Quasifiltering for time-series modeling

Tsyplakov, Alexander

Novosibirsk State University

10 July 2015

Quasifiltering for time-series modeling

Alexander Tsyplakov
Department of Economics, Novosibirsk State University

July 10, 2015

Abstract

In the paper a method for constructing new varieties of time-series models is proposed. The idea is to start from an unobserved components model in a state-space form and use it as an inspiration for development of another time-series model, in which time-varying underlying variables are directly observed. The goal is to replace a state-space model with an intractable likelihood function by another model, for which the likelihood function can be written in a closed form. If state transition equation of the parent state-space model is linear Gaussian, then the resulting model would belong to the class of score driven model (aka GAS, DCS).

1 Introduction

One can use relatively simple time-series models to bring richer dynamics into some other model. Direct observations for the former are not available, thus, the corresponding elementary dynamic processes are called unobserved components. This is a convenient way of formulating new time-series models. The unobserved components are frequently of Markov class. The most popular variant is a first-order autoregression with Gaussian errors.

One way of obtaining unobserved components models is to take some parameters, which are initially static, and make them time-varying. For example, a very simple level plus noise model can be modified by assuming time-varying level and variance. Coefficients of seasonal dummies can be made time-varying to take into account changing seasonal pattern. A typical application of time-varying parameters approach to macroeconomic modeling is Cogley and Sargent (2005). In Harvey (1989) a "construction set" approach to building time series models is advocated and the resulting models are called "structural time series models" (see also Harvey; 2006). The elements of the standard construction set are stochastic trends, seasonals, cycles, etc., which are directly interpretable in substantial terms. The term "unobserved components model" in a narrow sense is a synonym of a structural time series model, which can be decomposed into such elementary processes. However in this paper we use the term in a broader sense of a model based on underlying latent processes.

An unobserved components model can be cast it into a canonical form called state-space form. The variables of such a model are divided into two groups: observed $\mathbf{y}_t$ and unobserved $\mathbf{a}_t$. The dynamic behavior of the *state variable* $\mathbf{a}_t$ is governed by a process with a (conditionally) Markov structure, while the distribution of $\mathbf{y}_t$ depends only on $\mathbf{a}_t$ and its own previous history, but not on the previous history of $\mathbf{a}_t$.

Although for a time series model in a state-space form there exists a toolkit of standard methods, in general one needs some kind of numerical integration to deal with such a model (when the state variable is continuous). Only for very narrow classes of state-space models integration can be done in a closed form, notably for linear Gaussian models equipped with the

famous Kalman filter algorithm. Even a minor modification can bring a tractable model into an analytically intractable class. Numerical integration can be computationally demanding. Similar to any approximation, there is a tradeoff between the accuracy of approximation and the amount of computation. Monte Carlo techniques reduce the curse of dimensionality only partially.

In summary, from the point of view of an applied researcher unobserved components are very attractive means of model formulation. At the same time they burden the researcher with a load of computational problems.

An alternative approach is to add dynamic features in such a way that the resulting underlying variables are observable conditionally on previous observed history, static parameters and initial conditions. An illuminating example is given by volatility modeling with stochastic volatility (SV) models. Although the basic SV model has a slick and natural formulation, it does not possess a tractable likelihood, that is why in applications it is dominated by a somewhat less natural GARCH with modifications. Both models have their volatility variables, but SV volatility is unobservable, while GARCH volatility is governed rigidly by the explored time series, which makes GARCH more suitable for applied research.

Following categorization in Cox (1981) the models obtained by this second approach are labeled *observation driven* as opposed to *parameter driven*. An approach to formulation of such observation driven models is proposed in Creal et al. (2008), Creal et al. (2013) under the name of GAS (generalized autoregressive score) and, independently, Harvey and Chakravarty (2008), Harvey (2013) under the name of DCS (dynamic conditional score).

By connecting score driven models to unobserved components models, the current paper provides some theoretical grounds for the former. The grounds are mostly informal, but they make construction of score-driven models a less ad hoc process.

One of the drawbacks of the existing approach to score driven modeling is arbitrariness of scaling of the score in the dynamic process for the underlying factors. Creal et al. (2013) propose several variants of scaling matrices, however, the choice is largely ad hoc. The current paper proposes more rigid principles of choosing scaling matrices. The idea is to derive them from the parent unobserved component model in state-space form.

When constructing an observation driven model inspired by an unobserved component model one would typically do various simplifications to make the descendant model more tractable. The main goal is to obtain a model described by closed form recursive formulas without any computationally demanding aspects such as numerical integration or numerical optimization, but further simplifications are also permitted. If one believes the parent unobserved component model to be the *true* one, then the various approximations and simplifications can lead to the loss of estimators' consistency, deterioration of model fit and forecast ability and should be done only if one is ready to pay this price. However, for real-life data there is no such thing as "the true model". It may well be that a computationally simpler roughened model is better in terms of goodness of fit and/or forecast ability.

The various simplified models derived from unobserved components model in a state-space form can be called *quasifilters* due to their resemblance to the corresponding proper filtering techniques such as the Kalman filter. Naturally, most of the known score-driven models can be considered as quasifilters. Indeed, Harvey (2013) draws many explicit parallels with state-space models and Kalman filter.

The quasifilter roots can be found in several seemingly unrelated areas such as volatility models of GARCH type, the extended Kalman filter and exponential smoothing techniques. For example, quasifilter logic explains informally the need for using fat-tailed distributions in the models of GARCH type.

This paper introduces two types of approximations, which can be utilized in state updating and which thus underlie the construction of quasifilters from the parent state-space models.

2 Filtering in a general state-space model

2.1 Formulation of a general state-space model

Let $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_T)$ be an observed (univariate or multivariate) series. A typical observation $\mathbf{y}_t$ is a $k_t \times 1$ vector. The model for the $\mathbf{y}$ series is formulated in terms of the state series $\mathbf{a} = (\mathbf{a}_1, \ldots, \mathbf{a}_T)$, where $\mathbf{a}_t$ is a $m_t \times 1$ vector of unobserved components. The joint distribution of $\mathbf{y}$ and $\mathbf{a}$ is known up to some vector of parameters $\boldsymbol{\theta}$: $f(\mathbf{y}, \mathbf{a}) = f(\mathbf{y}, \mathbf{a} \mid \boldsymbol{\theta})$. Below we suppress the dependence on $\boldsymbol{\theta}$. We assume $\mathbf{a}$ to be continuous. To simplify exposition we accept the convention that $\mathbf{y}$ is also continuous. However, discrete or mixed $\mathbf{y}$ can be treated in a similar manner.

The overall density $f(\mathbf{y}, \mathbf{a})$ of a general state-space model is constructed from two series of densities (all of which are parametric and depend on $\boldsymbol{\theta}$):

- measurement density $f(\mathbf{y}_t \mid \mathbf{a}_{1:t}, \mathbf{y}_{1:t-1}) = f(\mathbf{y}_t \mid \mathbf{a}_t, \mathbf{y}_{1:t-1})$, $\quad t = 1, \ldots, T$;
- transition density $f(\mathbf{a}_t \mid \mathbf{a}_{1:t-1}, \mathbf{y}_{1:t-1}) = f(\mathbf{a}_t \mid \mathbf{a}_{t-1}, \mathbf{y}_{1:t-1})$, $\quad t = 2, \ldots, T$.

We also need $f(\mathbf{a}_1)$ to be specified. It can be viewed as a special case of the transition density for $t = 1$. Note that the measurement density does not depend on $\mathbf{a}_{1:t-1}$. Similarly, the transition density does not depend on $\mathbf{a}_{1:t-2}$ and thus the model has a conditionally Markov transition given the previous history $\mathbf{y}_{1:t-1}$.

2.2 Filtering in a general state-space model

What can be the objectives of filtering in a state-space model?

First, filtering can be used as a device for computing the values of the likelihood function for given values of parameters $\boldsymbol{\theta}$. This function can be used to obtain maximum likelihood estimates for $\boldsymbol{\theta}$. The likelihood function is the density $f(\mathbf{y})$ viewed as a function of $\boldsymbol{\theta}$. Filtering provides a factorization of the likelihood function

$$f(\mathbf{y}) = \prod_{t=1}^{T} f(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}),$$

where $f(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$ are contributions of individual observations to the overall likelihood.

Second, of interest can be the conditional densities for the state variables $f(\mathbf{a}_t \mid \mathbf{y}_{1:t})$, $f(\mathbf{a}_t \mid \mathbf{y}_{1:t-1})$ and various predictions obtained from them. Usually these predictions can be represented as expectations of functions of the state variable; for example,

$$\mathsf{E}[h(\mathbf{a}_t) \mid \mathbf{y}_{1:t-1}] = \int h(\mathbf{a}_t) f(\mathbf{a}_t \mid \mathbf{y}_{1:t-1}) d\mathbf{a}_t.$$

In what follows we are primarily interested in some analogues of $f(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$, while analogues of $f(\mathbf{a}_t \mid \mathbf{y}_{1:t})$ and $f(\mathbf{a}_t \mid \mathbf{y}_{1:t-1})$ play an auxiliary role.

For a general state-space model $f(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$, $f(\mathbf{a}_t \mid \mathbf{y}_{1:t})$ and $f(\mathbf{a}_t \mid \mathbf{y}_{1:t-1})$ can be obtained in a recursive way. Cf. Kitagawa (1987), Harvey (2006), Creal (2012). Suppose that at time t the previous filtering density $f(\mathbf{a}_{t-1} \mid \mathbf{y}_{1:t-1})$ is already known. Filtering recursion is usually represented as iterating prediction step and updating step.

Prediction step:

$$f(\mathbf{a}_t \mid \mathbf{y}_{1:t-1}) = \int f(\mathbf{a}_t \mid \mathbf{a}_{t-1}, \mathbf{y}_{1:t-1}) f(\mathbf{a}_{t-1} \mid \mathbf{y}_{1:t-1}) d\mathbf{a}_{t-1}.$$

Here $f(\mathbf{a}_{t-1} \mid \mathbf{y}_{1:t-1})$ comes from the previous period updating step, while $f(\mathbf{a}_t \mid \mathbf{a}_{t-1}, \mathbf{y}_{1:t-1})$ is specified by the model.

Updating step:

$$f(\mathbf{a}_t \mid \mathbf{y}_{1:t}) = \frac{f(\mathbf{y}_t \mid \mathbf{a}_t, \mathbf{y}_{1:t-1}) f(\mathbf{a}_t \mid \mathbf{y}_{1:t-1})}{f(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})},$$

where

$$f(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}) = \int f(\mathbf{y}_t \mid \mathbf{a}_t, \mathbf{y}_{1:t-1}) f(\mathbf{a}_t \mid \mathbf{y}_{1:t-1}) d\mathbf{a}_t$$

is the contribution to the likelihood. Here $f(\mathbf{a}_t \mid \mathbf{y}_{1:t-1})$ comes from the prediction step, while $f(\mathbf{y}_t \mid \mathbf{a}_t, \mathbf{y}_{1:t-1})$ is specified by the model.

2.3 Approximate filtering

In what follows we change notation and denote functions and variables associated with true densities by letters with circle subscript while the corresponding approximations by letters without such subscript.

