \alpha(\mathbf{z}_{t-1}) p(\mathbf{y}_t|\mathbf{z}_t) \\
&\quad \exp\left(\mathbb{E}_{q_\theta(\theta)} \log p(\mathbf{z}_t|\mathbf{z}_{t-1})\right) \beta(\mathbf{z}_t).
\end{aligned}$$

We approximate the joint distribution over $\mathbf{z}_{t-1}$ and $\mathbf{z}_t$ to a Gaussian. The second derivative of the logarithm of $\xi(\mathbf{z}_{t-1}, \mathbf{z}_t)$ is given by

$$-\begin{bmatrix} \langle A \rangle^\top \Sigma^{-1} \langle A \rangle + V_{t-1}^{-1} & -\langle A \rangle^\top \Sigma^{-1} \\ -\Sigma^{-1} \langle A \rangle & \Gamma_t^{*-1} \end{bmatrix}, \tag{26}$$

where we denote $W(\hat{\boldsymbol{\mu}}_t)$ by W_t. Using the Schur complement, we can obtain the cross covariance of $(\mathbf{z}_{t-1}, \mathbf{z}_t)$.

3.2. VBM step

In VBM step, we compute $q_\theta(\theta)$ by extracting all the terms in $\log p(\mathbf{z}_{0:T}, \mathbf{y}_{1:T}|\theta)$ that depend on θ and then taking the expectation over $\mathbf{z}_{0:T}$:

$$\log q_\theta(\theta) = \mathbb{E}_{q_{\mathbf{z}}(\mathbf{z}_{0:T})}\left[\log p(\mathbf{z}_{0:T}|\theta)\right] + \log p(\theta) + const.$$

where the first term on RHS is given by

$$-\tfrac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}\sigma_{ij}^2(\mathbf{a}_i^T W_A \mathbf{a}_j - 2\mathbf{a}_i^T S_A \mathbf{e}_j) \qquad (27)$$

where σ_{ij} is the i,jth element of ML estimate[1] of the latent noise covariance Σ, and $\mathbf{e}_j$ is the unit vector where the jth element is 1. The sufficient statistics of latent variables are denoted by W_A and S_A:

$$W_A = \sum_{t=1}^{T} < \mathbf{z}_{t-1}\mathbf{z}_{t-1}^T >, \quad S_A = \sum_{t=1}^{T} < \mathbf{z}_{t-1}\mathbf{z}_t^T > .$$

Clearly, the joint posterior over the rows of A does not factorize. However, to avoid computational burden finding the joint posterior for large data, we approximate the posterior over A as[2]

$$q_\theta(\theta) \quad = \quad \prod_{j=1}^{T}\mathcal{N}(\mathbf{a}_j|\boldsymbol{\mu}_{\mathbf{a}_j}, \Sigma_A),$$

where the covariance and mean are given by

$$\Sigma_A^{-1} \quad = \quad W_A + \mathrm{diag}(\boldsymbol{\alpha}), \qquad (28)$$
$$\boldsymbol{\mu}_{\mathbf{a}_j} \quad = \quad \Sigma_A S_A \mathbf{e}_j. \qquad (29)$$

3.3. Hyperparameter estimation

We update hyperparameters so as to maximize the variational lower bound on the marginal likelihood (eq. 11). The lower bound can be simplified as[3]

$$\log p(\mathbf{y}_{1:T}|\phi) \geq \log Z' - KL(q(\theta)||p(\theta)), \qquad (30)$$

where

$$Z' = \int d\mathbf{z}_{0:T}\exp\left(\mathbb{E}_{q_\theta(\theta)}\log p(\mathbf{z}_{0:T}, \mathbf{y}_{1:T}|\theta)\right). \qquad (31)$$

