2018 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 17–20, 2018, AALBORG, DENMARK

ONLINE VARIATIONAL MESSAGE PASSING IN THE HIERARCHICAL GAUSSIAN FILTER

İsmail Şenöz[1] *and Bert de Vries*[1,2]

[1]Eindhoven University of Technology and [2]GN Hearing, Eindhoven, the Netherlands
i.senoz@tue.nl, bdevries@ieee.org

ABSTRACT

We address the problem of online state and parameter estimation in hierarchical Bayesian nonlinear dynamic systems. We focus on the Hierarchical Gaussian Filter (HGF), which is a popular model in the computational neuroscience literature. For this filter, explicit equations for online state estimation (and offline parameter estimation) have been derived before. We extend this work by casting the HGF as a probabilistic factor graph and present variational message passing update rules that facilitate both online state and parameter estimation as well as online tracking of the free energy (or ELBO), which can be used as a proxy for Bayesian evidence. Due to the locality and modularity of the factor graph framework, our approach supports application of HGF's and variations as plug-in modules to a wide variety of dynamic modelling applications.

***Index Terms*—** Dynamical systems, Hierarchical Gaussian Filter, Variational Message Passing, free energy, Online state and parameter estimation.

1. INTRODUCTION

Online updating of non-linear dynamic models for possibly non-stationary time series remains a much-studied subject in various disciplines. In this paper we focus on these issues for the Hierarchical Gaussian Filter (HGF), which is a particularly successful model in the computational neuroscience community, e.g., [1][2]. In this community, the HGF is positioned as a generative probabilistic non-linear hierarchical model for sensory observations. In this view, perceptual processes are modelled as a Bayesian inference (state estimation) task and learning corresponds to Bayesian inference of the model parameters.

In inspiring work by [1], analytic equations for online state estimation in the HGF are derived. [1] recommends offline ("batch") variational Bayesian learning for the HGF model parameters. Unfortunately, offline parameter estimation is not well-suited to track non-stationarities that are part of real-world sensoria. Moreover, while an open-source toolbox for HGF-based modeling is available [3], the toolbox does not automatically update the inference equations if the HGF model specification were slightly modified. These are limiting factors to a wide application of HGF-inspired modeling for non-stationary processes.

In the current paper, we extend the applicability of HGF-based modeling by casting inference in HGF's as a message passing task in a factor graph.

Specifically, our contributions are as follows:

- In Fig.1a, we present a Forney-style Factor Graph (FFG) representation of the HGF. Crucially, we isolate the nonlinear components as a *composite* factor from the linear Gaussian parts of the model. This isolation facilitates easy re-application of the non-linear factor to alternative models as the FFG framework allows re-use of factors as plug-in modules.

- The FFG formalism takes advantage of the factorization of the model by implementing inference as a message passing algorithm. In Sec. 2.5, we specify analytic update equations for both online state and parameter tracking in the HGF via a Laplace approximation to Variational Message Passing (VMP).

- VMP is based on minimization of variational free energy (FE), which is an upperbound to negative log-evidence and as such comprises a useful performance criterion for the HGF model. In Sec. 2.4 we show how FE is additively distributed in factor graphs and derive the local FE contributions of the critical factor in the HGF.

As a result of casting the HGF in an FFG framework, it is possible to *automate* the computation of online state and parameter estimation as well as updating the involved FE changes that result from message passing. Moreover, due to the modularity of the factor graph representation, the constituent factors in an HGF model can be re-combined or applied in alternative models such that inference and FE computations remain automated. As a result, the proposed methods provide a very versatile toolset for HGF-based modeling.

978-1-5386-5477-4/18/$31.00

2. METHODS

2.1. Model Specification

Consider a sequence of observations $\mathbf{y} \triangleq \mathbf{y}_{1:T} = (y_1, \ldots, y_T)$. The hierarchical Gaussian filter (HGF) is an N-layered generative model for this sequence of the form

$$p(\mathbf{y}, \mathbf{x}) = \prod_{t=1}^{T} \underbrace{p(y_t | x_t^{(1)})}_{\text{observation}} \prod_{i=1}^{N} \underbrace{p(x_t^{(i)} | x_{t-1}^{(i)}, x_t^{(i+1)})}_{\text{state transition}} \tag{1a}$$

$$p(x_t^{(i)} | x_{t-1}^{(i)}, x_t^{(i+1)}) = \delta\left(x_t^{(i)} - x_{t-1}^{(i)} - f_i(x_t^{(i+1)})\right) \tag{1b}$$