Conditional density of the state series $\mathbf{a}$ given the observed series $\mathbf{y}$, that is, $f_\circ(\mathbf{a} \mid \mathbf{y}) = f_\circ(\mathbf{y}, \mathbf{a})/f_\circ(\mathbf{y})$, is called the smoothing density. Smoothing uses all observations available at time T. A (full data) smoothing approximation is some function $f(\mathbf{a} \mid \mathbf{y})$, which approximates $f_\circ(\mathbf{a} \mid \mathbf{y})$.

Filtering refers to a situation when observations of $\mathbf{y}_t$ arrive one by one. At time t only $\mathbf{y}_{1:t}$ is used for inference about $\mathbf{a}_{1:t}$. Similarly to the full data smoothing one can consider a series of partial smoothing problems based on observations $1, \ldots, t$. Approximate filtering can be based on a series of approximations $f(\mathbf{a}_{1:t} \mid \mathbf{y}_{1:t})$ to $f_\circ(\mathbf{a}_{1:t} \mid \mathbf{y}_{1:t}) = f_\circ(\mathbf{y}_{1:t}, \mathbf{a}_{1:t})/f_\circ(\mathbf{y}_{1:t})$ with last-period approximate filtering densities $f(\mathbf{a}_t \mid \mathbf{y}_{1:t})$, predictive densities $f(\mathbf{a}_t \mid \mathbf{y}_{1:t-1})$ and contributions to the likelihood $f(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$ produced as a byproduct.

However, dealing directly with batch approximations $f(\mathbf{a}_{1:t} \mid \mathbf{y}_{1:t})$ can be difficult due to growing dimensionality. A simpler piecemeal approach to approximate filtering does not keep track of densities $f(\mathbf{a}_{1:t} \mid \mathbf{y}_{1:t})$ explicitly. With this approach in the approximate filtering step of time t only $f_\circ(\mathbf{a}_t \mid \mathbf{y}_{1:t-1})$, $f_\circ(\mathbf{a}_t \mid \mathbf{y}_{1:t})$ and $f_\circ(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$ are approximated by $f(\mathbf{a}_t \mid \mathbf{y}_{1:t-1})$, $f(\mathbf{a}_t \mid \mathbf{y}_{1:t})$ and $f(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$ given the previous period approximation $f(\mathbf{a}_{t-1} \mid \mathbf{y}_{1:t-1})$. The price of such a piecemeal approach is that the approximation error can accumulate from period to period.

Many different methods of approximate piecemeal filtering were proposed in the literature. These include approximating densities by step functions (ordinary numerical integration), by (weighted) averages of Dirac delta-functions corresponding to random samples (particle filters) and so on.

For the goals of genuine approximate filtering the approximations used should be accurate and closely reproduce true densities. For quasifiltering which we consider further there is no such goal. Quasifiltering is some loose imitation of the genuine filtering.

3 Basic quasifilter recursion

In the derivation of our basic quasifilter we assume that the conditional densities of the state variables are approximately Gaussian, so that $f_\circ(\mathbf{a}_{t-1} \mid \mathbf{y}_{1:t-1})$ and $f_\circ(\mathbf{a}_t \mid \mathbf{y}_{1:t-1})$ are approximated by $\varphi(\mathbf{a}_{t-1} - \bar{\mathbf{a}}_{t-1}, \bar{\mathbf{P}}_{t-1})$ and $\varphi(\mathbf{a}_t - \tilde{\mathbf{a}}_t, \tilde{\mathbf{P}}_t)$ respectively, where $\varphi(\mathbf{x}, \Sigma)$ is the density at $\mathbf{x}$ of the multivariate normal distribution with zero mean and covariance matrix Σ. Transition distribution is assumed to be Gaussian with the conditional mean which is linear in $\mathbf{a}_{t-1}$, that is,

$$\mathbf{a}_t \mid \mathbf{a}_{t-1}, \mathbf{y}_{1:t-1} \sim \mathcal{N}(\mathbf{R}_{at} + \mathbf{R}_{aat}\mathbf{a}_{t-1}, \Omega_{at})$$

In Section 8 we extend the quasifilter approach to the case of mildly nonlinear and/or non-Gaussian transition.

Prediction step The prediction step of the basic quasifilter is known from the Kalman filter and is given by

$$\begin{aligned}\tilde{\mathbf{a}}_t &= \mathbf{R}_{at} + \mathbf{R}_{aat}\bar{\mathbf{a}}_{t-1},\\ \tilde{\mathbf{P}}_t &= \mathbf{R}_{aat}\bar{\mathbf{P}}_{t-1}\mathbf{R}_{aat}^{\intercal} + \Omega_{at}.\end{aligned}$$

Updating step The Gaussian approximation $f(\mathbf{a}_t \mid \mathbf{y}_{1:t-1}) = \varphi(\mathbf{a}_t - \tilde{\mathbf{a}}_t, \tilde{\mathbf{P}}_t)$ for $f_\circ(\mathbf{a}_t \mid \mathbf{y}_{1:t-1})$ produces an approximate contribution to the likelihood for time t given by

$$f_\sharp(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}) = \int f_\circ(\mathbf{y}_t \mid \mathbf{a}_t, \mathbf{y}_{1:t-1})\varphi(\mathbf{a}_t - \tilde{\mathbf{a}}_t, \tilde{\mathbf{P}}_t)d\mathbf{a}_t.$$

We introduce the following notation for the corresponding log-density, which can be viewed an approximation to the log-likelihood $\ell_{\circ t} = \ln f_\circ(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$ for observation t:

$$\ell_{\sharp t} = \ln f_\sharp(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}).$$

Below we are primarily interested in dependence of $\ell_{\sharp t}$ on $\tilde{\mathbf{a}}_t$, so $\ell_{\sharp t} = \ell_{\sharp t}(\tilde{\mathbf{a}}_t)$ with dependence on $\mathbf{y}_t$, static parameters $\boldsymbol{\theta}$, $\tilde{\mathbf{P}}_t$ and $\mathbf{y}_{1:t-1}$ from the measurement density being implicit.

By analogy with

$$f_\circ(\mathbf{a}_t \mid \mathbf{y}_{1:t}) = \frac{f_\circ(\mathbf{y}_t \mid \mathbf{a}_t, \mathbf{y}_{1:t-1})f_\circ(\mathbf{a}_t \mid \mathbf{y}_{1:t-1})}{f_\circ(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})}$$

we can write

$$f_\sharp(\mathbf{a}_t \mid \mathbf{y}_{1:t}) = \exp(-\ell_{\sharp t})f_\circ(\mathbf{y}_t \mid \mathbf{a}_t, \mathbf{y}_{1:t-1})\varphi(\mathbf{a}_t - \tilde{\mathbf{a}}_t, \tilde{\mathbf{P}}_t),$$

where $f_\sharp(\mathbf{a}_t \mid \mathbf{y}_{1:t})$ is the approximation to filtering density implied by $\varphi(\mathbf{a}_t - \tilde{\mathbf{a}}_t, \tilde{\mathbf{P}}_t)$ as an approximation of the prediction density $f_\circ(\mathbf{a}_t \mid \mathbf{y}_{1:t-1})$. By construction it is a proper density function with unit integral.

The moments of the approximate filtering distribution are obtained by integration with respect to $f_\sharp(\mathbf{a}_t \mid \mathbf{y}_{1:t})$. In particular, the filtering estimate of $\mathbf{a}_t$ implied by $\varphi(\mathbf{a}_t - \tilde{\mathbf{a}}_t, \tilde{\mathbf{P}}_t)$ is given by

$$\mathsf{E}_{\sharp t}\,\mathbf{a}_t = \int f_\sharp(\mathbf{a}_t \mid \mathbf{y}_{1:t})\mathbf{a}_t d\mathbf{a}_t,$$

where $\mathsf{E}_{\sharp t}$ denotes the corresponding expectation operator. The corresponding variance-covariance matrix is

$$\mathsf{var}_{\sharp t}\,\mathbf{a}_t = \mathsf{E}_{\sharp t}[(\mathbf{a}_t - \mathsf{E}_{\sharp t}\,\mathbf{a}_t)(\mathbf{a}_t - \mathsf{E}_{\sharp t}\,\mathbf{a}_t)^{\intercal}].$$

The following proposition provides an informal foundation for our basic quasifilter by suggesting a non-obvious relation between the approximate log-likelihood $\ell_{\sharp t}$ and the approximate filtering distribution with density $f_\sharp(\mathbf{a}_t \mid \mathbf{y}_{1:t})$.[1]

Proposition 1. *The mean and covariance matrix of the approximate filtering distribution can be expressed as*

$$\mathsf{E}_{\sharp t}\,\mathbf{a}_t = \tilde{\mathbf{a}}_t + \tilde{\mathbf{P}}_t\nabla\ell_{\sharp t}(\tilde{\mathbf{a}}_t)$$

and

$$\mathsf{var}_{\sharp t}\,\mathbf{a}_t = \tilde{\mathbf{P}}_t + \tilde{\mathbf{P}}_t\nabla^2\ell_{\sharp t}(\tilde{\mathbf{a}}_t)\tilde{\mathbf{P}}_t.$$

[1] This resembles a result obtained in Masreliez (1975).

The derivation is placed in Appendix. In this proposition $\nabla \ell_{\sharp t}(\tilde{\mathbf{a}}_t) = \partial \ell_{\sharp t}(\tilde{\mathbf{a}}_t)/\partial \tilde{\mathbf{a}}_t$ can be recognized as the score vector and $\nabla^2 \ell_{\sharp t}(\tilde{\mathbf{a}}_t)$ as the Hessian matrix corresponding to the time t approximate log-likelihood $\ell_{\sharp t}$. It is important that application of these formulas does not require the knowledge of measurement density of the parent model $f_\circ(\mathbf{y}_t \mid \mathbf{a}_t, \mathbf{y}_{1:t-1})$. One needs only $\ell_{\sharp t}$.

In general we do not know closed-form formulas for $\ell_{\sharp t}$. Instead a suitable approximation $\ell_t = \ell_t(\tilde{\mathbf{a}}_t)$ would be used in a quasifilter. The corresponding filtering approximation is given by $\mathcal{N}(\bar{\mathbf{a}}_t, \bar{\mathbf{P}}_t)$, where

$$\bar{\mathbf{a}}_t = \tilde{\mathbf{a}}_t + \tilde{\mathbf{P}}_t \mathbf{s}_t, \qquad \mathbf{s}_t = \nabla \ell_t(\tilde{\mathbf{a}}_t) \tag{1}$$

and

$$\bar{\mathbf{P}}_t = \tilde{\mathbf{P}}_t - \tilde{\mathbf{P}}_t \mathbf{N}_t \tilde{\mathbf{P}}_t. \tag{2}$$

Here $\mathbf{N}_t$ can be the negated Hessian of ℓ_t, that is,

$$\mathbf{N}_t = -\nabla^2 \ell_t(\tilde{\mathbf{a}}_t),$$

or some other suitable approximation. Since $\bar{\mathbf{P}}_t$ represents the covariance matrix of the approximate filtering distribution, $\mathbf{N}_t$ should be chosen in such a way that $\bar{\mathbf{P}}_t$ is positive definite whenever $\tilde{\mathbf{P}}_t$ is positive definite.