The KL divergence between $q(\theta)$ and $p(\theta)$ is given by

$$\sum_{j=1}^{k}\int d\mathbf{a}_j \mathcal{N}(\mathbf{a}_j|\boldsymbol{\mu}_{\mathbf{a}_j}, \Sigma_A)\log\frac{\mathcal{N}(\mathbf{a}_j|\boldsymbol{\mu}_{\mathbf{a}_j}, \Sigma_A)}{\mathcal{N}(\mathbf{a}_j|\mathbf{0}, \mathrm{diag}(\boldsymbol{\alpha}^{-1}))},$$
$$= \sum_{j=1}^{k}(-\tfrac{1}{2}\log|\mathrm{diag}(\boldsymbol{\alpha})\Sigma_A| +$$
$$\tfrac{1}{2}\mathrm{Tr}\left[\mathrm{diag}(\boldsymbol{\alpha})(\Sigma_A - \mathrm{diag}(\boldsymbol{\alpha})^{-1} + \boldsymbol{\mu}_{\mathbf{a}_j}\boldsymbol{\mu}_{\mathbf{a}_j}{}^\top)\right]).$$

[1]The formula for ML estimate of the noise covariance is given in (Bishop, 2006)

[2]This corresponds to assuming the latent noise covariance to be diagonal.

[3]See Ch.5 in (Beal, 2003) for derivation in detail

The first derivative expression of $\boldsymbol{\alpha}$ gives us the following update:

$$\boldsymbol{\alpha}^{-1} \quad = \quad \tfrac{1}{k}\mathrm{diag}\left[\sum_{j=1}^{k}(\Sigma_A + \boldsymbol{\mu}_{\mathbf{a}_j}\boldsymbol{\mu}_{\mathbf{a}_j}{}^\top)\right]. \qquad (32)$$

Similarly, we update the hyperparameters for initial states by

$$\boldsymbol{\mu}_0 \quad = \quad \omega_0, \qquad (33)$$
$$\Sigma_0 \quad = \quad \Upsilon_{0,0}. \qquad (34)$$

The summary of the entire algorithm is given below:

Algorithm 1 VBEM for dynamic proportion models

Given data $\mathcal{D}$ and initial $q(\theta)$, iterate the following:

1. VBE step: Given $q(\theta)$, compute forward (α), backward (β) and marginal (γ) messages and the cross-covariation of messages at each time.
2. VBM Step: Given $q(\mathbf{z}_{0:T})$, update $q(\theta)$.
3. Update hyperparameters.

Until convergence.

4. Prediction

Given $p(\mathbf{z}_T|\mathbf{y}_{1:T}) \approx \mathcal{N}(\hat{\boldsymbol{\mu}}_T, \hat{V}_T)$, we want to make a prediction on the time series in each level of hierarchy by

$$p(\mathbf{y}_{T+1}|\mathbf{y}_{1:T}) = \int p(\mathbf{y}_{T+1}|\mathbf{z}_{T+1})p(\mathbf{z}_{T+1}|\mathbf{y}_{1:T})\mathrm{d}\mathbf{z}_{T+1}, \quad (35)$$

where the second part of the integrand is

$$p(\mathbf{z}_{T+1}|\mathbf{y}_{1:T}) \quad = \quad \int \exp\left(\mathbb{E}_{q_\theta(\theta)}\log p(\mathbf{z}_{T+1}|\mathbf{z}_T)\right)$$
$$\mathcal{N}(\mathbf{z}_T|\hat{\boldsymbol{\mu}}_T, \hat{V}_T)\, d\mathbf{z}_T, \qquad (36)$$
$$= \quad \mathcal{N}(\mathbf{z}_{T+1}|\tilde{\boldsymbol{\mu}}_{T+1}, \tilde{V}_{T+1}) \qquad (37)$$

where the mean and covariance are given by

$$\tilde{V}_{T+1}^{-1} \quad = \quad \Sigma + \langle A\rangle\hat{V}_T\langle A\rangle^\top,$$
$$\tilde{\boldsymbol{\mu}}_{T+1} \quad = \quad \langle A\rangle\hat{\boldsymbol{\mu}}_T.$$

The integral in eq. 35 is not analytically tractable due to the non-Gaussian likelihood term. One can do is to draw samples of $\mathbf{z}_{T+1}^i$ from eq. 36, and approximate

$$p(\mathbf{y}_{T+1}|\mathbf{y}_{1:T}) \quad \approx \quad \sum_i p(\mathbf{y}_{T+1}|\mathbf{y}_{1:T}, \mathbf{z}_{T+1}^i). \quad (38)$$

It is straightforward to extend this to multiple steps ahead prediction.

