The Switching Hierarchical Gaussian Filter

İsmail Şenöz, Albert Podusenko, Semih Akbayrak
TU Eindhoven, the Netherlands
{i.senoz,a.podusenko,s.akbayrak}@tue.nl

Christoph Mathys
Aarhus University, DK
chmathys@cas.au.dk

Bert de Vries
GN Hearing & TU Eindhoven, NL
bert.de.vries@tue.nl

2021 IEEE International Symposium on Information Theory (ISIT) | 978-1-5386-8209-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISIT45174.2021.9518229

***Abstract*—In this paper we discuss variational message passing-based (VMP) inference in a switching Hierarchical Gaussian Filter (HGF). An HGF is a flexible hierarchical state space model that supports closed-form VMP-based approximate inference for tracking of both states and slowly time-varying parameters. Since natural signals often submit to regime-switching dynamics, there is a need for low-complexity closed-form inference in switching state space models. Here we extend the HGF model with parameter switching mechanics and derive closed-form VMP update rules for plug-in applications in factor graph-based models. These VMP rules support both tracking of latent variables and variational free energy as a model performance measure. We show that the switching HGF performs better than a non-switching HGF on modelling of a stock market data set.**

I. Introduction

Hierarchical Dynamic Models (HDM) have often been used to explain the variation of parameters and states of natural processes [1]–[5]. The Hierarchical Gaussian filter (HGF) is a specific type of HDM that is popular in the neuroscience community, which is partly due to the availability of an open source modelling toolbox [6]–[8]. The HGF is a multi-layer nonlinear state space model where the variance of state transitions at a particular layer is controlled by the states at a higher layer. In the literature, parameters and hidden states of the HGF can be recovered by closed-form variational message passing updates. These properties makes the HGF an interesting model for modeling of natural signals [9]–[11].

However, in many practical applications the observed signal can be subject to Markovian regime-switching behavior [12], [13]. The "classical" HGF model will fail to accurately describe a time series when the underlying dynamics are ruled by parameter regime switches.

While it is not difficult to describe the forward mechanics of regime-switching behavior in a generative model, inference for states and parameters in these models is problematic. In [14], a switching state space model (SSSM) that employs a variational inference technique for tracking the posterior of the hidden states was introduced. This work takes a pivotal position in the literature and was followed by diverse further developments on state inference for SSSMs [15]–[18]. Examples include efficient Gaussian Sum Filtering to track a Gaussian Mixture state posterior [19, Ch. 25] and Rao-Blackwellised particle filters for state tracking by analytical marginalization of continuous variables conditioned on sampled discrete latent variables [20].

In this paper, we develop a state and parameter inference framework for a *Switching* Hierarchical Gaussian Filter (SHGF) that extends the original HGF by supporting a selector mechanism for the model's parameters. The SHGF is a complex generative model that features hierarchical regime-switching dynamics, together with non-linear couplings between the layers. Since our target applications require real-time inference on wearable devices, we are interested in developing closed-form inference updates for states and (both slowly time-varying and regime-switching) parameters, along with tracking of a Bayesian evidence performance measure. Inference by Monte Carlo sampling is computationally too expensive for these applications. We build on previous work for the HGF by representing the model as a factor graph and execute message passing-based inference via divergence minimization [10], [11]. The contributions of this paper include:

- In Section II, we present a new switching hierarchical dynamical model, the SHGF. We map the SHGF onto a Forney-style Factor Graph (FFG), which supports a fully modular message passing-based approach to inference.
- In Section III, we identify and isolate a "Gaussian with controlled switching variance (GCSV) node" as the module that causes inference issues.
- In Section IV-B, we derive new variational update rules for the GCSV node and combine these rules with Expectation Propagation algorithm to show that the non-conjugate operations can be handled by quadrature based moment-matching [21] yielding a hybrid algorithm [22].
- We experimentally verify the proposed inference procedure on synthetic data for a 2-layer SHGF in Section V. We also provide a real-world example on a stock market data set where we compare the SHGF to a (non-switching) HGF model.

II. Model Specification

Let $y_t \in \mathbb{R}$ represent observations. We denote latent continuously valued states at layer i by $x_t^{(i)} \in \mathbb{R}$ and categorical states by $s_t^{(i)} \in \{1, \ \dots \ , M_i\}$. State transitions of categorical variables are governed by transition matrices $\boldsymbol{A}^{(i)} \in \mathbb{R}^{M_i \times M_i}$ and continuous state transitions are parameterized by $\boldsymbol{\kappa}^{(i)} \in \mathbb{R}^{M_i}$ and $\boldsymbol{\omega}^{(i)} \in \mathbb{R}^{M_i}$.

One layer of an N-layer switching hierarchical Gaussian filter is defined by the state transitions

$$p\left(x_t^{(i)}|x_{t-1}^{(i)}, s_t^{(i)}, g_t^{(i)}\right) = \prod_{m=1}^{M_i} \mathcal{N}\left(x_t^{(i)}|x_{t-1}^{(i)}, g_t^{(i)}\right)^{[s_t^{(i)}=m]} \tag{1a}$$

$$p\left(s_t^{(i)}|s_{t-1}^{(i)}, \boldsymbol{A}^{(i)}\right) = \prod_{k=1}^{M_i}\prod_{m=1}^{M_i} \left(\alpha_{km}^{(i)}\right)^{[s_t^{(i)}=k][s_{t-1}^{(i)}=m]} \tag{1b}$$

where we used the following definition and constraint, respectively, for every $i = 1, \ldots, N-1$:

$$g_t^{(i)}\left(x_t^{(i+1)}, \kappa_m^{(i)}, \omega_m^{(i)}\right) \triangleq \exp\left(\kappa_m^{(i)} x_t^{(i+1)} + \omega_m^{(i)}\right) \quad (2)$$

$$\sum_{k=1}^{M_i} \alpha_{km}^{(i)} = 1\,. \quad (3)$$

We use Iverson bracket notation in (1), which is defined as

$$[s_t = m] = \begin{cases} 1 & \text{if } s_t = m \text{ is true} \\ 0 & \text{otherwise} \end{cases}.$$