Matrix $\tilde{\mathbf{P}}_t$ is used to scale score vector $\mathbf{s}_t$ in the state updating formula. Since in quasifiltering $\bar{\mathbf{P}}_t$ and $\tilde{\mathbf{P}}_t$ can be some very loose approximations to the true covariance matrices, we call them just scaling matrices.

4 Possible approaches and examples

4.1 Log-likelihood approximations

The key ingredient of a quasifilter is the contribution to the log-likelihood. We do not know the true contribution to to the log-likelihood of the parent state-space model $\ell_{\circ t}$ and use some suitable approximation ℓ_t instead. The piecemeal nature of quasifiltering implies that we do not have enough information to assess the quality of ℓ_t as an approximation to $\ell_{\circ t}$. However, we have some information to assess the quality of ℓ_t as an approximation to $\ell_{\sharp t} = \ln f_\sharp(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$, where

$$f_\sharp(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}) = \int f_\circ(\mathbf{y}_t \mid \mathbf{a}_t, \mathbf{y}_{1:t-1}) \varphi(\mathbf{a}_t - \tilde{\mathbf{a}}_t, \tilde{\mathbf{P}}_t) d\mathbf{a}_t.$$

This is also an approximation to the true $f_\circ(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$ with Gaussian density $\varphi(\mathbf{a}_t - \tilde{\mathbf{a}}_t, \tilde{\mathbf{P}}_t)$ supplanting unknown $f_\circ(\mathbf{a}_t \mid \mathbf{y}_{1:t-1})$. As such it can only give a suggestion for choosing ℓ_t. However, such a suggestion can be very valuable as it can help to choose the functional form of ℓ_t.

In general a closed form expression for $f_\sharp(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$ would be unavailable. For some models the moments of $f_\sharp(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$ could be known in a closed form. In general for exploratory purposes one can use simulations. For example, for a sample $\mathbf{a}_t^1, \ldots, \mathbf{a}_t^S$ from $\mathcal{N}(\tilde{\mathbf{a}}_t, \tilde{\mathbf{P}}_t)$ a Monte Carlo approximation to $f_\sharp(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$ is given by

$$f_\sharp(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}) \approx \frac{1}{S} \sum_{s=1}^{S} f_\circ(\mathbf{y}_t \mid \mathbf{a}_t^s, \mathbf{y}_{1:t-1}).$$

There are numerous possibilities in deriving ℓ_t from $\ell_{\sharp t}$.

- Derive ℓ_t as an approximation to $\ell_{\sharp t}$ by matching characteristics of ℓ_t to these of $\ell_{\sharp t}$ in a pure analytic manner.

- Use a parametric family for $\ell_t = \ell_t(\boldsymbol{\psi})$ and estimate the corresponding parameters $\boldsymbol{\psi}$ using a Monte Carlo sample. For example, choose $\boldsymbol{\psi}$ to (approximately) solve the maximization problem

$$\max_{\boldsymbol{\psi}} \int \ell_t(\boldsymbol{\psi}; \mathbf{y}_t) f_\sharp(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}) d\mathbf{y}_t.$$

  The objective function here is related to the Kulback–Leibler distance between $f_\sharp(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$ and $\exp(\ell_t)$. The estimation should be done beforehand and parameters $\boldsymbol{\psi}$ should be expressed by closed-form formulas so that quasifilter is not slowed down by simulations. Note that in general ℓ_t depends on $\tilde{\mathbf{a}}_t$, $\tilde{\mathbf{P}}_t$, static parameters $\boldsymbol{\theta}$ and previous observed history $\mathbf{y}_{1:t-1}$, so that optimized $\boldsymbol{\psi}$ can be a function of all these variables.

- Use a parametric family for ℓ_t with parameters $\boldsymbol{\psi}$ and append these parameters to the parameters of the initial state-space model $\boldsymbol{\theta}$ so that $(\boldsymbol{\psi}, \boldsymbol{\theta})$ is the resulting parameter vector for the quasifilter model to be estimated jointly given the observed data.

All of these approaches need some additional efforts. A quick-and-dirty alternative is to use the measurement log-density at $\tilde{\mathbf{a}}_t$ as the contribution to the log-likelihood

$$\ell_t = \lambda_t(\tilde{\mathbf{a}}_t),$$

where

$$\lambda_t(\mathbf{a}_t) = \ln f_\circ(\mathbf{y}_t \mid \mathbf{a}_t, \mathbf{y}_{1:t-1}).$$

This can be a reasonable approximation if $\tilde{\mathbf{P}}_t$ is relatively small. However, as we will see below, for some models the result can be rather poor.

4.2 Time-varying scale model

As an example we consider a time-varying scale model (known as stochastic volatility model) given by

$$y_t = e^{h_t/2}\epsilon_t,$$
$$h_t \mid \mathbf{h}_{1:t-1}, \mathbf{y}_{1:t-1} \sim \mathcal{N}(\omega + \delta h_{t-1}, \sigma_h^2).$$

where ϵ_t is an independent identically distributed white noise series with unit variance, e^{h_t} is the time-varying error variance and $\delta \in (0, 1)$ (although $\delta = 1$ is also possible).

Note that if $h_t \mid \mathbf{y}_{1:t-1} \sim \mathcal{N}(\tilde{h}_t, \tilde{p}_t)$ and ϵ_t standard normal or leptokurtic, then $f_\circ(y_t \mid \mathbf{y}_{1:t-1})$ corresponds to a distribution which is symmetric around zero and leptokurtic. The value of $\tilde{h}_t$ determines only the scale of the distribution, but not the shape. Indeed,

$$\tilde{\mathsf{E}}_{t-1} y_t = \tilde{\mathsf{E}}_{t-1} e^{h_t/2} \tilde{\mathsf{E}}_{t-1} \epsilon_t = 0,$$

and

$$\tilde{\mathsf{var}}_{t-1} y_t = \tilde{\mathsf{E}}_{t-1}(e^{h_t}\epsilon_t^2) = \tilde{\mathsf{E}}_{t-1} e^{h_t} \tilde{\mathsf{E}}_{t-1} \epsilon_t^2 = e^{\tilde{h}_t + \tilde{p}_t/2}$$

where expectations are with respect to $f_\circ(y_t \mid h_t, \mathbf{y}_{1:t-1})\varphi(h_t - \tilde{h}_t, \tilde{p}_t)$ and $e^{\tilde{h}_t + \tilde{p}_t/2}$ is the mean of a log-normal variable e^{h_t}. The standardized variant of y_t is thus $e^{(h_t - \tilde{h}_t)/2} e^{-\tilde{p}_t/4}\epsilon_t$, where the conditional distribution of $h_t - \tilde{h}_t$ is $\mathcal{N}(0, \tilde{p}_t)$ and does not depend on $\tilde{h}_t$. The conditional kurtosis of y_t is given by

$$\tilde{\mathsf{E}}_{t-1}[(e^{(h_t - \tilde{h}_t)/2} e^{-\tilde{p}_t/4}\epsilon_t)^4] = \tilde{\mathsf{E}}_{t-1}[e^{2(h_t - \tilde{h}_t)}] e^{-\tilde{p}_t} \tilde{\mathsf{E}}_{t-1}(\epsilon_t^4) = e^{2\tilde{p}_t} e^{-\tilde{p}_t} \tilde{\mathsf{E}}_{t-1}(\epsilon_t^4) = e^{\tilde{p}_t} \tilde{\mathsf{E}}_{t-1}(\epsilon_t^4).$$

This demonstrates that the conditional kurtosis of y_t is almost surely greater than the conditional kurtosis of ϵ_t. As SV-generated quasifilter is in a class of models similar to GARCH,

this observation suggests an explanation to the widespread use of fat-tailed disturbances in GARCH-type models (cf. Bollerslev, 1997).

Following the approach popular in GARCH modeling we approximate the conditional distribution of y_t by the Student's t distribution with ν_t degrees of freedom, and scale $q_t e^{\tilde{h}_t/2}$, where q_t is some coefficient. Denote the scaled residuals by

$$T_t = \frac{y_t}{q_t} e^{-\tilde{h}_t/2}.$$

Here we assume that T_t has the ordinary Student's distribution. Then the contribution to the log-likelihood is

$$\ell_t = \ln \Gamma\left(\frac{\nu_t + 1}{2}\right) - \ln \Gamma\left(\frac{\nu_t}{2}\right) - \frac{1}{2}\ln(\pi\nu_t) - \frac{\nu_t + 1}{2}\ln\left(1 + \frac{T_t^2}{\nu_t}\right) - \frac{\tilde{h}_t}{2} - \ln q_t$$

and the basic quasifilter recursions are

$$\tilde{h}_{t+1} = \omega + \delta(\tilde{h}_t + \tilde{p}_t s_t), \qquad s_t = \frac{\partial \ell_t}{\partial \tilde{h}_t} = \frac{1}{2}\frac{\nu_t(T_t^2 - 1)}{\nu_t + T_t^2},$$

$$\tilde{p}_{t+1} = \delta^2(\tilde{p}_t - \tilde{p}_t^2 N_t) + \sigma_h^2, \qquad N_t = -\frac{\partial^2 \ell_t}{\partial \tilde{h}_t^2} = \frac{1}{2}\frac{\nu_t(\nu_t + 1)}{(\nu_t + T_t^2)^2} T_t^2.$$

Possible strategies include:

A Assume $\epsilon_t \sim \mathcal{N}(0, 1)$ and $\ell_t = \ln f_\circ(y_t \mid h_t, \mathbf{y}_{1:t-1})\,\big|_{h_t = \tilde{h}_t} = \ln \varphi(y_t, e^{\tilde{h}_t/2})$ ("quick-and-dirty" approach), which corresponds to $\nu_t = +\infty$ and $q_t = 1$.

B Express ν_t and q_t as functions of $\tilde{p}_t$ by estimating the corresponding parametric models on Monte Carlo data prior to estimating the model itself.

C Express ν_t and q_t as functions of $\tilde{p}_t$ and estimate parameters of these functions together with other parameters of the model (ω, δ, σ_h).

D Fix $\nu_t = \nu$, $q_t = 1$ and treat ν as a parameter of the model.

Tables 1 and 2 show maximum likelihood estimation results for the four models, corresponding to these strategies. For approaches B and C we take

$$\ln \nu_t = \psi_1 + \psi_2 \ln \tilde{p}_t,$$
$$\ln q_t = \psi_3 \tilde{p}_t + \psi_4/\nu_t.$$

All of the strategies potentially have a problem with positivity of the variance variable $\tilde{p}_t$. However, only for approach A this problem does materialize. The estimates in column A were actually produced with

$$\tilde{p}_{t+1} = \delta^2 \min\{\tilde{p}_t - \tilde{p}_t^2 N_t, 0\} + \sigma_h^2$$

recursion for the variance, which is obviously quite an ugly workaround.