$$f_i(x_t^{(i+1)}) \sim \mathcal{N}\left(0, \exp(\kappa^{(i)} x_t^{(i+1)} + \omega^{(i)})\right), \tag{1c}$$

where $x_t^{(i)}$ denotes the (latent) state at time t in layer i and $\theta^{(i)} = \{\kappa^{(i)}, \omega^{(i)}\}$ comprises model parameters at layer i [4]. Eq. 1a factorizes the model into layers with Markovian dynamics at each layer. Eq. 1b specifies that the state transition model is a Gaussian random walk with *time-varying* variance that is determined by the state of the superior layer. Eq. 1c describes how the variance in the random walk step depends on the state of the superior layer. The exponent in Eq. 1c enforces a non-negative variance that contains a phasic (time-varying) component $\exp(\kappa x_t)$ and a tonic (time-invariant) component $\exp(\omega)$.

In principle, the HGF state transition model can be combined with any observation model $p(y_t | x_t^{(1)})$. For instance, for continuously valued observations, a Gaussian variation of the first-layer state is an option:

$$p(y_t | x_t^{(1)}) = \mathcal{N}\left(y_t \middle| x_t^{(1)}, \exp(\omega^{(0)})\right). \tag{2}$$

In order to properly run this model (i.e., generate $\mathbf{y}$), we will assume the initialization $x_0^{(i)} = 0$ for all layers and set $\kappa^{(N)} = 0$, since $x_t^{(N+1)}$ is not specified.

A factor graph-based illustration of the model with three layers is displayed in Fig. 1 (notational details of factor graphs will be discussed in Sec. 2.3).

2.2. Signal Processing as Inference

Once the generative model is specified, there are generally three tasks we are interested in when the model is applied in a signal processing context. These tasks are (recursive, online) state estimation, parameter estimation and evaluation of model performance. Let $\mathbf{x}_t = (x_t^{(1)}, \ldots, x_t^{(N)})$ collect the states from all layers at time t. Given past and current observations $\mathbf{y}_{1:t}$, application of Bayes rule leads to the recursive state estimation equations, which amounts to obtaining the posterior, $p(\mathbf{x}_t | \mathbf{y}_{1:t})$ through computing[1]

[1] Here, for notational simplicity we removed the model parameters from the equations.

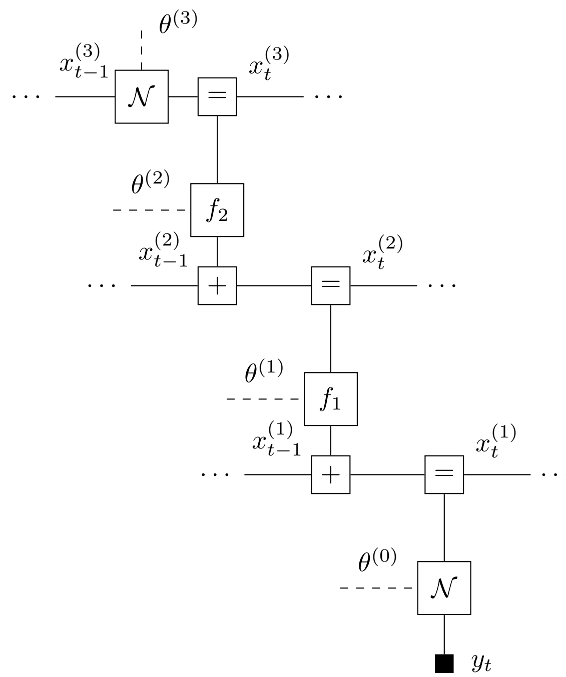


Fig. 1. One time segment of the FFG corresponding to a 3-layer HGF model as defined by Eqs. 1 and 2. We use small black nodes for observed variables, midsize nodes for deterministic factors and large nodes for stochastic factors. Solid and dashed edges are associated with states and parameters respectively. The dotted edges indicate that the graph can be extended in the same way for the other time steps.

$$\underbrace{p(\mathbf{x}_t | \mathbf{y}_{1:t})}_{\text{posterior}} = \underbrace{\frac{1}{p(y_t | \mathbf{y}_{1:t-1})}}_{\text{evidence}} \underbrace{p(y_t | x_t^{(1)})}_{\text{likelihood}} \cdot \int \prod_{i=1}^{N} \underbrace{p(x_t^{(i)} | x_{t-1}^{(i)}, x_t^{(i+1)})}_{\text{state transition}} \underbrace{p(x_{t-1}^{(i)} | \mathbf{y}_{1:t-1})}_{\text{prior}} \, \mathrm{d}x_{t-1}^{(i)}. \tag{3}$$

Online parameter updating consists of tracking $p(\theta_t^{(i)} | \mathbf{y}_{1:t})$, which, from a Bayesian viewpoint, can be treated as online estimation of an "extended" state.