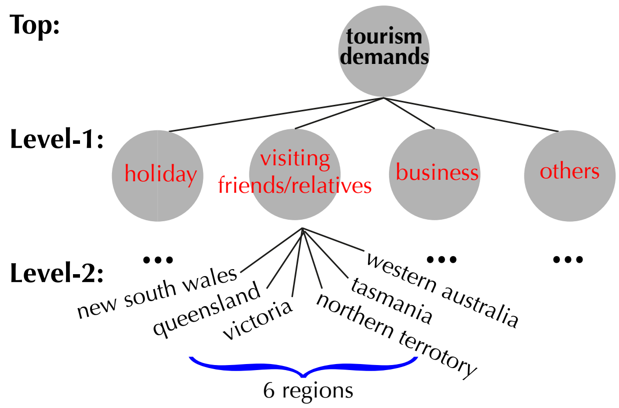

Figure 2. **Example of two-level hierarchical time series: Australian domestic tourism demands**. Total domestic tourism demands are in the top level, which are disaggregated by four different purpose of travel: holiday, visiting friends/relatives, business, and others. Each of Level-1 time series is then disaggregated by 6 different regions: New South Wales, Queensland, Victoria, Western Australia, Tasmania and the Northern territory.

Table 1. **Forecasting performance (MAPE)** Our method (DPM) outperforms other methods. The DPM achieved the lowest average MAPE across the three levels, 22.80. The second best method (TD) achieved average MAPE 28.78.

METHOD	Q1	Q2	Q3	Q4	AVG
TOP LEVEL					
INDEP (TD/DPM)	9.32	34.01	26.58	20.17	22.52
BU	4.14	42.83	29.84	23.10	24.98
OPTIMAL	6.13	34.92	26.58	19.98	21.90
LEVEL-1					
INDEP	14.07	50.76	38.99	33.50	33.58
TD	13.36	38.65	31.46	28.05	27.88
BU	10.86	49.19	37.51	32.27	32.46
OPTIMAL	13.55	41.56	33.56	29.72	29.60
DPM	9.36	30.43	23.78	20.86	**21.11**
LEVEL-2					
INDEP	26.11	58.91	46.52	40.51	43.01
TD	24.62	47.16	38.15	33.87	35.95
BU	26.11	58.91	46.52	40.51	43.01
OPTIMAL	27.21	52.00	41.90	37.00	39.53
DPM	10.23	29.96	27.16	31.74	**24.77**

5. Experiments

We apply our method to forecasting time series using data generated from a toy two-level hierarchical network given in Fig. 2. This network represents the hierarchy present in Australian domestic tourism data. This data is an indicator of tourism activity: the number of visitor nights per quarter consists of time series. There are two levels of hierarchy in the data as shown in Fig. 2. The aggregated domestic tourism demand for the entire Australia consists of the top level time series. At (sub) level 1, the top level time series is disaggregated by four different purpose of travel: Holiday, Visiting friends and relatives, Business, and Others. At (sub) level 2, the level 1 time series is disaggregated by seven different states and territories they visited: New South Wales, Queensland, Victoria, Western Australia, Tasmania, and the Northern Territory. Therefore, there are 4 time series at level 1, 24 time series at level 2.

We generated each time series from the AR-1 processes with random parameters. To make sure the sum of sub-level time series matches the upper-level time series, we divide the sub-level time series by the sum of sub-level time series. This gives us a proportion at the sub-level, which we multiply by the upper-level time series to obtain the revised sub-level time series.