Table 1: Generated SV

	A	B	C	D	E	F
ψ_1	—	$0.975^{\dagger}$	2.261 (0.72)	—	—	—
ψ_2	—	$-0.92^{\dagger}$	−0.003 (0.56)	—	—	—
ψ_3	$0^{\dagger}$	$0.227^{\dagger}$	0.285 (0.69)	$0^{\dagger}$	$0^{\dagger}$	$0^{\dagger}$
ψ_4	$0^{\dagger}$	$-1^{\dagger}$	−3.80 (4.5)	$0^{\dagger}$	$0^{\dagger}$	$0^{\dagger}$
ω	0.0101 (0.0024)	0.0059 (0.0027)	0.0169 (0.0156)	0.0044 (0.0026)	0.0105 (0.0027)	0.0048 (0.0027)
δ	0.980 (0.0028)	0.982 (0.0031)	0.980 (0.0036)	0.982 (0.0033)	0.978 (0.0029)	0.980 (0.0033)
σ_h.	0.153 (0.0077)	0.192 (0.0113)	0.201 (0.0166)	0.192 (0.0122)	0.175 (0.0104)	0.199 (0.0132)
ν	$+\infty^{\dagger}$	*9.01 (1.345)*	*9.63 (0.005)*	9.76 (1.0)	$+\infty^{\dagger}$	9.59 (1.0)
Max LL	−9882.2	−9805.4	−9802.8	−9803.5	−9886.4	−9806.2
AIC	3.295	3.269	3.270	3.269	3.296	3.270
BIC	3.298	3.273	3.278	3.274	3.300	3.275

Note: The SV series was generated with $\omega = 0$, $\delta = 0.98$, $\sigma_h = 0.2$, standard normal ϵ_t, $T = 6000$ observations. Standard errors in brackets; † marks fixed parameters. For models B and C the numbers in the ν row are means and standard deviations of the model ν_t series (in italics). Infinite ν was approximated by $\nu = 1000000$. Model E (F) is similar to A (respectively, D), but uses the information matrix instead of the negated Hessian (which is explained in subsection 4.3).

Table 2: FTSE100

	A	B	C	D	E	F
ψ_1	—	$0.975^{\dagger}$	5.221 (0.96)	—	—	—
ψ_2	—	$-0.92^{\dagger}$	1.662 (0.54)	—	—	—
ψ_3	$0^{\dagger}$	$0.227^{\dagger}$	1.977 (0.68)	$0^{\dagger}$	$0^{\dagger}$	$0^{\dagger}$
ψ_4	$0^{\dagger}$	$-1^{\dagger}$	0.39 (1.4)	$0^{\dagger}$	$0^{\dagger}$	$0^{\dagger}$
ω	−0.0021 (0.0015)	−0.0030 (0.0017)	−0.0228 (0.0102)	−0.0037 (0.0018)	−0.0021 (0.0015)	−0.0036 (0.0018)
δ	0.984 (0.0027)	0.986 (0.0027)	0.981 (0.0042)	0.987 (0.0028)	0.983 (0.0026)	0.986 (0.0028)
σ_h	0.121 (0.0078)	0.133 (0.0095)	0.164 (0.0186)	0.131 (0.0102)	0.119 (0.0087)	0.131 (0.0105)
ν	$+\infty^{\dagger}$	*13.13 (1.730)*	*14.48 (3.777)*	13.92 (1.8)	$+\infty^{\dagger}$	15.76 (2.3)
Max LL	−9338.7	−9297.6	−9286.5	−9294.7	−9324.3	−9282.8
AIC	2.724	2.712	2.710	2.711	2.720	2.708
BIC	2.727	2.715	2.717	2.715	2.723	2.712

Note: FTSE100 daily returns for the period from 1984-05-03 to 2011-06-30, 6859 observations. See the note to Table 1 for further explanation.

4.3 I-scaling: using information matrix instead of negated Hessian

For some models it is convenient to use the information matrix corresponding to ℓ_t as $\mathbf{N}_t$. The matrix is given by the expectation of the negated Hessian $-\nabla^2\ell_t$ under the distribution of $\mathbf{y}_t$ implied by ℓ_t, that is,

$$\mathbf{I}_t = -\int \frac{\partial^2 \ell_t(\tilde{\mathbf{a}}_t; \mathbf{y}_t)}{\partial \tilde{\mathbf{a}}_t \partial \tilde{\mathbf{a}}_t^\top} \exp(\ell_t(\tilde{\mathbf{a}}_t; \mathbf{y}_t)) d\mathbf{y}_t.$$

Here the dependence of ℓ_t on $\mathbf{y}_t$ has to be shown explicitly. Alternatively, it can be obtained as the covariance matrix of the score vector $\mathbf{s}_t = \nabla\ell_t$ under the same distribution of $\mathbf{y}_t$, that is,

$$\mathbf{I}_t = \int \frac{\partial \ell_t(\tilde{\mathbf{a}}_t; \mathbf{y}_t)}{\partial \tilde{\mathbf{a}}_t} \frac{\partial \ell_t(\tilde{\mathbf{a}}_t; \mathbf{y}_t)}{\partial \tilde{\mathbf{a}}_t^\top} \exp(\ell_t(\tilde{\mathbf{a}}_t; \mathbf{y}_t)) d\mathbf{y}_t.$$

That these two alternative expressions give the same result is the information matrix identity known from the maximum likelihood estimation theory. The use of the information matrix $\mathbf{I}_t$ instead of the negated Hessian in quasifilter scaling recursions can be called I-scaling as opposed to H-scaling.

There are at least two reasons for using the information matrix instead of the negated Hessian. First, using the information matrix can ensure positive definiteness of $\bar{\mathbf{P}}_t$ for some models and choices of ℓ_t. Second, frequently, the expression for the information matrix is much simpler than the expression for the negated Hessian. For example, we can obtain block-diagonal $\mathbf{N}_t$, which allows to keep the scaling matrices $\tilde{\mathbf{P}}_t$ and $\bar{\mathbf{P}}_t$ block-diagonal for some models.

For the time-varying scale example of subsection 4.2 one can set $N_t = I_t$, where

$$I_t = \frac{\nu_t}{2(\nu_t + 3)},$$

since

$$\mathsf{E}\left[\frac{T^2}{(\nu + T^2)^2}\right] = \frac{1}{(\nu + 1)(\nu + 3)} \quad \text{for } T \sim t_\nu.$$

Setting $\nu_t = \nu$, $q_t = 1$ as in approach D above gives a model, which is simpler than the model produced by D (column F in Tables 1 and 2). In the same way one can simplify the model produced by approach A, which corresponds to $\nu_t = +\infty$, $q_t = 1$ (column E).

Note that if we assume that $\nu_t \geq 1$ and $\sigma_h^2 < 1/2$, then $\tilde{p}_t \in (0, 2)$ implies $\bar{p}_t = \tilde{p}_t - \tilde{p}_t^2 I_t > 0$ and $\tilde{p}_{t+1} \in (0, 2)$. Thus, the use of information matrix can ensure that the scaling series remain positive. In particular, unlike model A, model E for our two empirical examples is not affected by the problem of negative variances.

4.4 I-scaling for Gaussian nonlinear measurement

Another example of I-scaling illustrates simplification of covariance matrix recursions. Suppose that the measurement density is Gaussian, that is,

$$\mathbf{y}_t \mid \mathbf{a}_t, \mathbf{y}_{1:t-1} \sim \mathcal{N}(\mathbf{g}_{yt}(\mathbf{a}_t), \Omega_{yt}),$$

where $\mathbf{g}_{yt}(\mathbf{a}_t)$ is a smooth nonlinear function. If $\mathbf{a}_t \mid \mathbf{y}_{1:t-1} \sim \mathcal{N}(\tilde{\mathbf{a}}_t, \tilde{\mathbf{P}}_t)$, then by using linearization around $\tilde{\mathbf{a}}_t$ we obtain that approximately

$$\mathbf{y}_t \mid \mathbf{y}_{1:t-1} \sim \mathcal{N}(\mathbf{g}_{yt}, \Sigma_{yt}),$$

where $\mathbf{g}_{yt} = \mathbf{g}_{yt}(\tilde{\mathbf{a}}_t)$, $\Sigma_{yt} = \Sigma_{yt}(\tilde{\mathbf{a}}_t) = \nabla\mathbf{g}_{yt}^\top \tilde{\mathbf{P}}_t \nabla\mathbf{g}_{yt} + \Omega_{yt}$, $\nabla\mathbf{g}_{yt} = \nabla\mathbf{g}_{yt}(\tilde{\mathbf{a}}_t) = \partial\mathbf{g}_{yt}^\top(\tilde{\mathbf{a}}_t)/\partial\tilde{\mathbf{a}}_t$. Thus, the approximate log-likelihood is

$$\ell_t = \varphi(\mathbf{y}_t - \mathbf{g}_{yt}, \Sigma_{yt}) = -\frac{1}{2}\ln|\Sigma_{yt}| - \frac{1}{2}(\mathbf{y}_t - \mathbf{g}_{yt})^\top \Sigma_{yt}^{-1}(\mathbf{y}_t - \mathbf{g}_{yt}) + const.$$

The elements of $\mathbf{s}_t$ are given by

$$\frac{\partial \ell_t}{\partial \tilde{a}_{tj}} = -\frac{1}{2}\operatorname{tr}\left(\frac{\partial \Sigma_{yt}}{\partial \tilde{a}_{tj}}\Sigma_{yt}^{-1}\right) + \frac{1}{2}(\mathbf{y}_t - \mathbf{g}_{yt})^\intercal \Sigma_{yt}^{-1}\frac{\partial \Sigma_{yt}}{\partial \tilde{a}_{tj}}\Sigma_{yt}^{-1}(\mathbf{y}_t - \mathbf{g}_{yt}) + (\mathbf{y}_t - \mathbf{g}_{yt})^\intercal \Sigma_{yt}^{-1}\frac{\partial \mathbf{g}_{yt}}{\partial \tilde{a}_{tj}}.$$

The expression for the Hessian matrix is quite complicated. However, one can simplify things by using the information matrix instead with elements given by

$$\mathbf{I}_t = \frac{1}{2}\operatorname{tr}\left(\frac{\partial \Sigma_{yt}}{\partial \tilde{a}_{tj}}\Sigma_{yt}^{-1}\frac{\partial \Sigma_{yt}}{\partial \tilde{a}_{tk}}\Sigma_{yt}^{-1}\right) + \frac{\partial \mathbf{g}_{yt}^\intercal}{\partial \tilde{a}_{tj}}\Sigma_{yt}^{-1}\frac{\partial \mathbf{g}_{yt}}{\partial \tilde{a}_{tk}}.$$

Note that these formulas differ from the well-known extended Kalman filter. To reproduce the formulas of the EKF one should assume that the derivatives $\partial \Sigma_{yt}/\partial \tilde{a}_{tj}$ are relatively small so that the last terms would dominate in the expressions for the score and information matrix:

$$\mathbf{s}_t \approx \nabla \mathbf{g}_{yt}^\intercal \Sigma_{yt}^{-1}(\mathbf{y}_t - \mathbf{g}_{yt}), \qquad \mathbf{I}_t \approx \nabla \mathbf{g}_{yt}^\intercal \Sigma_{yt}^{-1}\nabla \mathbf{g}_{yt}.$$