A non-switching HGF layer is recovered for $M_i = 1$. At the top layer ($i = N$), we assume non-switching random walk dynamics with transition variance ξ:

$$p\left(x_t^{(N)} | x_{t-1}^{(N)}\right) = \mathcal{N}\left(x_t^{(N)} | x_{t-1}^{(N)}, \xi\right). \quad (4)$$

While other likelihood functions are compatible with the SHGF, for simplicity we will assume that observations are generated by a Gaussian likelihood from the first (bottom) layer hidden states with variance τ:

$$p\left(y_t | x_t^{(1)}\right) = \mathcal{N}\left(y_t | x_t^{(1)}, \tau\right). \quad (5)$$

As (2) shows, the essential characteristic of an HGF model is that the variance of state transitions $g_t^{(i)}$ for the continuously valued states at layer i are controlled by a non-linear mapping of the continuously valued state at layer $i+1$. In the extension to a *switching* HGF, the m^{th} component of the parameters $\boldsymbol{\kappa}^{(i)}$ and $\boldsymbol{\omega}^{(i)}$ of the nonlinear transformation (2) are selected by a discrete categorical state $s_t^{(i)} = m$ that evolves according to Markovian dynamics given by (1b). After selection of the component of parameters $\boldsymbol{\kappa}^{(i)}$ and $\boldsymbol{\omega}^{(i)}$, the corresponding transition in (1a) is selected by the categorical variable. Columns of transition matrices $\boldsymbol{A}^{(i)}$ define probability distributions that lie in $M_i - 1$ dimensional simplex (3).

A Forney-style factor graph (FFG) representation that corresponds to the SHGF model and a description of the graphical notation is given in Figure 1. An FFG is a representation of a global factorized function, where nodes correspond to factors and edges correspond to variables [23] [24]. For a detailed explanation of the FFG formalism we refer to [23], [24].

III. Problem Statement

For a given SHGF model m and collection of data $\boldsymbol{y} \triangleq \boldsymbol{y}_{1:T} = [y_1 \ldots y_T]$, we are interested in obtaining the posterior distributions for every layer i for the states $p(x_t^{(i)}|\boldsymbol{y})$, $p(s_t^{(i)}|\boldsymbol{y})$, and parameters $p(\boldsymbol{\kappa}^{(i)}|\boldsymbol{y})$, $p(\boldsymbol{\omega}^{(i)}|\boldsymbol{y})$, $p(\boldsymbol{A}^{(i)}|\boldsymbol{y})$. Furthermore, to score model performance, we are interested in computing Bayesian evidence $p(\boldsymbol{y}|m)$.

To make matters concrete, suppose that we are interested in obtaining $p(x_t^{(i)}|\boldsymbol{y})$, then the corresponding Bayesian smoothing equations are given by [25]

$$p\left(x_t^{(i)}|\boldsymbol{y}\right) = p(x_t^{(i)}|\boldsymbol{y}_{1:t}) \int \frac{p(x_{t+1}^{(i)}|x_t^{(i)}) p(x_{t+1}^{(i)}|\boldsymbol{y})}{p(x_{t+1}^{(i)}|\boldsymbol{y}_{1:t})} \mathrm{d}x_{t+1}^{(i)} \quad (6)$$

where the filtering equation is evaluated as

$$p\left(x_t^{(i)}|\boldsymbol{y}_{1:t}\right) = \frac{p\left(x_t^{(i)}|\boldsymbol{y}_{1:t-1}\right) p\left(x_t^{(i)}|y_t\right)}{\int p\left(x_t^{(i)}|\boldsymbol{y}_{1:t-1}\right) p\left(x_t^{(i)}|y_t\right) \mathrm{d}x_t^{(i)}}. \quad (7)$$

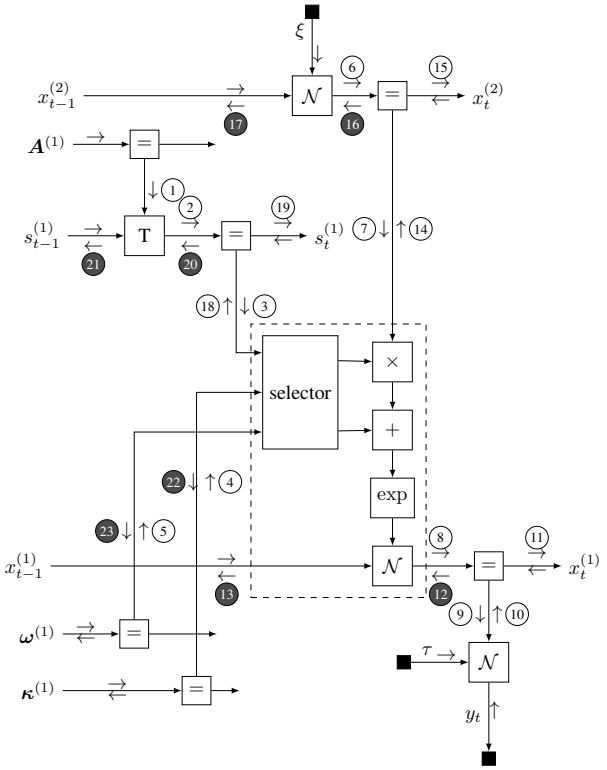

Fig. 1: One time segment of an FFG corresponding to a 2-layer SHGF model. Nodes represent the factors. An abbreviation of the underlying functional form of the factors are given in the nodes. Small dark squares indicate an observation constraint and they send point mass messages. Arrows represent messages flowing at an edge. Circled numbers indicate a computation schedule that can be chosen differently. A dark circle refers to a backwards message (from observations to latent variables). If dark messages are set proportional to 1 (i.e., uninformative), then this message passing schedule results in filtering, otherwise the schedule leads to smoothing. The selector node chooses the components of $\boldsymbol{\kappa}^{(1)}$ and $\boldsymbol{\omega}^{(1)}$, depending on the value of the categorical selector. The output of the selector node are then passed to the non-linearity block (2). The dashed box corresponds to the "composite" GCSV node $f_t^{(i)}$ defined by (8c).