For forecasting individual time series independently, we used an ARMA model using the automatic algorithm developed in (Hyndman & Khandakar, 2008). We used the first 120 observations (1980:Q1-2010:Q4) as a training set and predicted tourism demands up to 4-steps ahead (i.e., Q1, Q2, Q3, and Q4 in 2011). We computed the MAPE (mean absolute percentage error) values from these results and computed the average in Table 1. For the top-down method (TD), we used the forecasted proportion (FP) method. For the bottom-up method (BU), we used the simple summation of the individual forecasts at the bottom level. As shown in Table 1, our method outperforms other methods.

6. Discussion

In this paper, we modeled the proportion changes of hierarchically structured time series using linear dynamical systems with multinomial observations. We developed the variational Bayesian expectation maximization algorithm for posterior inference and parameter estimation. The sequential forward/backward type algorithm allows us to parallelize the posterior inference, which would be beneficial when dealing with large datasets. Simulation results on a toy dataset show the effectiveness of our approach.

A potential criticism of our approach would be that our method highly relies on top level prediction. However, top level time series are often periodic and less abruptly changing over time (since they are sum of many sub-level time series) compared to individual time series at the bottom level. The prediction performance at the top level typically outperforms sub-level prediction (Table 1 shows the same trend). It would be interesting to test our method to large hierarchical time series datasets in future work.

References

Athanasopoulos, George, Ahmed, Roman A., and Hyndman, Rob J. Hierarchical forecasts for australian domestic tourism. *International Journal of Forecasting*, 25(1): 146 – 166, 2009.

Beal, M.J. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

Bishop, Christopher M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., 2006.

Dangerfield, Byron J. and Morris, John S. Top-down or bottom-up: Aggregate versus disaggregate extrapolations. *International Journal of Forecasting*, 8(2):233 – 241, 1992.

Fischer, U., Schildt, C., Hartmann, C., and Lehner, W. Forecasting the data cube: A model configuration advisor for multi-dimensional data sets. pp. 853–864, 2013.

Fliedner, G. An investigation of aggregate variable time series forecast strategies with specific subaggregate time series statistical correlation. *Computers and Operations Research*, 26(10-11):1133–1149, 1999.

Gross, Charles W. and Sohl, Jeffrey E. Disaggregation Methods to Expedite Product Line Forecasting. *Journal of Forecasting*, 9(1):233–254, 1990.

Hyndman, R.J., Ahmed, R.A., Athanasopoulos, G., and Shang, H.L. Optimal combination forecasts for hierarchical time series. *Computational Statistics and Data Analysis*, 55(9):2579–2589, 2011.

Hyndman, Rob J. and Khandakar, Yeasmin. Automatic time series forecasting: The forecast package for r. *Journal of Statistical Software*, 27(3):1–22, 7 2008. ISSN 1548-7660.

Kalchschmidt, M., Verganti, R., and Zotteri, G. Forecasting demand from heterogeneous customers. *International Journal of Operations and Production Management*, 26 (6):619–638, 2006.

Kerkkaenen, A., Korpela, J., and Huiskonen, J. Demand forecasting errors in industrial context: Measurement and impacts. *International Journal of Production Economics*, 118(1):43–48, 2009.

Moon, S., Hicks, C., and Simpson, A. The development of a hierarchical forecasting method for predicting spare parts demand in the south korean navy - a case study. *International Journal of Production Economics*, 140(2): 794–802, 2012.

Moon, S., Simpson, A., and Hicks, C. The development of a classification model for predicting the performance of forecasting methods for naval spare parts demand. *International Journal of Production Economics*, 143(2):449–454, 2013.

Sbrana, Giacomo and Silvestrini, Andrea. Forecasting aggregate demand: Analytical comparison of top-down and bottom-up approaches in a multivariate exponential smoothing framework. *International Journal of Production Economics*, 146(1):185 – 198, 2013.

Zellner, Arnold and Tobias, Justin. A note on aggregation, disaggregation and forecasting performance. *Journal of Forecasting*, 19(5):457–465, 2000.

Zotteri, G. and Kalchschmidt, M. A model for selecting the appropriate level of aggregation in forecasting processes. *International Journal of Production Economics*, 108(1-2):74–83, 2007.