4.5 C-scaling

Consider the ordinary linear Gaussian state-space model

$$\begin{aligned}\mathbf{y}_t \mid \mathbf{a}_t, \mathbf{y}_{1:t-1} &\sim \mathcal{N}(\mathbf{R}_{yt} + \mathbf{R}_{yat}\mathbf{a}_t, \Omega_t^y),\\ \mathbf{a}_t \mid \mathbf{a}_{t-1}, \mathbf{y}_{1:t-1} &\sim \mathcal{N}(\mathbf{R}_{at} + \mathbf{R}_{aat}\mathbf{a}_{t-1}, \Omega_{at}).\end{aligned}$$

In the Kalman filter corresponding to this model we have the following recursion for the covariance matrices:

$$\tilde{\mathbf{P}}_{t+1} = \mathbf{R}_{aa,t+1}(\tilde{\mathbf{P}}_t - \tilde{\mathbf{P}}_t\mathbf{R}_{yat}^\intercal(\mathbf{R}_{yat}\tilde{\mathbf{P}}_t\mathbf{R}_{yat}^\intercal + \Omega_{yt})^{-1}\mathbf{R}_{yat}\tilde{\mathbf{P}}_t)\mathbf{R}_{aa,t+1}^\intercal + \Omega_{a,t+1}.$$

If time variation of the coefficients matrices $\mathbf{R}_{yat}$, Ω_{yt}, $\mathbf{R}_{aat}$, Ω_{at} has some suitable pattern, the recursions in the limit can produce covariance matrices with a stable pattern. That is, $\tilde{\mathbf{P}}_t \approx \mathbf{S}_t\tilde{\mathbf{P}}\mathbf{S}_t^\intercal$ for some fixed positive definite matrix $\tilde{\mathbf{P}}$ and a sequence of known matrices $\mathbf{S}_t$, so that the difference between $\tilde{\mathbf{P}}_t$ and $\mathbf{S}_t\tilde{\mathbf{P}}\mathbf{S}_t^\intercal$ vanishes as $t \to \infty$. Then one can replace $\tilde{\mathbf{P}}_t$ by $\mathbf{S}_t\tilde{\mathbf{P}}\mathbf{S}_t^\intercal$ in the Kalman filter recursions. In particular we can have $\tilde{\mathbf{P}}_t \approx \tilde{\mathbf{P}}$ (setting $\mathbf{S}_t = \mathbf{I}$), when $\mathbf{R}_{yat}$, Ω_{yt}, $\mathbf{R}_{aat}$, Ω_{at} are time-invariant, so that

$$\tilde{\mathbf{P}} = \mathbf{R}_{aa}(\tilde{\mathbf{P}} - \tilde{\mathbf{P}}\mathbf{R}_{ya}^\intercal(\mathbf{R}_{ya}\tilde{\mathbf{P}}\mathbf{R}_{ya}^\intercal + \Omega_y)^{-1}\mathbf{R}_{ya}\tilde{\mathbf{P}})\mathbf{R}_{aa}^\intercal + \Omega_a$$

is an equation, for which $\tilde{\mathbf{P}}$ is a solution. This is so called discrete-time algebraic Riccati equation. Replacing $\tilde{\mathbf{P}}_t$ by $\tilde{\mathbf{P}}$ is a standard approximation used in Kalman filtering. It produces a steady-state filter (discussed, for example, in Simon; 2006).

Similar simplifications can be utilized in quasifilters based on some nonlinear and/or non-Gaussian state-space models. Harvey (2013) propose to use this idea in DSC models.

Consider a simple case when the state variable is univariate. The transition equation is given by

$$a_t = \omega + \delta a_{t-1} + \sigma_a \eta_t,$$

with independent standard normal innovations η_t. That is,

$$a_t \mid \mathbf{a}_{1:t-1}, \mathbf{y}_{1:t-1} \sim \mathcal{N}(\omega + \delta a_{t-1}, \sigma_a^2).$$

The quasifilter recursions for such a model can be written as

$$\tilde{a}_{t+1} = \omega + \delta(\tilde{a}_t + \tilde{p}_t s_t) \qquad \text{for } s_t = \partial \ell_t/\partial a_t,$$

$$\tilde{p}_{t+1} = \delta^2(\tilde{p}_t - \tilde{p}_t^2 N_t) + \sigma_a^2,$$

If N_t depends only on $\tilde{p}_t$, that is, $N_t = N(\tilde{p}_t)$, then the steady-state variance $\tilde{p}$ (if it exists) is a solution to the following equation:

$$\tilde{p} = \delta^2(\tilde{p} - \tilde{p}^2 N(\tilde{p})) + \sigma_a^2.$$

Replacing $\tilde{p}_t$ by $\tilde{p}$ we obtain

$$\tilde{a}_{t+1} = \omega + \delta(\tilde{a}_t + \tilde{p}s_t).$$

In such a model we can use $\tilde{p}$ rather than σ_a^2 as a parameter to be estimated. Another possibility is to estimate γ in

$$\tilde{a}_{t+1} = \omega + \delta\tilde{a}_t + \gamma s_t.$$

In particular, one can use this trick in the time-varying scale example above. It can be readily seen that if one uses $\nu_t = \nu$, $q_t = 1$ and I-scaling, then the result is equivalent to beta-t-EGARCH model of Harvey and Chakravarty (2008) and Harvey (2013). A similar model is used as an illustration of GAS in Creal et al. (2013). Note that in Tables 1 and 2 the case F is indistinguishable from beta-t-EGARCH, because the recursions for the state variance quickly converge to a steady-state value.

Even in the cases when $\tilde{\mathbf{P}}_t$ would not converge to a steady-state value, it can be useful to set $\tilde{\mathbf{P}}_t = \tilde{\mathbf{P}}$ and thereby simplify the model by economizing on the number of recursive equations.

One can use a known function to represent $\tilde{\mathbf{P}}$ and estimate parameters of this function. Using such function for quasifilter scaling can be called C-scaling (which stands for "constant scaling"). More generally, the use of scaling matrix $\tilde{\mathbf{P}}_t$ which is a known function of t can be also called, by extension, C-scaling. The main difference form I-scaling and H-scaling is that C-scaling is not based on recursions.

4.6 C-scaling for time-varying level and seasonality model

Consider the following simple model of time-varying level and seasonality (with M seasons)

$$\begin{aligned} y_t &= \mu_t + \gamma_{t1} + \sigma_y \epsilon_t, \\ \mu_t &= \mu_{t-1} + \sigma_\mu \eta_{\mu t}, \\ \boldsymbol{\gamma}_t &= \mathbf{R}_{\gamma\gamma}\boldsymbol{\gamma}_{t-1} + \sigma_\gamma \boldsymbol{\eta}_{\gamma t}, \end{aligned}$$

Here μ_t represents the time-varying level and $\boldsymbol{\gamma}_t = (\gamma_{t1}, \ldots, \gamma_{tM})^\intercal$ represents M seasonal components. Matrix

$$\mathbf{R}_{\gamma\gamma} = \begin{pmatrix} \mathbf{0}_{M-1}^\intercal & 1 \\ \mathbf{I}_{M-1} & \mathbf{0}_{M-1} \end{pmatrix}$$

circularly permutes the seasonal components, so that the current season corresponds to the first component. The error terms are independent, ϵ_t and $\eta_{\mu t}$ are standard normal, while $\boldsymbol{\eta}_{\gamma t}$ is a zero-sum vector distributed as $\boldsymbol{\eta}_{\gamma t} \sim \mathcal{N}(\mathbf{0}_M, \mathbf{I}_M - \frac{1}{M}\mathbf{1}_{M\times M})$. If the sum of the seasonal components is zero at $t = 1$, then the sum remains zero for all future periods $t = 2, 3, \ldots$ by construction. We can further assume that in the first period $\tilde{\boldsymbol{\gamma}}_t$, which is the estimate of $\boldsymbol{\gamma}_t$, has zero sum.

This model is linear Gaussian and can be readily estimated by the ordinary Kalman filter. Table 3 shows the estimates for the logarithms of the monthly dairy products production in Spain for the period 1980–2013. In this example the covariance matrix $\tilde{\mathbf{P}}_t$ converges quickly enough to a steady-state limit $\tilde{\mathbf{P}}$. This observation suggests using C-scaling.

Unfortunately, there seems to be no easy way to find the steady-state scaling matrix $\tilde{\mathbf{P}}$ except for solving the corresponding Riccati equation, which is also not straightforward. Potentially C-scaling can be implemented by estimating a $(M+1)\times(M+1)$ matrix $\tilde{\mathbf{P}}$. However, even after taking into account the necessary restrictions on $\tilde{\mathbf{P}}$ we are left with too many unknown parameters. This compares unfavorably with the original formulation, where there are just two major transition parameters (σ_μ and σ_γ).

According to the Kalman filter formulas, we have $y_t \mid \mathbf{y}_{1:t-1} \sim \mathcal{N}(\tilde{y}_t, \tilde{F}_t)$, where

$$\tilde{y}_t = \mathbf{R}_{ya}\tilde{\mathbf{a}}_t,$$
$$\tilde{F}_t = \mathbf{R}_{ya}\tilde{\mathbf{P}}_t\mathbf{R}_{ya}^\intercal + \sigma_y^2,$$
$$\mathbf{R}_{ya} = (1, 1, \mathbf{0}_{M-1}^\intercal).$$

The score vector for time t is given by

$$\mathbf{s}_t = \mathbf{R}_{ya}^\intercal \frac{1}{\tilde{F}_t}(y_t - \tilde{y}_t)$$

and thus the updating equation is

$$\bar{\mathbf{a}}_t = \tilde{\mathbf{a}}_t + \tilde{\mathbf{P}}_t\mathbf{s}_t = \tilde{\mathbf{a}}_t + \tilde{\mathbf{P}}_t\mathbf{R}_{ya}^\intercal \frac{1}{\tilde{F}_t}(y_t - \tilde{y}_t).$$

As $\tilde{\mathbf{P}}_t$ converges to $\tilde{\mathbf{P}}$, the vector of coefficients $\tilde{\mathbf{P}}_t\mathbf{R}_{ya}^\intercal\frac{1}{\tilde{F}_t}$ converges to

$$\mathbf{n} = \tilde{\mathbf{P}}\mathbf{R}_{ya}^\intercal \frac{1}{\tilde{F}}, \quad \tilde{F} = \mathbf{R}_{ya}\tilde{\mathbf{P}}\mathbf{R}_{ya}^\intercal + \sigma_y^2.$$

Thus, with C-scaling we obtain

$$\bar{\mathbf{a}}_t = \tilde{\mathbf{a}}_t + (y_t - \tilde{y}_t)\mathbf{n}.$$

We can estimate the elements of $\mathbf{n}$, but this still gives too many unknown parameters for large M.