In Eq. 3, the unknown parts are the posterior and evidence terms, since the prior is inherited from the posterior of the previous time step and the state transition and likelihood terms are given by the model specification Eqs. 1 and 2. While Eq. 3 comprises an exact recipe for online state estimation, due to the integration over states and non-conjugate prior-posterior pairing, executing Eq. 3 is generally intractable (and indeed this is the case for the HGF). Hence, we need a numerical approximation for evidence and posterior updating. In this paper, we present variational message passing (VMP) on factor graph as an approximate online inference method.

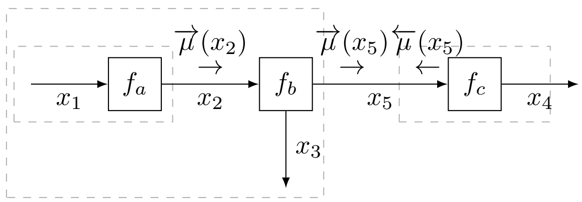


Fig. 2. An FFG corresponding to the model given by Eq. 4, including messages as per Eq. 6.

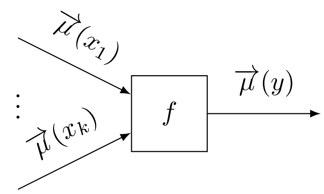


Fig. 3. Sum-product message passing for a generic factor $f(x_1, \ldots x_k, y)$. The outgoing message is given by Eq. 7.

2.3. Forney-style Factor Graphs

We give a brief introduction to Forney-style Factor Graphs (FFG). An FFG is a graphical representation of a factorization of a global function. Within the context of probabilistic modeling, an FFG represents a factorized probability distribution. Nodes in an FFG represent factors and edges are associated with variables. An edge is connected to a node if and only if the (edge) variable is part of the argument list of the (node) function [5].

Since an edge can not be connected to no more than two nodes, there seems to be a problem if a variable name appears in more than two factors. However, this restriction is easily resolved via introduction of auxiliary variables whose marginals are forced to be equal through adding equality factors to the model.

Consider an example model with factorization given by

$$p(x_1, \ldots, x_5) = f_a(x_1, x_2) f_b(x_2, x_3, x_5) f_c(x_4, x_5). \quad (4)$$

This factorization is visually displayed as an FFG in Fig. 2. Assume that we are interested in obtaining the marginal distribution of x_5, which corresponds to evaluating the integral

$$p(x_5) = \int \ldots \int p(x_1, \ldots, x_5)\, \mathrm{d}x_1 \ldots \mathrm{d}x_4. \quad (5)$$

If we plug the factorized form of Eq. 4 into Eq. 5 we can arrange sums-of-products into products-of-sums by the distribute law, yielding

$$p(x_5) = \underbrace{\iint f_b(x_2, x_3, x_5) \mathrm{d}x_3 \underbrace{\int f_a(x_1, x_2) \mathrm{d}x_1}_{\overrightarrow{\mu}(x_2)} \mathrm{d}x_2}_{\overrightarrow{\mu}(x_5)} \cdot \underbrace{\int f_c(x_4, x_5) \mathrm{d}x_4}_{\overleftarrow{\mu}(x_5)}. \quad (6)$$

Thus, a high-dimensional integral over four variables in Eq. 5 reduces to a set of smaller integrals. These sub-integrals can be locally computed by the nodes, effectively doing inference by passing on messages through the graph, see Fig. 2. In this example, application of the distributive law leads to the *sum-product* rule. For a generic node $f(x_1, \ldots, x_k, y)$ (see Fig. 3) with incoming messages $\mu_{X_1}(x_1), \ldots, \mu_{X_k}(x_k)$, the outgoing sum-product message is given by

$$\overrightarrow{\mu}_Y(y) = \int \ldots \int f(y, x_1, \ldots, x_N) \prod_{i=1}^{N} \overrightarrow{\mu}_{X_i}(x_i)\, \mathrm{d}x_i. \quad (7)$$

For a detailed explanation of sum-product message passing in FFGs, we refer to [5, 6].