The filtering process (7) requires evaluating

$$p\left(x_t^{(i)}|\boldsymbol{y}_{1:t-1}\right) = \int p\left(x_t^{(i)}|x_{t-1}^{(i)}\right) p\left(x_{t-1}^{(i)}|\boldsymbol{y}_{1:t-1}\right) \mathrm{d}x_{t-1}^{(i)}$$

$$p\left(x_t^{(i)}|y_t\right) = \mathbb{E}_{\setminus\{x_t^{(i)}\}}\left[f_t^{(i-1)}\right] \quad (8a)$$

$$p\left(x_t^{(i)}|x_{t-1}^{(i)}\right) = \mathbb{E}_{\setminus\{x_t^{(i)}, x_{t-1}^{(i)}\}}\left[f_t^{(i)}\right] \quad (8b)$$

$$f_t^{(i)} \triangleq p\left(x_t^{(i)}|x_{t-1}^{(i)}, s_t^{(i)}, \boldsymbol{\kappa}^{(i)}, \boldsymbol{\omega}^{(i)}, x_t^{(i+1)}\right) \quad (8c)$$

and the factor $f_t^{(i)}$ in (8c) is further specified by (1a) and (2). Note that computing the smoothing posterior (6) involves computing the filtering posterior (7), which in turn involves an exponentially growing number of summation terms due to the expectations in (8a) and (8b) with respect to categorical states. For example, if there are M categories, then there will be M indexed Gaussians at time $t = 1$, M^2 Gaussians at $t = 2$, and M^k Gaussians at $t = k$ [19, Ch. 25]. This explosion of terms make inference intractable. In addition, the non-linear couplings between the continuous state transitions in (8c) cause the complexity of functional dependencies for evaluating (8a) and (8b) to grow quickly. In short, the smoothing and filtering solutions (6) and (7) are not analytically tractable. In this work, we address approximating the filtering (7) and smoothing solutions (6) for the SHGF model. Above we iden-

tified the non-linearities inside the factor $f_t^{(i)}$ and expectations over internal categorical states in this factor as the problematic issues. We call this factor $f_t^{(i)}$ a "Gaussian with Controlled Switching Variance" (GCSV), see the dashed box in Figure 1.

Our solution to smoothing and filtering relies on exploiting the factorized structure of the SHGF model and uses variational message passing-based inference on factor graphs. This method supports solving the inference issues inside the GCSV node in isolation and then use the GCSV factor as a plug-in node in any factor graph, including the graph for the SHGF model. To this end, next we introduce factor graphs and variational message passing.

IV. Variational Message Passing

A. Variational Optimization

Consider a generative model over observations $\boldsymbol{y}$ and latent variables $\boldsymbol{z} = \{\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{\kappa}, \boldsymbol{\omega}, \boldsymbol{A}\}$ that has been specified in the (prior-times-likelihood) form as

$$p(\boldsymbol{y}, \boldsymbol{z}) = p(\boldsymbol{y}|\boldsymbol{z})p(\boldsymbol{z})\,. \tag{9}$$

Variational methods approximate the intractable posterior $p(\boldsymbol{z}|\boldsymbol{y})$ by an instrumental distribution $q(\boldsymbol{z})$ by minimizing the Kullback-Leibler divergence criterion between both distributions [26, Ch. 10]. Because direct computation of a KL-divergence requires access to the unavailable posterior, a common practice is to minimize a free-energy functional

$$F[q] \triangleq \mathbb{E}_{\boldsymbol{z}}\left[\log \frac{q\left(\boldsymbol{z}\right)}{p(\boldsymbol{y}, \boldsymbol{z})}\right] \tag{10}$$

that is an upper-bound to the negative log-evidence and requires access to the model specification (9).[1] To make the optimization tractable, $q(\boldsymbol{z})$ is usually constrained. Depending on the constraints, one can obtain various algorithms to find the stationary solutions of the optimization problem [27]

$$q^* = \arg\min_q F[q]\,. \tag{11}$$

In the SHGF model, we constrain the approximating distribution to be factorized into normalized terms over hierarchical layers. We utilize a structured factorization that reflects the first-order Markov assumption $q(\boldsymbol{z})$ over the layers:

$$\prod_i q\left(x_t^{(i)}, x_{t-1}^{(i)}\right) q\left(s_t^{(i)}, s_{t-1}^{(i)}\right) q\left(\boldsymbol{\kappa}^{(i)}\right) q\left(\boldsymbol{\omega}^{(i)}\right) q\left(\boldsymbol{A}^{(i)}\right)$$

such that each factor

$$q\left(x_t^{(i)}\right) = \int q\left(x_t^{(i)}, x_{t-1}^{(i)}\right) \mathrm{d}x_{t-1}^{(i)} \approx p\left(x_t^{(i)}|\boldsymbol{y}\right) \tag{12}$$

approximates the desired smoothing marginal (6) by imposing a marginalization constraint on the joint. By means of variational calculus it can be shown that the stationary solutions to the optimization problem (11) under the specified constraints have the functional form

$$q^*\left(x_t^{(i)}, x_{t-1}^{(i)}\right) = \frac{1}{Z_{x_{t,t-1}}^{(i)}} \exp\left(\mathbb{E}_{\setminus\{x_t^{(i)}, x_{t-1}^{(i)}\}}\left[\log p(\boldsymbol{y}, \boldsymbol{z})\right]\right) \tag{13}$$

where $Z_{x_{t,t-1}}^{(i)}$ is a normalization constant. Due to the factorized model structure, the stationary marginals (12) can efficiently be obtained as multiplication of messages on a factor graph corresponding to the SHGF model.