A possible simplification is to set

$$\tilde{\mathbf{P}} = \tilde{F}\begin{pmatrix} \alpha & \mathbf{0}_M^\intercal \\ \mathbf{0}_M & \beta\left(\mathbf{I} - \frac{1}{M}\mathbf{1}_{M\times M}\right)\end{pmatrix}, \quad \tilde{F} = \sigma_y^2/\left(1 - \alpha - \left(1 - \frac{1}{M}\right)\beta\right),$$

which produces

$$\mathbf{n} = \left(\alpha, \left(1 - \frac{1}{M}\right)\beta, -\frac{1}{M}\beta, \ldots, -\frac{1}{M}\beta\right)^\intercal,$$

and estimate unknown α and β. The structure of this vector somewhat resembles the structure of the original $\mathbf{n}$. Figure 1 plots the values of $\mathbf{n}$ for the Spanish dairy products example. The value of the first seasonal coefficient is large, while other seasonal coefficients are relatively small. Moreover, the seasonal coefficients sum to zero. Otherwise the pattern is different, because in the original $\mathbf{n}$ the small seasonal coefficients are described by some nonlinear curve. Some of the coefficients are negative while other are positive and in general they are far from being equal.

With this simplification the updating equations are as follows:

$$\bar{\mu}_t = \tilde{\mu}_t + \alpha(y_t - \tilde{y}_t),$$
$$\bar{\gamma}_{1t} = \tilde{\gamma}_{1t} + \left(1 - \frac{1}{M}\right)\beta(y_t - \tilde{y}_t),$$
$$\bar{\gamma}_{jt} = \tilde{\gamma}_{jt} - \frac{1}{M}\beta(y_t - \tilde{y}_t), \qquad j = 2, \ldots, M$$

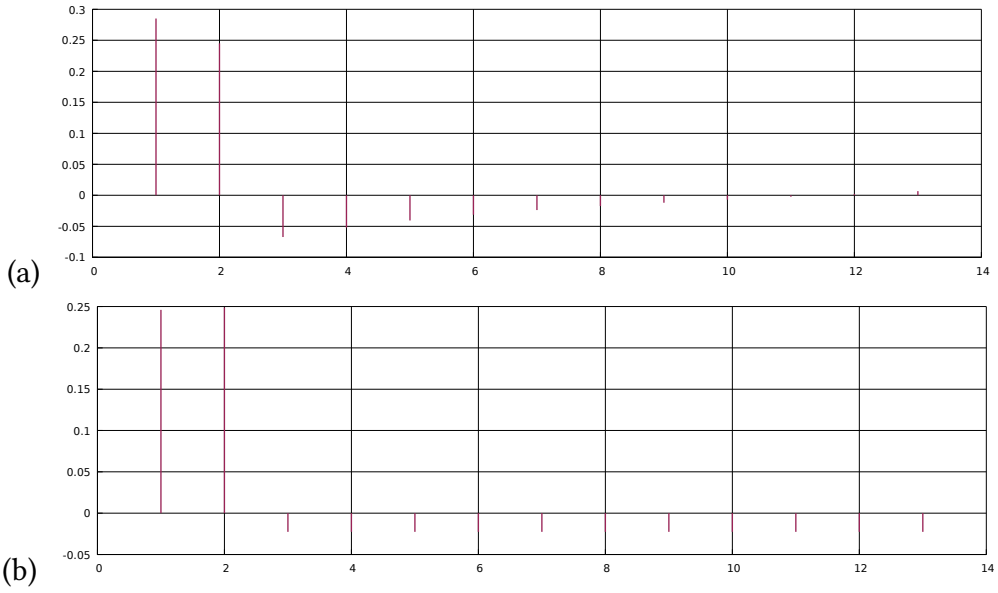

Figure 1: The elements of vector $\mathbf{n}$ for the Spanish dairy products example; (a) the original limiting vector based on the Kalman filter; (b) estimated vector with C-scaling the equal-weight simplification.

If $\sum_{j=1}^{M} \tilde{\gamma}_{jt} = 0$, then $\sum_{j=1}^{M} \bar{\gamma}_{jt} = 0$. Thus the proposed simplification preserves the sum of seasonal components to be zero. Such equal-weight "normalization" (method of keeping the seasonality centered) is known from the exponential smoothing literature; see Archibald and Koehler (2003) and references therein. See also paragraph 3.6.4 in Harvey (2013), where a similar ad hoc device is suggested for DCS models.

Note that we have one seasonal variable for each season. One can further simplify the model by reducing the number of seasonal variables. Define recursively a variable which accumulates the terms required for correcting seasonality:

$$r_{t+1} = r_t + \frac{1}{M}\beta(y_t - \tilde{y}_t), \qquad r_1 = 0,$$

and define uncorrected variables for the level and seasonality:

$$\mu_t^* = \tilde{\mu}_t - r_t,$$
$$\gamma_t^* = \tilde{\gamma}_{1t} + r_t.$$

These uncorrected variables can be described by the following recursions:

$$\mu_{t+1}^* = \mu_t^* + \Big(\alpha - \frac{1}{M}\beta\Big)(y_t - \tilde{y}_t),$$
$$\gamma_{t+M}^* = \gamma_t^* + \beta(y_t - \tilde{y}_t).$$

Here we have only one uncorrected seasonal variable. See Archibald and Koehler (2003) for a similar correction in an exponential smoothing model with time-varying level, trend and seasonality (a modification of the additive Holt–Winters model).

Therefore, with the above ad hoc simplification of the vector of coefficients $\mathbf{n}$, we obtain recursions, which in essence represent a kind of additive exponential smoothing in the Holt–Winters style. The links between exponential smoothing and state-space models have long been recognized; cf. Harvey (2006). Interestingly, the quasifilter logic goes in a reverse direction than the logic in Hyndman et al. (2008), a monograph specifically emphasizing the links between two kinds of models. Hyndman et al. (2008) represent an exponential smoothing

Table 3: Estimates for the time-varying level and seasonality model

Kalman filter		simplified C-scaling	
σ_y	0.032 (0.0018)	σ_y	0.034 (0.0020)
σ_μ	0.013 (0.0017)	α	0.246 (0.031)
σ_γ	0.0038 (0.00055)	β	0.272 (0.043)
Max LL	673.0	Max LL	660.8
AIC	−3.285	AIC	−3.220
BIC	−3.255	BIC	−3.180

Note: The data used is the manufacture of dairy products in Spain reported by Eurostat, 2010 = 100. Based on $T = 408$ monthly observations for the period from 1980-01 to 2013-12. Standard errors in brackets.

model as a state-space model with single source of randomness (or "innovations state space model"). Here we start from a state-space model with multiple sources of randomness and obtain recursions driven by a single innovations series $y_t - \tilde{y}_t$, which can be viewed as a kind of exponential smoothing.

Our conjecture is that many popular exponential smoothing models can be viewed as quasifilter models. The quasifilter logic can be used to derive exponential smoothing models from more natural unobserved components models.

The results of estimation of two models on the Spanish dairy products series are shown in Table 3. The initial state-space model estimated using the Kalman filter is characterized by a considerably better fit, while the number of parameters of the two models are the same. However, the internal state updating formulas for the quasifilter model with the simplified C-scaling are much simpler. Thus, here we have a trade-off between model fit and simplicity.

5 A quasifilter for time-varying regression

Consider a time-varying regression model

$$y_t = \mathbf{x}_t^\intercal \boldsymbol{\beta}_t + e^{h_t/2}\epsilon_t,$$

$$\begin{pmatrix}\boldsymbol{\beta}_t \\ h_t\end{pmatrix}\Big|\boldsymbol{\beta}_{1:t-1}, \mathbf{h}_{1:t-1}, \mathbf{y}_{1:t-1} \sim \mathcal{N}\left(\begin{pmatrix}\boldsymbol{\beta}_{t-1} \\ h_{t-1}\end{pmatrix}, \begin{pmatrix}\Omega_{\beta t} & \mathbf{0} \\ \mathbf{0}^\intercal & \omega_{ht}^2\end{pmatrix}\right),$$

where ϵ_t is an independent identically distributed white noise series with unit variance, e^{h_t} is the time-varying error variance. Additional variables $\mathbf{x}_t$ are assumed to be fixed for simplicity (however, random exogenous variables or lags of y_t are not a problem). More generally one can assume here a richer dynamics for the coefficients, replacing $\boldsymbol{\beta}_{t-1}$ in the role of the conditional mean of $\boldsymbol{\beta}_t$ by

$$\mathbf{R}_{\beta t} + \mathbf{R}_{\beta\beta t}\boldsymbol{\beta}_{t-1}.$$

For example, one can "dampen" the dynamics of the regression coefficients by multiplying them by coefficients, which are less than 1, and adding constants to the transition equations.

If the conditional distribution of $\mathbf{a}_t = \begin{pmatrix}\boldsymbol{\beta}_t \\ h_t\end{pmatrix}$ given previous history is multivariate normal,

$$\begin{pmatrix}\boldsymbol{\beta}_t \\ h_t\end{pmatrix}\Big|\mathbf{y}_{1:t-1} \sim \mathcal{N}\left(\begin{pmatrix}\tilde{\boldsymbol{\beta}}_t \\ \tilde{h}_t\end{pmatrix}, \begin{pmatrix}\tilde{\mathbf{P}}_{\beta t} & \cdot \\ \cdot & \tilde{p}_{ht}\end{pmatrix}\right)$$

(with dots replacing covariances, which are not important for our derivation), then the first two conditional moments of y_t are

$$\tilde{\mathsf{E}}_{t-1}y_t = \mathbf{x}_t^\intercal \tilde{\mathsf{E}}_{t-1}\boldsymbol{\beta}_t + \tilde{\mathsf{E}}_{t-1}e^{h_t/2}\tilde{\mathsf{E}}_{t-1}\epsilon_t = \mathbf{x}_t^\intercal \tilde{\boldsymbol{\beta}}_t$$

and

$$\begin{aligned}\tilde{\mathrm{var}}_{t-1} y_t &= \tilde{\mathsf{E}}_{t-1}[(y_t - \mathbf{x}_t^\intercal \tilde{\boldsymbol{\beta}}_t)^2] = \tilde{\mathsf{E}}_{t-1}[(\mathbf{x}_t^\intercal(\boldsymbol{\beta}_t - \tilde{\boldsymbol{\beta}}_t) + e^{h_t/2}\epsilon_t)^2] \\ &= \mathbf{x}_t^\intercal \tilde{\mathsf{E}}_{t-1}[(\boldsymbol{\beta}_t - \tilde{\boldsymbol{\beta}}_t)(\boldsymbol{\beta}_t - \tilde{\boldsymbol{\beta}}_t)^\intercal]\mathbf{x}_t + 2\mathbf{x}_t^\intercal \tilde{\mathsf{E}}_{t-1}[(\boldsymbol{\beta}_t - \tilde{\boldsymbol{\beta}}_t)e^{h_t/2}]\tilde{\mathsf{E}}_{t-1}\epsilon_t + \tilde{\mathsf{E}}_{t-1}e^{h_t}\tilde{\mathsf{E}}_{t-1}\epsilon_t^2 \\ &= e^{\tilde{h}_t + \tilde{p}_{ht}/2} + \mathbf{x}_t^\intercal \tilde{\mathbf{P}}_{\boldsymbol{\beta} t}\mathbf{x}_t = c_t^2,\end{aligned}$$

where expectations are with respect to $f_\circ(\mathbf{y}_t \mid \mathbf{a}_t, \mathbf{y}_{1:t-1})\varphi(\mathbf{a}_t - \tilde{\mathbf{a}}_t, \tilde{\mathbf{P}}_t)$, $e^{\tilde{h}_t + \tilde{p}_{ht}/2}$ is the mean of a log-normal variable e^{h_t} and

$$c_t = \sqrt{e^{\tilde{h}_t + \tilde{p}_{ht}/2} + \mathbf{x}_t^\intercal \tilde{\mathbf{P}}_{\boldsymbol{\beta} t}\mathbf{x}_t}.$$