2.4. Free Energy Minimization and Variational Message Passing

If the integrals in Eq. 7 are not analytically tractable then sum-product messages can not be computed. There are however alternative message passing formulations that lead to approximate Bayesian inference in FFGs. Next, we discuss Variational Message Passing (VMP), which is very popular for inference in models that are composed of factors of the exponential family of probability distributions [7]. VMP recognizes that exact Bayesian inference is not tractable and instead aims to approximate the posterior for the hidden states $\mathbf{x}_t$ by a *recognition distribution* $q(\mathbf{x}_t)$. Consider a so-called variational free energy functional (also known as negative-ELBO), given by

$$F_t[q] \triangleq \int q(\mathbf{x}_t) \log \frac{q(\mathbf{x}_t)}{p(\mathbf{x}_t, y_t | \mathbf{y}_{1:t-1})} \mathrm{d}\mathbf{x}_t \quad (8a)$$

$$= \underbrace{- \log p(y_t | \mathbf{y}_{1:t-1})}_{-\text{log-evidence}} + \underbrace{\int q(\mathbf{x}_t) \log \frac{q(\mathbf{x}_t)}{p(\mathbf{x}_t | \mathbf{y}_{1:t})} \mathrm{d}\mathbf{x}_t}_{\text{KL-divergence}} \quad (8b)$$

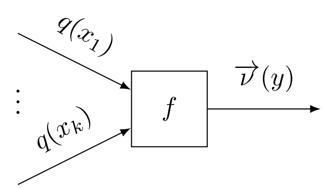
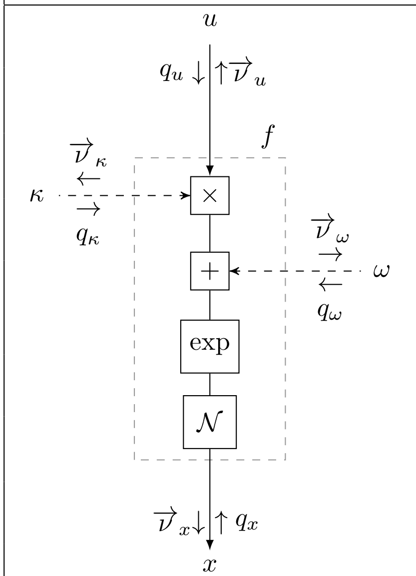


Fig. 4. Variational message passing for a generic node representing function $f(x_1, \ldots x_k, y)$.

Table 1. Message update rules for the Gaussian-with-Controlled-Variance (GCV) node $f(x,u,\kappa,\omega)$. Assume incoming messages $q_u \sim \mathcal{N}(m_u, v_u)$, $q_x \sim \mathcal{N}(m_x, v_x)$, $q_\omega \sim \mathcal{N}(m_\omega, v_\omega)$ and $q_\kappa \sim \mathcal{N}(m_\kappa, v_\kappa)$. Then, outgoing Gaussian variational messages per Eq. 9 as indicated in the table. The middle row specifies the local free energy contribution per Eq. 10 by the GCV node. The final row specifies auxiliary variables $\gamma_\bullet$ that are used in the message update rules.

Node	**Message**	**Update equation**
	$\overrightarrow{\nu}(x)$	$\mathcal{N}(0, \gamma_2)$
	$\overrightarrow{\nu}(u)$	$\mathcal{N}\left(\frac{\log(\gamma_3)}{m_\kappa}, \frac{2}{m_\kappa^2}\right)$
	$\overrightarrow{\nu}(\kappa)$	$\mathcal{N}\left(\frac{\log(\gamma_3)}{m_u}, \frac{2}{m_u^2}\right)$
	$\overrightarrow{\nu}(\omega)$	$\mathcal{N}(\log \gamma_4, 1)$
	$F[q]$	$\frac{1}{2}(\log(2\pi) + (m_\kappa m_u + m_\omega)) + \frac{1}{2}\left(m_x^2 + v_x\right)\gamma_2^{-1} - \log\left((2\pi e)^2 \sqrt{v_x v_u v_\kappa v_\omega}\right)$
	γ_1	$m_u^2 v_\kappa + m_\kappa^2 v_u + v_u v_\kappa$
	γ_2	$\exp\left(m_\kappa m_u + m_\omega - \left(\frac{\gamma_1 + v_\omega}{2}\right)\right)$
	γ_3	$\left(m_x^2 + v_x\right)\exp\left(-m_\omega + \frac{v_\omega}{2}\right)$
	γ_4	$\left(m_x^2 + v_x\right)\exp\left(-m_\kappa m_u + \frac{\gamma_1}{2}\right)$