[1] In this paper, all expectations are with respect to the q distribution, so $\mathbb{E}_{\boldsymbol{z}}$ is short for $\mathbb{E}_{q(\boldsymbol{z})}$. Expectations taken with respect to all other variables but $\boldsymbol{z}$ are denoted by $\mathbb{E}_{\setminus \boldsymbol{z}}$.

B. Factor Graphs, Message Passing and the GCSV Node

Forney-style factor graphs (FFG) are particularly useful for signal processing purposes [28]. Inference on a model can be interpreted as message passing on the corresponding FFG [28]. One can obtain a variational message passing algorithm to compute the stationary solutions (13). Due to the factorized model structure, the computation of (13) localizes over the time-segments of the FFG for the SHGF [10]. This means that the model induces a factorization on the approximate posterior q that supports computation of the local marginal approximating the smoothing solution (6) by multiplication of

$$q\left(x_t^{(i)}\right) \propto \overrightarrow{\nu}\left(x_t^{(i)}\right) \overleftarrow{\nu}\left(x_t^{(i)}\right) \uparrow\!\nu\left(x_t^{(i)}\right) \tag{14}$$

where the messages are obtained via

$$\overrightarrow{\nu}\left(x_t^{(i)}\right) \propto \int \overrightarrow{\nu}\left(x_{t-1}^{(i)}\right) \tilde{p}\left(x_t^{(i)}, x_{t-1}^{(i)}\right) \mathrm{d}x_{t-1}^{(i)} \tag{15a}$$

$$\overleftarrow{\nu}\left(x_t^{(i)}\right) \propto \int \overleftarrow{\nu}\left(x_{t+1}^{(i)}\right) \tilde{p}\left(x_t^{(i)}, x_{t+1}^{(i)}\right) \mathrm{d}x_{t+1}^{(i)} \tag{15b}$$

$$\tilde{p}\left(x_t^{(i)}, x_{t-1}^{(i)}\right) \triangleq \exp\left(\mathbb{E}_{\setminus\{x_t^{(i)}, x_{t-1}^{(i)}\}}\left[\log f_t^{(i)}\right]\right) \tag{15c}$$

$$\uparrow\!\nu\left(x_t^{(i)}\right) \propto \exp\left(\mathbb{E}_{\setminus\{x_t^{(i)}\}}\left[\log f_t^{(i-1)}\right]\right)\,. \tag{15d}$$

See Figure 1 for the messages around the GCSV node. Here (15a) and (15b) correspond to forward and backward messages that are sent by the GCSV node. The upward message to the upper layers is computed by (15d) and finally the marginal is computed at an equality node by (14) through multiplying forward, backward and upwards messages. By iteratively computing equation set (15), we obtain a structured variational message passing algorithm [29] [10]. In the FFG corresponding to the SHGF, messages around nodes other than GCSV can already be found in [30]. Around the GCSV node in Figure 1, the computation of messages ⑧, ⑬, ⑭, ⑮, ⑯ and ⑱ is the bottleneck to inference in the SHGF model. Due to space constraints we supply the algebraic operations to compute the entire list of messages and marginals for the SHGF model in a supplementary note[2] and present the results for the GCSV node in Table I under the assumptions of Table II. Messages ⑧ and ⑬ are Gaussian and message ⑱ is Categorical. The functional forms of the remaining messages do not correspond to known parametric exponential family distributions. We note that the messages associated with the continuously valued states have variances comprised of mixture of terms weighted by the discrete state probabilities. This mixture behaviour is the main difference with the HGF update equations.

Owing to the functional forms of messages ⑬, ⑭ and ⑮, the computation of marginals by multiplication, for example (14) is no longer a conjugate multiplication. If the multiplication is not approximated by a parametric exponential family distribution, then the complexity of the variational algorithm grows and quickly becomes infeasible. Fortunately, there are various ways to approximate non-conjugate multiplications.

[2] https://biaslab.github.io/pdf/isit2021/SHGF_derivations.pdf

TABLE I: Summary of message computations for the GCSV node. Computations require quantities that are defined in Table II. See https://biaslab.github.io/pdf/isit2021/SHGF_derivations.pdf for derivations.