In a typical application ϵ_t would be normal or just symmetric and moderately leptokurtic. If $\tilde{\mathbf{P}}_t$ is block-diagonal between $\boldsymbol{\beta}_t$ and h_t, then the two components of the state vector are conditionally independent. Then the conditional distribution of y_t is symmetric around $\mathbf{x}_t^\intercal \tilde{\boldsymbol{\beta}}_t$ and conditionally leptokurtic since the conditional kurtosis of y_t is given by

$$\frac{\tilde{\mathsf{E}}_{t-1}[(\mathbf{x}_t^\intercal(\boldsymbol{\beta}_t - \tilde{\boldsymbol{\beta}}_t) + e^{h_t/2}\epsilon_t)^4]}{(e^{\tilde{h}_t + \tilde{p}_{ht}/2} + \mathbf{x}_t^\intercal \tilde{\mathbf{P}}_{\boldsymbol{\beta} t}\mathbf{x}_t)^2} = 3 + \frac{e^{2\tilde{h}_t + \tilde{p}_{ht}}}{(e^{\tilde{h}_t + \tilde{p}_{ht}/2} + \mathbf{x}_t^\intercal \tilde{\mathbf{P}}_{\boldsymbol{\beta} t}\mathbf{x}_t)^2}(e^{\tilde{p}_{ht}}\tilde{\mathsf{E}}_{t-1}[{\epsilon_t}^4] - 3).$$

.

We approximate the conditional distribution of y_t by the Student's t distribution with ν degrees of freedom, location $\mathbf{x}_t^\intercal \tilde{\boldsymbol{\beta}}_t$ and scale Ac_t, where A is some coefficient. The distribution has variance $\frac{\nu}{\nu-2}A^2c_t^2$. To equate the variances, we could set $A = \sqrt{\frac{\nu-2}{\nu}}$, but below a more convenient choice of A is proposed.

Denote

$$T_t = \frac{y_t - \mathbf{x}_t^\intercal \tilde{\boldsymbol{\beta}}_t}{Ac_t}.$$

Here T_t has the ordinary Student's distribution. The contribution to the log-likelihood is

$$\ell_t = \ln \mathrm{tden}(T_t, \nu) - \ln c_t - \ln A,$$

where $\mathrm{tden}(\cdot)$ is the Student's t density, so that

$$\ln \mathrm{tden}(T, \nu) = \ln \Gamma\left(\frac{\nu+1}{2}\right) - \ln \Gamma\left(\frac{\nu}{2}\right) - \frac{1}{2}\ln(\pi\nu) - \frac{\nu+1}{2}\ln\left(1 + \frac{T^2}{\nu}\right).$$

The components of the corresponding score vector are

$$\frac{\partial \ell_t}{\partial \tilde{\boldsymbol{\beta}}_t} = \frac{(\nu+1)T_t}{\nu + T_t^2}\frac{1}{Ac_t}\mathbf{x}_t,$$

$$\frac{\partial \ell_t}{\partial \tilde{h}_t} = \frac{1}{2}\frac{\nu(T_t^2 - 1)}{\nu + T_t^2}\frac{1}{c_t^2}e^{\tilde{h}_t + \tilde{p}_{ht}/2}.$$

In order to simplify the model we use I-scaling. The blocks of the information matrix are given by

$$\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\beta} t} = \frac{\nu+1}{\nu+3}\frac{1}{A^2c_t^2}\mathbf{x}_t\mathbf{x}_t^\intercal, \qquad \mathbf{I}_{\boldsymbol{\beta} ht} = \mathbf{0}, \qquad I_{hht} = \frac{1}{2}\frac{\nu}{\nu+3}\frac{1}{c_t^4}e^{2\tilde{h}_t + \tilde{p}_{ht}},$$

where we used that for a t_ν-distributed T

$$\mathsf{E}\left[\frac{T^2}{(\nu + T^2)^2}\right] = \frac{1}{(\nu+1)(\nu+3)}, \qquad \mathsf{E}\left[\frac{(T^2-1)^2}{(\nu + T^2)^2}\right] = \frac{2}{\nu(\nu+3)}.$$

Due to the block-diagonality of the information matrix choosing $\mathbf{P}_1$ to be block-diagonal leads to block-diagonality of $\mathbf{P}_t$. This allows to simplify the quasi-filter recursions:

$$\tilde{\boldsymbol{\beta}}_{t+1} = \tilde{\boldsymbol{\beta}}_t + \tilde{\mathbf{P}}_{\boldsymbol{\beta}t}\frac{(\nu+1)T_t}{\nu+T_t^2}\frac{1}{Ac_t}\mathbf{x}_t,$$

$$\tilde{\mathbf{P}}_{\boldsymbol{\beta},t+1} = \tilde{\mathbf{P}}_{\boldsymbol{\beta}t} - \tilde{\mathbf{P}}_{\boldsymbol{\beta}t}I_{\boldsymbol{\beta}\boldsymbol{\beta}t}\tilde{\mathbf{P}}_{\boldsymbol{\beta}t} + \boldsymbol{\Omega}_{\boldsymbol{\beta},t+1} = \tilde{\mathbf{P}}_{\boldsymbol{\beta}t} - \frac{\nu+1}{\nu+3}\frac{1}{A^2}\tilde{\mathbf{P}}_{\boldsymbol{\beta}t}\mathbf{x}_t\frac{1}{c_t^2}\mathbf{x}_t^\intercal\tilde{\mathbf{P}}_{\boldsymbol{\beta}t} + \boldsymbol{\Omega}_{\boldsymbol{\beta},t+1}.$$

$$\tilde{h}_{t+1} = \tilde{h}_t + \frac{\tilde{p}_{ht}}{2}\frac{\nu(T_t^2-1)}{\nu+T_t^2}\frac{1}{c_t^2}e^{\tilde{h}_t+\tilde{p}_{ht}/2},$$

$$\tilde{p}_{h,t+1} = \tilde{p}_{ht} - \tilde{p}_{ht}^2 I_{hht} + \omega_{h,t+1}^2 = \tilde{p}_{ht} - \frac{1}{2}\tilde{p}_{ht}^2\frac{\nu}{\nu+3}\frac{1}{c_t^4}e^{2\tilde{h}_t+\tilde{p}_{ht}} + \omega_{h,t+1}^2.$$

One can take

$$A = \sqrt{\frac{\nu+1}{\nu+3}}$$

to simplify formulas and to ensure preservation of positive definiteness of $\tilde{\mathbf{P}}_{\boldsymbol{\beta}t}$ matrices. Then

$$\tilde{\mathbf{P}}_{\boldsymbol{\beta},t+1} = \tilde{\mathbf{P}}_{\boldsymbol{\beta}t} - \tilde{\mathbf{P}}_{\boldsymbol{\beta}t}\mathbf{x}_t\frac{1}{c_t^2}\mathbf{x}_t^\intercal\tilde{\mathbf{P}}_{\boldsymbol{\beta}t} + \boldsymbol{\Omega}_{\boldsymbol{\beta},t+1}$$

or, in a more compact form,

$$\tilde{\mathbf{P}}_{\boldsymbol{\beta},t+1} = (\tilde{\mathbf{P}}_{\boldsymbol{\beta}t}^{-1} + e^{-\tilde{h}_t-\tilde{p}_{ht}/2}\mathbf{x}_t\mathbf{x}_t^\intercal)^{-1} + \boldsymbol{\Omega}_{\boldsymbol{\beta},t+1}$$

and

$$\tilde{\boldsymbol{\beta}}_{t+1} = \tilde{\boldsymbol{\beta}}_t + \frac{\sqrt{(\nu+1)(\nu+3)}T_t}{\nu+T_t^2}\frac{1}{c_t}\tilde{\mathbf{P}}_{\boldsymbol{\beta}t}\mathbf{x}_t, \quad \text{where } T_t = \sqrt{\frac{\nu+3}{\nu+1}}\,\frac{y_t - \mathbf{x}_t^\intercal\tilde{\boldsymbol{\beta}}_t}{c_t}.$$

Further simplification can be achieved by setting $\tilde{p}_{ht} = 2\rho$, where ρ is a static parameter. This gives

$$c_t = \sqrt{e^{\tilde{h}_t+\rho} + \mathbf{x}_t^\intercal\tilde{\mathbf{P}}_{\boldsymbol{\beta}t}\mathbf{x}_t}, \qquad T_t = \sqrt{\frac{\nu+3}{\nu+1}}\,\frac{y_t - \mathbf{x}_t^\intercal\tilde{\boldsymbol{\beta}}_t}{c_t},$$

$$\ell_t = \ln \mathrm{tden}(T_t, \nu) - \ln c_t - \frac{1}{2}\ln\left(\frac{\nu+1}{\nu+3}\right),$$

$$\tilde{h}_{t+1} = \tilde{h}_t + \rho\frac{\nu(T_t^2-1)}{\nu+T_t^2}\frac{1}{c_t^2}e^{\tilde{h}_t+\rho},$$

$$\tilde{\mathbf{P}}_{\boldsymbol{\beta},t+1} = \tilde{\mathbf{P}}_{\boldsymbol{\beta}t} - \tilde{\mathbf{P}}_{\boldsymbol{\beta}t}\mathbf{x}_t\frac{1}{c_t^2}\mathbf{x}_t^\intercal\tilde{\mathbf{P}}_{\boldsymbol{\beta}t} + \boldsymbol{\Omega}_{\boldsymbol{\beta},t+1},$$

and

$$\tilde{\boldsymbol{\beta}}_{t+1} = \tilde{\boldsymbol{\beta}}_t + \frac{\sqrt{(\nu+1)(\nu+3)}T_t}{\nu+T_t^2}\frac{1}{c_t}\tilde{\mathbf{P}}_{\boldsymbol{\beta}t}\mathbf{x}_t.$$

6 State-quadratic updating approach

6.1 The general idea of state-quadratic updating

An alternative approach to the updating quasifiler step is based on a quadratic (second order) expansion of the measurement log-density

$$\lambda_t(\mathbf{a}_t) = \ln f_\circ(\mathbf{y}_t \mid \mathbf{a}_t, \mathbf{y}_{1:t-1})$$

treated as a function of the state vector $\mathbf{a}_t$. The expansion is around the point $\mathbf{a}_t = \tilde{\mathbf{a}}_t$ (which already available at the updating step). It produces the following approximation:

$$\lambda_{*t}(\mathbf{a}_t) = \lambda_t(\tilde{\mathbf{a}}_t) + \tilde{\mathbf{s}}_t^\intercal(\mathbf{a}_t - \tilde{\mathbf{a}}_t) - \frac{1}{2}(\mathbf{a}_t - \tilde{\mathbf{a}}_t)^\intercal \tilde{\mathbf{N}}_t(\mathbf{a}_t - \tilde{\mathbf{a}}_t).$$

where

$$\tilde{\mathbf{s}}_t = \nabla\lambda_t(\tilde{\mathbf{a}}_t), \qquad \tilde{\mathbf{N}}_t = -\nabla^2\lambda_t(\tilde{\mathbf{a}}_t).$$