Since the KL-divergence term in Eq. 8b is always non-negative, the free energy is an upper-bound to the negative log-evidence, i.e $F_t[q] \geq -\log p(y_t|\mathbf{y}_{1:t-1})$, with equality holding if and only if $q(\mathbf{x}_t) = p(\mathbf{x}_t|\mathbf{y}_{1:t})$. This means that minimizing $F_t[q]$ leads to approximations for both the evidence and posterior state distribution, i.o.w., approximation of the online state estimation problem as stated by Eq. 3. In the Variational Bayes method, it is common to use the free energy to score the performance of a model.

In an FFG framework, the free energy functional can be minimized through Variational Message Passing. Consider a generic node $f(x_1, \ldots, x_K, y)$ with incoming (marginal) messages $q(x_k)$, see Fig.4. It can be shown by variational calculus that minimization of FE results from sending an outgoing message

$$\overrightarrow{\nu}(y) \propto \exp\left(\int q(\mathbf{x}) \log f(y, x_1, \ldots, x_k)\, \mathrm{d}\mathbf{x}\right), \qquad (9)$$

followed by updating the marginal for y by $q(y) = \overrightarrow{\nu}(y)\overleftarrow{\nu}(y)$ [7].

Note that the free energy can be computed as a sum of local free energies in an FFG. To show this, assume a factorized probabilistic model $p(\mathbf{x}) = \prod_{a=1}^{M} p_a(x_a)$, where a is some index set, x_a is a subset of variables that are arguments for p_a and each p_a is a stochastic factor. Using a *mean field* assumption for the recognition distribution $q(\mathbf{x}) = \prod_{i=1}^{N} q(x_i)$, it follows that

$$F[q] = \sum_{a=1}^{M} \underbrace{\int -\log p_a(x_a) \prod_{j \in N(a)} q(x_j)\mathrm{d}x_a}_{U[p_a]} + \sum_{i=1}^{N} \underbrace{\int q(x_i)\log q(x_i)\mathrm{d}x_i}_{-H[q_i]}, \qquad (10)$$

where $N(a)$ denotes the set of variables that are arguments of p_a. Hence, each factor (node) p_a contributes an average energy term $U[p_a]$ and each (edge) q_i contributes a negative-entropy term $H[q_i]$ to the free energy functional.

In summary, the FFG framework provides a provides a flexible framework to develop online state and parameter estimation as well as performance scoring for a wide range of dynamic models.

2.5. Variational Message Passing for the Hierarchical Gaussian Filter

Working out sum-product or variational update rules for the nodes in the HGF factor graph (Fig. 1) leads only to a problem for the non-linear nodes f_i, whose factors are specified by

$$f(x, u, \kappa, \omega) = \mathcal{N}(x \,|\, 0, \exp(\kappa u + \omega)). \qquad (11)$$

In this factor, u controls the variance of a Gaussian, and hence we call this primitive structure the Gaussian--with-Controlled-Variance (GCV) factor, see Table 1 for the internal FFG structure of this factor. In principle, working out approximate message update rules for all interfaces (i.e., the edges x, u, κ and ω) would allow the GCV node to exchange messages in freely definable models, including the HGF. In Table 1, we provide the full set of these variational messages (as well as the local free energy contributions) that have been derived through Laplace approximation of the integrals in Eq. 9.[2]With these rules, together with the sum-product rules for the addition, equality and standard Gaussian nodes (see [8]), it is possible to conduct online state and parameter estimation and free energy tracking in HGF models.

3. EXAMPLE: MODELING CURRENCY EXCHANGE RATES

3.1. Experimental Setup

In order to demonstrate the proposed message passing rules, we modeled the daily currency exchange rate between USD and CHF during the period 2010-2011.[3] We employed a two-level HGF with Gaussian observation model and discuss here the state estimation process. The model parameters were obtained via TAPAS [3] as: $\kappa^{(2)} = 0$, $\kappa^{(1)} = 1$, $\kappa^{(0)} = 0$, $\omega^{(0)} = -16.03$, $\omega^{(1)} = -11.84$ and $\omega^{(2)} = -5.90$.