Messages	**Functional form**
⑧	$\mathcal{N}\left(x_t^{(1)}\vert\overrightarrow{m}_{t-1}^{(1)},\overrightarrow{v}_{t-1}^{(1)}+\sum_{k=1}^{M_i}\pi_{t,k}^{(1)}\exp\left(\gamma_{t,k}^{(1)}+\beta_{t,k}^{(1)}\right)^{-1}\right)$
⑬	$\mathcal{N}\left(x_t^{(1)}\vert\overleftarrow{m}_{t+1}^{(1)},\overleftarrow{v}_{t+1}^{(1)}+\sum_{k=1}^{M_i}\pi_{t,k}^{(1)}\exp\left(\gamma_{t,k}^{(1)}+\beta_{t,k}^{(1)}\right)^{-1}\right)$
⑭	$\exp\left(-0.5\sum_j\pi_{t,j}^{(1)}\left(\left(\boldsymbol{\mu}_t^{(1)}\right)_j x_t^{(2)}+h\left(x_t^{(2)}\right)\right)\right)$
㉒	$\exp\left(-0.5\sum_k\pi_{t,k}^{(1)}\left(\kappa_k^{(1)}m_t^{(2)}+r\left(\boldsymbol{\kappa}^{(i)}\right)\right)\right)$
㉓	$\exp\left(-0.5\sum_k\pi_{t,k}^{(1)}\left(\omega_k^{(1)}+\zeta_t^{(1)}\exp\left(-\omega_k^{(1)}\right)\right)\right)$
⑱	$\prod_j\exp\left(-0.5\left(\eta_{t,j}^{(1)}+\zeta_t^{(1)}\exp\left(\gamma_{t,j}^{(1)}+\beta_{t,j}^{(1)}\right)\right)\right)^{\left[s_t^{(1)}=j\right]}$
Auxilary	**Definition by moment statistics**
$\eta_{t,j}^{(i)}$	$\left(\boldsymbol{\mu}_t^{(i)}\right)_j m_t^{(i+1)}+\left(\boldsymbol{\vartheta}_t^{(i)}\right)_j$
$\Psi_t^{(i)}$	$\left(\boldsymbol{\Sigma}_{t,t-1}^{(i)}\right)_{11}+\left(\boldsymbol{\Sigma}_{t,t-1}^{(i)}\right)_{22}-\left(\boldsymbol{\Sigma}_{t,t-1}^{(i)}\right)_{12}-\left(\boldsymbol{\Sigma}_{t,t-1}^{(i)}\right)_{21}$
$\Phi_{t,j}^{(i)}$	$\left(\boldsymbol{\mu}_t^{(i)}\right)_j^2 v_t^{(i+1)}+\left(\boldsymbol{\Omega}_t^{(i)}\right)_{jj}\left(m_t^{(i+1)}\right)^2+v_t^{(i+1)}\left(\boldsymbol{\Omega}_t^{(i)}\right)_{jj}^2$
$\zeta_t^{(i)}$	$\left(\left(\boldsymbol{m}_{t,t-1}^{(i)}\right)_1-\left(\boldsymbol{m}_{t,t-1}^{(i)}\right)_2\right)^2+\Psi_t^{(i)}$
$\gamma_{t,j}^{(i)}$	$-\left(\boldsymbol{\mu}_t^{(i)}\right)_j m_t^{(i+1)}+0.5\Phi_{t,j}^{(i)}$
$\beta_{t,j}^{(i)}$	$-\left(\boldsymbol{\vartheta}_t^{(i)}\right)_j+0.5\left(\boldsymbol{\Xi}_t^{(i)}\right)_{jj}$
$h\left(x_t^{(i)}\right)$	$\zeta_t^{(i)}\exp\left(-\left(\boldsymbol{\mu}_t^{(i-1)}\right)_j x_t^{(i)}+0.5\left(x_t^{(i)}\right)^2\left(\boldsymbol{\Omega}_t^{(i-1)}\right)_{jj}\right)$
$r\left(\boldsymbol{\kappa}^{(i)}\right)$	$\zeta_t^{(i)}\exp\left(-m_t^{(i+1)}\kappa_k^{(i)}+0.5v_t^{(i+1)}\left(\kappa_k^{(i)}\right)^2\right)$

For instance, Laplace approximation requires expanding the multiplication into a Taylor series and finding a stationary point where the gradient almost vanishes [26, Ch. 4.4]. Then the multiplication is approximated with a Gaussian distribution whose mean is a point where gradient vanishes and covariance is the inverse Hessian [26, Ch. 4.4]. Another approach is moment matching [25, Ch. 6] which gives rise to notable algorithms such expectation propagation [31] and assumed density filtering [32]. In [21], moment computations in expectation propagation is achieved by quadrature methods. Along the lines of moment matching, [10] implements a quadrature-based approximation to the non-conjugate multiplication and shows that quadrature-based moment matching outperforms Laplace's method. Here, we choose the quadrature-based moment approximation of [10] to handle message multiplications.

The quadrature-based moment matching approximation of [10] starts by determining the normalization constant that corresponds to the marginal computed by (14). The computation assumes that the messages $\overleftarrow{\nu}(x_t^{(i)})$ and $\overrightarrow{\nu}(x_t^{(i)})$ are Gaussian. This allows us to write the normalization constant in the form of a Gaussian integral with limits at infinity, i.e.,

$$Z_{x_t}^{(i)}=\int_{\infty}^{\infty}\uparrow\nu\left(x_t^{(i)}\right)\mathcal{N}\left(x_t^{(i)}\vert\tilde{m}_t^{(i)},\tilde{v}_t^{(i)}\right)\mathrm{d}x_t^{(i)} \tag{16}$$

where $\tilde{m}_t^{(i)}$ and $\tilde{v}_t^{(i)}$ are the corresponding statistics for the Gaussian resulting from the multiplication of $\overleftarrow{\nu}(x_t^{(i)})$ and $\overrightarrow{\nu}(x_t^{(i)})$. Using Hermite polynomials, integration in (16) can be obtained by Gaussian quadrature such that

$$Z_{x_t}^{(i)}\approx\frac{1}{\sqrt{\pi}}\sum_k w_k^{(i)}\uparrow\nu\left(\psi_k^{(i)}\sqrt{2\tilde{v}_t^{(i)}}+\tilde{m}_t^{(i)}\right) \tag{17}$$

where $\psi_k^{(i)}$ are points that are the roots of Hermite polynomials and $w_k^{(i)}$ are the corresponding weights [25, Ch. 6]. Once the normalization constant (17) has been determined, the moments of the distribution corresponding to the non-conjugate multiplication (14) can be evaluated by

$$\mathbb{E}\left[\left(x_t^{(i)}\right)^n\right]=\frac{\sum_k\uparrow\nu\left(\psi_k^{(i)}\sqrt{2\tilde{v}_t^{(i)}}+\tilde{m}_t^{(i)}\right)\left(\psi_k^{(i)}\sqrt{2\tilde{v}_t^{(i)}}+\tilde{m}_t^{(i)}\right)^n}{\sqrt{\pi}Z_{x_t}^{(i)}}.$$

Using the first two moments we can now approximate the non-conjugate multiplication by a Gaussian distribution. Due to one dimensional nature of the problem Gauss-Hermite integration does not suffer from curse of dimensionality and is computationally feasible. In our experiments we fix the order of Gauss-Hermite polynomials to 11 and plan to address the effect of polynomial order in further research.

TABLE II: Messages and marginals required in Table I.

Messages	**Functional form**
$\overrightarrow{\nu}\left(x_{t-1}^{(i)}\right)$	$\mathcal{N}\left(x_{t-1}^{(i)}\vert\overrightarrow{m}_{t-1}^{(i)},\overrightarrow{v}_{t-1}^{(i)}\right)$
$\overleftarrow{\nu}\left(x_t^{(i)}\right)$	$\mathcal{N}\left(x_t^{(i)}\vert\overleftarrow{m}_t^{(i)},\overleftarrow{v}_t^{(i)}\right)$.
Marginals	**Functional form**
$q\left(x_t^{(i-1)},x_{t-1}^{(i-1)}\right)$	$\mathcal{N}\left(\boldsymbol{x}_{t,t-1}^{(i+1)}\vert\boldsymbol{m}_{t,t-1}^{(i+1)},\boldsymbol{\Sigma}_{t,t-1}^{(i+1)}\right)$
$q\left(x_t^{(i+1)}\right)$	$\mathcal{N}\left(x_t^{(i+1)}\vert m_t^{(i+1)},v_t^{(i+1)}\right)$
$q_t\left(\boldsymbol{\kappa}^{(i)}\right)$	$\mathcal{N}\left(\boldsymbol{\kappa}^{(i)}\vert\boldsymbol{\mu}_t^{(i)},\boldsymbol{\Omega}_t^{(i)}\right)$
$q_t\left(\boldsymbol{\omega}^{(i)}\right)$	$\mathcal{N}\left(\boldsymbol{\omega}^{(i)}\vert\boldsymbol{\vartheta}_t^{(i)},\boldsymbol{\Xi}_t^{(i)}\right)$
$q\left(s_t^{(i)}\right)$	$\prod_{k=1}^{M_i}\left(\pi_{t,k}^{(i)}\right)^{\left[s_t^{(i)}=k\right]}$

V. Experiments

All experiments have been implemented with the Julia package `ForneyLab` [33]. The source code for the experiments can be found at https://github.com/biaslab/SGCV.

A. Verification

To verify the proposed inference algorithm, we built a 2-layer (2-L) SHGF model (see Fig. 1) where $\boldsymbol{\omega}^{(1)},\boldsymbol{\kappa}^{(1)}\in\mathbb{R}^3$. We generated $N=100$ data sets with $T=500$ observation points in each set. We used weakly informative priors for $x_0^{(1)}$, $x_0^{(2)}$, $s_0^{(1)}$ and $\boldsymbol{A}^{(1)}$, but informative for $\boldsymbol{\omega}^{(1)}$, i.e. $\boldsymbol{\omega}^{(1)}\sim\mathcal{N}(\boldsymbol{\omega}^*,\mathbf{I})$ where $\boldsymbol{\omega}^*\sim\mathcal{N}(\boldsymbol{\omega}^{\text{true}},\mathbf{I})$ ($\boldsymbol{\omega}^{\text{true}}$ denotes ground-truth parameters). Note that in these experiments we did not learn $\boldsymbol{\kappa}^{(1)}$. As the update equations for $\boldsymbol{\kappa}^{(1)}$ and x_t are symmetrical, one of these random variables should be observed. Otherwise, learning of $\boldsymbol{\kappa}^{(1)}$ together with $\boldsymbol{\omega}^{(1)}$ and x_t would lead to identifiability issues. An approach to overcome identifiability is to constrain $\boldsymbol{\kappa}^{(1)}$ further. For example, constraining the support set of $\boldsymbol{\kappa}^{(1)}$ from $\mathbb{R}^{M_1}$ to $[0,1]^{M_1}$ and bounding the variance of state transitions that $\boldsymbol{\kappa}^{(1)}$ undergoes, is an approach that can help learning of $\boldsymbol{\kappa}^{(1)}$.

We ran the proposed message passing algorithm on the full SHGF graph with $T=500$ time segments and 500×10 nodes in total for the entire data set. The update schedule for one-time segment of SHGF is shown in Fig. 1. Fig. 2 reports the results of the verification experiments. The verification results indicate that the hybrid VMP algorithm consistently decreases free-energy averaged over the entire data set and converges to stationary solutions of the minimization problem (11).

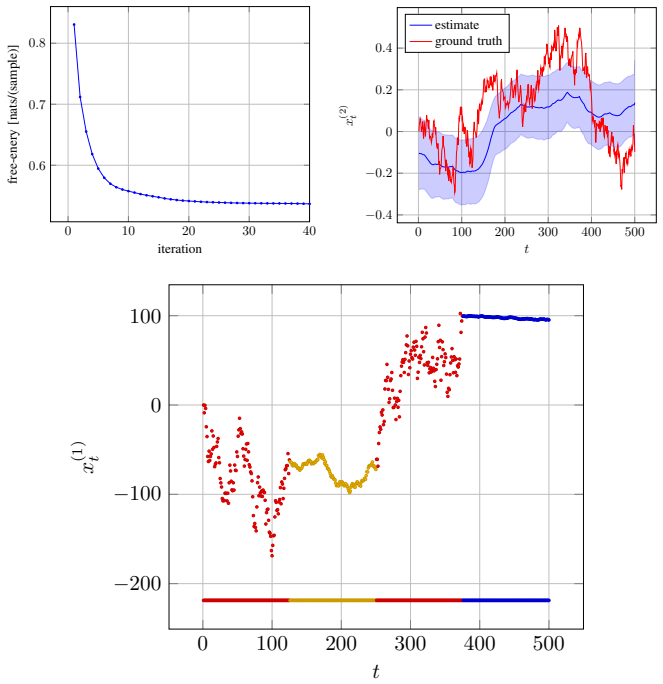

Fig. 2: Verification results. (top-left) Evolution of free energy per variational iteration averaged over data sets ($N = 100$) and number of observations ($T = 500$). The curve indicates that the proposed algorithm consistently minimizes free energy and converges to stationary solutions. (top-right) An example of the inference of the upper layer random walk. The red signal indicates the second layer continuous state $x_t^{(2)}$ that corresponds to the observations at the bottom figure. The blue curve is the estimate of the state obtained by the VMP algorithm. The estimate recovers the trend of the second layer state. (bottom) An example of observations from one of the data sets that are used to verify the algorithm. Each color represents a particular regime (switch). Observations are color-coded according to the regimes they are generated from. The mode of categorical distributions corresponding to the switch variables for the entire time points are marked below the signal. The plot indicates that recovery of switching regimes matches the ground truth.