Fig. 5 displays one time-step of the message passing schedule for the two-layer HGF model. Messages 1 and 7 carry state predictions from time-step $t-1$ and the likelihood node is terminated by observation y_t. The message passing schedule 1—16 implements recursive inference on the states, hence messages 12 and 16 carry approximate estimates $q(x_t^{(1)}) \approx p(x_t^{(1)}|\mathbf{y}_{1:t})$ and $q(x_t^{(2)}) \approx p(x_t^{(2)}|\mathbf{y}_{1:t})$ respectively. Note that the "problematic" GCV node f_1 communicates with the rest of the model through the update rules as specified by Table 1.For each time segment we initialize $l = 1$ with $q^{(l)}(x_t^{(i)}) \propto \mathcal{N}(0.0, 100.0)$ and iterate messages 1—16 over l and update the sufficient statistics (see Fig. 6 for free energy vs. number of iterations).

We implement our method in ForneyLab [9], an FFG toolbox that is being developed by our research group. [4]

3.2. Analysis

The exchange rate data (observations) is plotted in the first row of Fig. 7. As a baseline comparison, we also simulated the two-layer HGF with the TAPAS toolbox [3]. The belief over the "volatility" $x_t^{(2)}$ for the TAPAS and FFG simulations are plotted in the second and third rows of Fig. 7 respectively. Note the strong similarity between the state estimates.

We are also interested in measuring model performance and used the variational free energy as a performance metric. In Fig. 6 we plot the free energy as a function of VMP iteration number for both the TAPAS- and FFG-based simulations. Note that a lower free energy in principle points to a better model performance. The performance of both simulations level out after about 6 iterations with a performance ad-

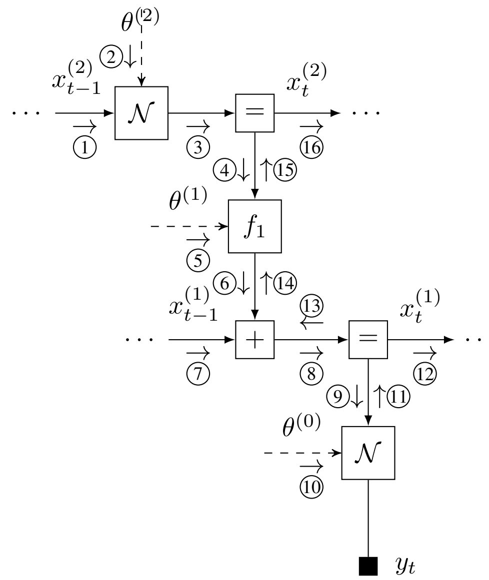


Fig. 5. Message passing schedule for online state estimation in the two-layer HGF. Messages are computed in the order as indicated by the numbers.

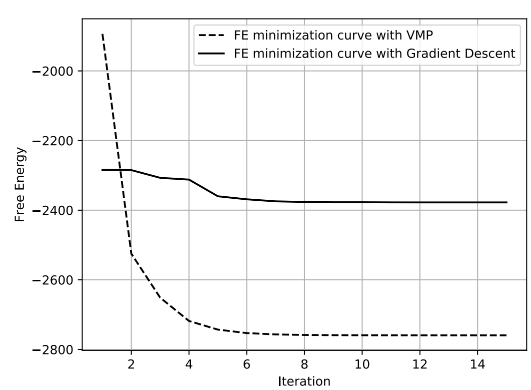


Fig. 6. Model performance scoring by the Free energy as a function of iteration number for the two-layer HGF filter for online state estimation.

[2]Full derivations of update rules can be found at http://biaslab.github.io/pdf/mlsp2018/senoz_mlsp_2018_supplement.pdf

[3]Data is taken from http://www.macrotrends.net/2558/us-dollar-swiss-franc-exchange-rate-historical-chart