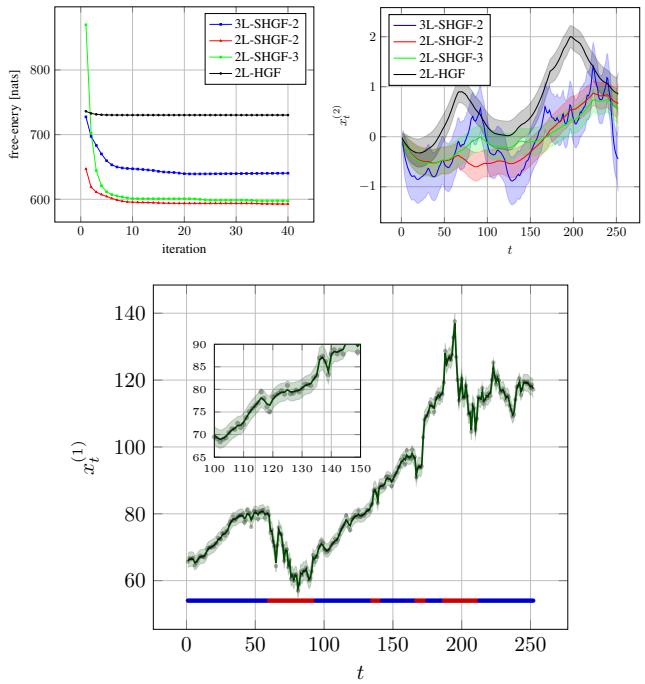

Fig. 3: Validation results. (top-left) Free energy plots corresponding to a 3-layer SHGF with 2 categories, a 2-layer SHGF with 3 categories, a 2-layer SHGF with 2 categories and a 2-layer HGF, where the observations are AAPL stock prices. All SHGF models outperform the 2-layer HGF, indicating that the prices indeed exhibit switching behavior. Since the 2-layer SHGFs performs better than a 3-layer SHGF, we conclude that the extra model complexity of the 3rd layer outweighs the increased accuracy due to the 3rd layer. The 2-layer SHGF with 3 categories performs almost as well as the one with 2 categories. The inference results for 3 categories indicate that the model actually assigns a vanishing probability to the third category, meaning that it actually settles for 2 categories. (top-right) Second layer state trajectories for the 3-layer SHGF, 2-layer SHGFs and 2-layer HGF models obtained from the AAPL stocks. The 3-layer SHGF is quite active due to the presence of an extra layer. The 2-layer SHGFs and HGF are smoother. These three trajectories share a similar trend where the volatility makes two peaks around $t = 60$ and $t = 200$. On average, the HGF model attributes a higher volatility to the stock prices than the SHGF model. (bottom) Black dots correspond to the stock prices. The green curve represents the belief trajectory for $x_t^{(1)}$ obtained by the 2-layer SHGF model. In order to avoid clutter in the plot, we only present the model with the lowest free energy (2L-SHGF-2). We display a zoomed version on the smooth behavior of the belief trajectory. The obtained switches are color coded and displayed beneath the prices. Based on the model, there are 2 underlying regimes governing the prices.

B. Validation

In order to validate our model, we applied the SHGF model to a real-world data set. The data set corresponds to AAPL stock prices (downloaded from https://finance.yahoo.com/quote/AAPL/). We wanted to test if the stock price evolution exhibits regime-switching dynamics over a consecutive period of $T = 252$ days. We used the minimized variational free energy as a model performance score and compared 4 different models: a 2-layer HGF [10], 2-layer SHGFs with 2 and 3 categories, and a 3-layer SHGF. To keep the comparison fair we used identical priors for the states and parameters where possible. Fig. 3 highlights the results of validation experiments. The 2-layer SHGF with 2 categories results in lower free energy than the 3-layer SHGF (too complex) and HGF (too simple). The 2-layer SHGF with 3 categories assigns vanishing probability to the 3rd category, so the 2-layer SHGF with 2 regimes is optimal. This indicates that the underlying prices submit to two-category regime switching dynamics.

VI. Conclusions

We introduced a Switching Hierarchical Gaussian Filter (HGF) to model regime-switching non-stationary time series. The proposed model extends the classical HGF by assuming that the parameters in each layer are selected by a discrete state, which in turn evolves according to a hidden Markov model. We presented a closed-form variational message passing framework to track all states and the transition matrix (a matrix of parameters) for the discrete states. The presented message passing framework relies on hybridization of conjugate and non-conjugate variational message update rules. We verified that the proposed inference algorithm finds the stationary solutions of the minimization problem and consistently minimizes free energy. After verification, we showed that the SHGF provides improved results over the HGF on modelling of stock prices. Crucially, the closed-form update rules for the problematic GCSV factor allow it to be used as a plug-in node in any factor graph, thus enabling message passing-based inference for alternative hierarchical dynamic models.

Acknowledgments

This work was partly financed by research program ZERO with project number P15-06, which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO).

REFERENCES

[1] I. B. Yildiz, K. von Kriegstein, and S. J. Kiebel, "From Birdsong to Human Speech Recognition: Bayesian Inference on a Hierarchy of Nonlinear Dynamical Systems," *PLoS Computational Biology*, vol. 9, no. 9, p. e1003219, Sep. 2013.

[2] R. Ranganath, D. Tran, and D. M. Blei, "Hierarchical Variational Models," *Journal of Machine Learning Research*, vol. 48, p. 10, 2016.

[3] T. Dean, "Scalable Inference in Hierarchical Generative Models." in *ISAIM*. Citeseer, 2006.

[4] K. Friston, "Hierarchical models in the brain," *PLOS Computational Biology*, vol. 4, no. 11, pp. 1–24, 11 2008.

[5] E. Moench, S. Ng, and S. M. Potter, "Dynamic hierarchical factor models," Federal Reserve Bank of New York, Staff Reports 412, 2009.

[6] C. D. Mathys, "Hierarchical Gaussian filtering," Ph.D. dissertation, Diss., Eidgenoessische Technische Hochschule ETH Zuerich, Nr. 20909, 2012.

[7] C. D. Mathys, E. I. Lomakina, J. Daunizeau, S. Iglesias, K. H. Brodersen, K. J. Friston, and K. E. Stephan, "Uncertainty in perception and the Hierarchical Gaussian Filter," *Frontiers in Human Neuroscience*, vol. 8, Nov. 2014.

[8] C. D. Mathys, "HGF Toolbox release 3.3.0 (package of TAPAS toolbox), https://github.com/translationalneuromodeling/tapas/," accessed 2021-01-28.

[9] İ. Şenöz and B. de Vries, "Online Variational Message Passing in the Hierarchical Gaussian Filter," in *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2018, pp. 1–6.

[10] İ. Şenöz and B. de Vries, "Online message passing-based inference in the hierarchical gaussian filter," in *2020 IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 2676–2681.

[11] İ. Şenöz, A. Podusenko, W. M. Kouw, and B. de Vries, "Bayesian joint state and parameter tracking in autoregressive models," in *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, ser. Proceedings of Machine Learning Research, A. M. Bayen, A. Jadbabaie, G. Pappas, P. A. Parrilo, B. Recht, C. Tomlin, and M. Zeilinger, Eds., vol. 120. The Cloud: PMLR, 10–11 Jun 2020, pp. 95–104.

[12] C. M. Carvalho and H. F. Lopes, "Simulation-based sequential analysis of Markov switching stochastic volatility models," *Computational Statistics & Data Analysis*, vol. 51, no. 9, pp. 4526–4542, May 2007.

[13] X. Liu, D. Margaritis, and P. Wang, "Stock market volatility and equity returns: Evidence from a two-state markov-switching model with regressors," *Journal of Empirical Finance*, vol. 19, no. 4, pp. 483 – 496, 2012.

[14] Z. Ghahramani and G. E. Hinton, "Variational learning for switching state-space models," *Neural computation*, vol. 12, no. 4, pp. 831–864, 2000.

[15] E. Fox, E. Sudderth, M. Jordan, and A. Willsky, "Nonparametric bayesian learning of switching linear dynamical systems," in *Advances in Neural Information Processing Systems*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., vol. 21. Curran Associates, Inc., 2009, pp. 457–464.

[16] J. Daunizeau, K. Friston, and S. Kiebel, "Variational bayesian identification and prediction of stochastic nonlinear dynamic causal models," *Physica D: Nonlinear Phenomena*, vol. 238, no. 21, pp. 2089 – 2118, 2009.

[17] V. P. Jilkov and X. R. Li, "Online bayesian estimation of transition probabilities for markovian jump systems," *IEEE Transactions on Signal Processing*, vol. 52, no. 6, pp. 1620–1630, 2004.

[18] E. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "Bayesian nonparametric inference of switching dynamic linear models," *IEEE Transactions on Signal Processing*, vol. 59, no. 4, pp. 1569–1585, 2011.

[19] D. Barber, *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.

[20] A. Doucet, N. De Freitas, K. Murphy, and S. Russell, "Rao-blackwellised particle filtering for dynamic bayesian networks," *arXiv preprint arXiv:1301.3853*, 2013.

[21] O. Zoeter and T. Heskes, "Gaussian Quadrature Based Expectation Propagation," *Tenth International Workshop on Artifical Intelligence and Statistics*, p. 9, 2005.

[22] D. Zhang, W. Wang, G. Fettweis, and X. Gao, "Unifying Message Passing Algorithms Under the Framework of Constrained Bethe Free Energy Minimization," *arXiv:1703.10932 [cs, math]*, Mar. 2017, arXiv: 1703.10932.

[23] H.-A. Loeliger, "An introduction to factor graphs," *Signal Processing Magazine, IEEE*, vol. 21, no. 1, pp. 28–41, 2004.

[24] G. Forney, "Codes on graphs: normal realizations," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 520–548, Feb. 2001.

[25] S. Särkkä, *Bayesian Filtering and Smoothing*. London ; New York: Cambridge University Press, Oct. 2013.

[26] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., 2006.

[27] J. S. Yedidia, W. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2282–2312, Jul. 2005.

[28] S. Korl, "A factor graph approach to signal modelling, system identification and filtering," Ph.D. dissertation, Swiss Federal Institute of Technology, Zurich, 2005.

[29] J. Dauwels, "On Variational Message Passing on Factor Graphs," in *IEEE International Symposium on Information Theory*, Jun. 2007, pp. 2546–2550.

[30] T. van de Laar, "Automated Design of Bayesian Signal Processing Algorithms," Ph.D. dissertation, Eindhoven University of Technology, Eindhoven, The Netherlands, 2019.

[31] T. P. Minka, "Expectation Propagation for Approximate Bayesian Inference," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 362–369.

[32] S. Ghosh, F. M. Delle Fave, and J. Yedidia, "Assumed density filtering methods for learning bayesian neural networks," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[33] M. Cox, T. van de Laar, and B. de Vries, "A factor graph approach to automated design of Bayesian signal processing algorithms," *International Journal of Approximate Reasoning*, vol. 104, pp. 185–204, Jan. 2019.