[4]ForneyLab.jl is available at https://github.com/biaslab/ForneyLab.jl

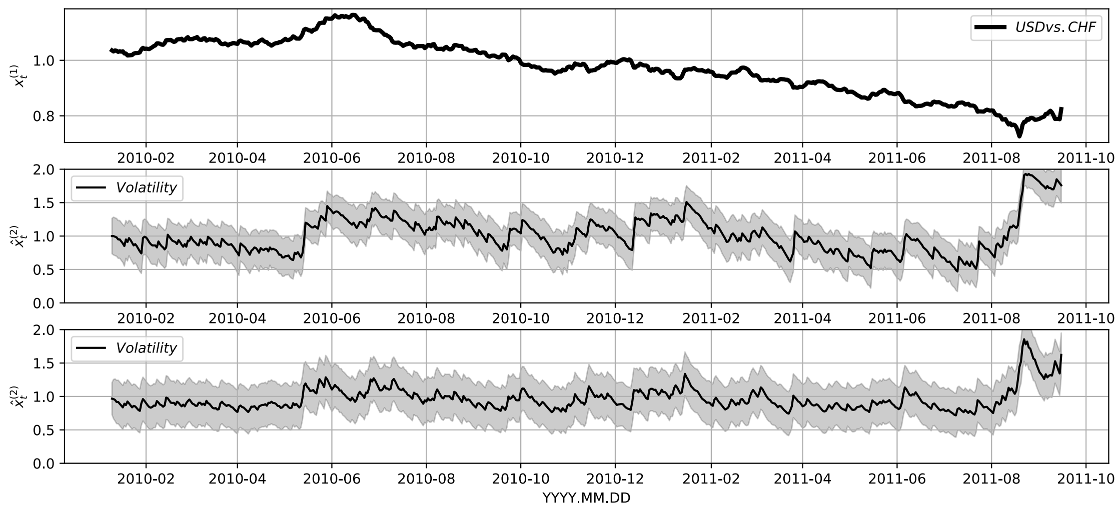


Fig. 7. Simulation results. The top row shows the daily USD-CHF currency exchange rate. The second and last row display the beliefs (mean $\pm$ standard deviation) over volatility $q(x_t^{(2)})$ for simulations by the TAPAS and FFG toolboxes respectively.

vantage for the FFG framework. This plot supports our claim that VMP update equations indeed minimize FE. Even though the bound obtained with the FFG framework is tighter than the bound obtained by the TAPAS simulations, this single experiment is not enough to conclude that our inference scheme is superior to that of the original scheme. We are planning to extend our investigations on this topic.

4. CONCLUSIONS

In this paper we have cast the Hierarchical Gaussian Filter in a factor graph framework and derived local (variational) message passing update rules for the nonlinear connection factor between the layers in this model. Moreover, we derived formulae for the local free energy contributions of the connection factor. As a result, both online state and parameter estimation as well as performance tracking of the HGF or any variants thereof can now be automatically simulated in a software toolbox that stores the results (Table 1) in a lookup table.

5. REFERENCES

[1] Christoph D. Mathys, *Hierarchical Gaussian filtering*, Ph.D. thesis, Diss., Eidgenoessische Technische Hochschule ETH Zuerich, Nr. 20909, 2012.

[2] Sandra Iglesias, Christoph D. Mathys, Kay H. Brodersen, Lars Kasper, Marco Piccirelli, Hanneke E. M. den Ouden, and Klaas E. Stephan, "Hierarchical Prediction Errors in Midbrain and Basal Forebrain during Sensory Learning," *Neuron*, vol. 80, no. 2, pp. 519–530, Oct. 2013.

[3] Christoph D. Mathys, "TAPAS - Translational Algorithms for Psychiatry-Advancing Science," `https://github.com/translationalneuromodeling/tapas`, Apr. 2008, original-date: 2017-01-14T13:14:23Z.

[4] Christoph D. Mathys, Jean Daunizeau, Karl J. Friston, and Stephan E. Klaas, "A Bayesian foundation for individual learning under uncertainty," *Frontiers in Human Neuroscience*, vol. 5, 2011.

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

[6] Sascha Korl, *A factor graph approach to signal modelling, system identification and filtering*, Ph.D. thesis, Swiss Federal Institute of Technology, Zurich, 2005.

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

[8] Hans-Andrea Loeliger, Justin Dauwels, Junli Hu, Sascha Korl, Li Ping, and Frank R. Kschischang, "The Factor Graph Approach to Model-Based Signal Processing," *Proceedings of the IEEE*, vol. 95, no. 6, pp. 1295–1322, June 2007.

[9] Marco Cox, Thijs van de Laar, and Bert de Vries, "ForneyLab.jl: Fast and flexible automated inference through message passing in Julia," *Under Review*, Apr. 2018.