One can use the expansion as a basis for a Gaussian approximation for $f_\circ(\mathbf{a}_t \mid \mathbf{y}_{1:t})$. The approximation provides close-form formulas suitable for use in a quasifilter.[2]

From $\lambda_{*t}(\mathbf{a}_t)$ and $f(\mathbf{a}_t \mid \mathbf{y}_{1:t-1})$ we can obtain $f(\mathbf{a}_t \mid \mathbf{y}_{1:t})$. By construction

$$\ln f(\mathbf{a}_t \mid \mathbf{y}_{1:t-1}) + \lambda_{*t}(\mathbf{a}_t) = \ln f(\mathbf{a}_t \mid \mathbf{y}_{1:t}) + \ell_t,$$

where $\ell_t = \ln f(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$ and $f(\mathbf{a}_t \mid \mathbf{y}_{1:t})$ corresponds to $\mathcal{N}(\bar{\mathbf{a}}_t, \bar{\mathbf{P}}_t)$. That is,

$$\begin{aligned}
-\frac{m_t}{2}\ln(2\pi) - \frac{1}{2}\ln|\tilde{\mathbf{P}}_t| - \frac{1}{2}(\mathbf{a}_t - \tilde{\mathbf{a}}_t)^\intercal \tilde{\mathbf{P}}_t^{-1}(\mathbf{a}_t - \tilde{\mathbf{a}}_t) + \lambda_t(\tilde{\mathbf{a}}_t) + \tilde{\mathbf{s}}_t^\intercal(\mathbf{a}_t - \tilde{\mathbf{a}}_t) - \frac{1}{2}(\mathbf{a}_t - \tilde{\mathbf{a}}_t)^\intercal \tilde{\mathbf{N}}_t(\mathbf{a}_t - \tilde{\mathbf{a}}_t) \\
= -\frac{m_t}{2}\ln(2\pi) - \frac{1}{2}\ln|\bar{\mathbf{P}}_t| - \frac{1}{2}(\mathbf{a}_t - \bar{\mathbf{a}}_t)^\intercal \bar{\mathbf{P}}_t^{-1}(\mathbf{a}_t - \bar{\mathbf{a}}_t) + \ell_t.
\end{aligned}$$

Equating the coefficients of the various powers of $\mathbf{a}_t$ we obtain

$$\begin{aligned}
-\frac{1}{2}\tilde{\mathbf{P}}_t^{-1} - \frac{1}{2}\tilde{\mathbf{N}}_t &= -\frac{1}{2}\bar{\mathbf{P}}_t^{-1}, \\
\tilde{\mathbf{P}}_t^{-1}\tilde{\mathbf{a}}_t + \tilde{\mathbf{s}}_t + \tilde{\mathbf{N}}_t\tilde{\mathbf{a}}_t &= \bar{\mathbf{P}}_t^{-1}\bar{\mathbf{a}}_t, \\
-\frac{1}{2}\ln|\tilde{\mathbf{P}}_t| - \frac{1}{2}\tilde{\mathbf{a}}_t^\intercal\tilde{\mathbf{P}}_t^{-1}\tilde{\mathbf{a}}_t + \lambda_t(\tilde{\mathbf{a}}_t) - \tilde{\mathbf{s}}_t^\intercal\tilde{\mathbf{a}}_t - \frac{1}{2}\tilde{\mathbf{a}}_t^\intercal\tilde{\mathbf{N}}_t\tilde{\mathbf{a}}_t &= -\frac{1}{2}\ln|\bar{\mathbf{P}}_t| - \frac{1}{2}\bar{\mathbf{a}}_t^\intercal\bar{\mathbf{P}}_t^{-1}\bar{\mathbf{a}}_t + \ell_t,
\end{aligned}$$

which produces the following recursions

$$\begin{aligned}
\bar{\mathbf{P}}_t &= (\tilde{\mathbf{P}}_t^{-1} + \tilde{\mathbf{N}}_t)^{-1}, \\
\bar{\mathbf{a}}_t &= \tilde{\mathbf{a}}_t + \bar{\mathbf{P}}_t\tilde{\mathbf{s}}_t.
\end{aligned}$$

Here $\tilde{\mathbf{s}}_t = \nabla\lambda_t(\tilde{\mathbf{a}}_t)$ is the *measurement score vector*, corresponding to "the log-likelihood function" $\tilde{\ell}_t = \ln f_\circ(\mathbf{y}_t \mid \mathbf{a}_t, \mathbf{y}_{1:t-1}) = \lambda_t(\mathbf{a}_t)$. This puts the resulting model in the score driven class of models. However, here the "score" is different from the one used in basic quasifilters.

The approximation to the log-likelihood implied by this approach is given by

$$\ell_t = \lambda_t(\tilde{\mathbf{a}}_t) + \frac{1}{2}\ln|\bar{\mathbf{P}}_t| - \frac{1}{2}\ln|\tilde{\mathbf{P}}_t| + \frac{1}{2}\tilde{\mathbf{s}}_t^\intercal\bar{\mathbf{P}}_t\tilde{\mathbf{s}}_t.$$

[2] Other state-quadratic approximations are possible, but they do not in general provide close-form formulas (e.g. require numerical optimization).

This quantity is in general unusable, since the approximation is usually poor and the implied conditional density of $\mathbf{y}_t$ does not integrate to one. The conclusion is that this alternative approach can not be applied independently. One has to supplement state-quadratic quasi-filter recursions with a suitable log-likelihood ℓ_t. Another reasonable possibility is a hybrid approach, that is, combining scaling matrix updating from the state-quadratic approach with state updating based on derivatives of the log-likelihood ℓ_t as in the basic approach. Which approach is better to use for the state estimate updating is an open question.

For the time-varying scale example of subsection 4.2 with a standard normal ϵ_t we have

$$\lambda_t = -\frac{1}{2}\ln(2\pi) - \frac{h_t}{2} - \frac{1}{2}e^{-h_t}y_t^2,$$

$$\tilde{s}_t = \lambda_t'(\tilde{h}_t) = \frac{1}{2}(e^{-\tilde{h}_t}y_t^2 - 1),$$

$$\tilde{N}_t = -\lambda_t''(\tilde{h}_t) = \frac{1}{2}e^{-\tilde{h}_t}y_t^2.$$

This gives the following updating step:

$$\bar{p}_t = \left(\tilde{p}_t^{-1} + \frac{1}{2}e^{-\tilde{h}_t}y_t^2\right)^{-1},$$

$$\bar{h}_t = \tilde{h}_t + \frac{\bar{p}_t}{2}(e^{-\tilde{h}_t}y_t^2 - 1).$$

6.2 I-scaling for state-quadratic updating

In addition to unusable log-likelihood, there is another problem with this approach, the same one that we have with the basic quasifilter. In general one cannot hope that the Hessian matrix of λ_t is negative definite at $\tilde{\mathbf{a}}_t$. Thus $\tilde{\mathbf{N}}_t$ is not in general positive semidefinite, which can result in scaling matrices $\bar{\mathbf{P}}_t$ and $\tilde{\mathbf{P}}_t$ which are not positive definite.

A crude amendment is to replace $-\nabla^2\lambda_t(\mathbf{a}_t)$ by its expected value, where expectation is with respect to $\mathbf{y}_t$ under the assumption that it is distributed according to $f_\circ(\mathbf{y}_t \mid \mathbf{a}_t, \mathbf{y}_{1:t-1}) = \exp(\lambda_t(\mathbf{a}_t; \mathbf{y}_t))$ for $\mathbf{a}_t = \tilde{\mathbf{a}}_t$:

$$\tilde{\mathbf{I}}_t(\mathbf{a}_t) = -\int \frac{\partial^2\lambda_t(\mathbf{a}_t; \mathbf{y}_t)}{\partial\mathbf{a}_t\partial\mathbf{a}_t^\intercal} \exp(\lambda_t(\mathbf{a}_t; \mathbf{y}_t))d\mathbf{y}_t,$$

where the dependence of λ_t on $\mathbf{y}_t$ has to be shown explicitly. Following analogy with the method of maximum likelihood, this is the "information matrix" corresponding to "the log-likelihood function" $\lambda_t(\mathbf{a}_t)$, which we can call the *measurement information matrix*. According to the information identity the measurement information matrix coincides with the covariance matrix of the measurement score vector $\tilde{\mathbf{s}}_t$ and thus is positive semi-definite:

$$\tilde{\mathbf{I}}_t(\mathbf{a}_t) = \int \frac{\partial\lambda_t(\mathbf{a}_t; \mathbf{y}_t)}{\partial\mathbf{a}_t} \frac{\partial\lambda_t(\mathbf{a}_t; \mathbf{y}_t)}{\partial\mathbf{a}_t^\intercal} \exp(\lambda_t(\mathbf{a}_t; \mathbf{y}_t))d\mathbf{y}_t.$$

The updating step for the covariance matrix is given by

$$\bar{\mathbf{P}}_t = (\tilde{\mathbf{P}}_t^{-1} + \tilde{\mathbf{I}}_t)^{-1}, \qquad \tilde{\mathbf{I}}_t = \tilde{\mathbf{I}}_t(\tilde{\mathbf{a}}_t).$$

For the time-varying scale example if $y_t \sim \mathcal{N}(0, e^{\tilde{h}_t})$, then $\mathsf{E}[\frac{1}{2}e^{-\tilde{h}_t}y_t^2] = \frac{1}{2}$ and the updating step for the state variance simplifies to

$$\bar{p}_{ht} = \left(\tilde{p}_{ht}^{-1} + \frac{1}{2}\right)^{-1} = \frac{2\tilde{p}_{ht}}{2 + \tilde{p}_{ht}}.$$

Table 4: State-quadratic and hybrid updating for time varying scale example

	SV			FTSE-100		
	G	H	I	G	H	I
ω	0.0088 (0.0043)	−0.0190 (0.0033)	0.0045 (0.0026)	−0.0032 (0.0025)	−0.0140 (0.0022)	−0.0031 (0.0018)
δ	0.982 (0.0032)	0.979 (0.0032)	0.982 (0.0032)	0.987 (0.0028)	0.984 (0.0027)	0.986 (0.0030)
σ_h	0.199 (0.0146)	0.142 (0.0104)	0.197 (0.0118)	0.135 (0.0128)	0.100 (0.0089)	0.138 (0.0100)
ν	9.06 (1.0)	8.59 (0.9)	9.75 (1.1)	12.18 (1.4)	13.73 (1.9)	16.99 (3.0)
Max LL	−9805.5	−9817.2	−9803.6	−9298.4	−9281.8	−9307.2
AIC	3.270	3.274	3.269	2.712	2.708	2.715
BIC	3.274	3.278	3.274	2.716	2.712	2.719

Note: Models G and H use state-quadratic updating and H-scaling (I-scaling). Model I uses hybrid updating and H-scaling.
See the notes to Tables 1 and 2 for further explanation.

Table 4 explores the consequences of using different updating approaches for the time varying scale example. (One can additionally consider a model based on hybrid updating and I-scaling, but it almost coincides with model F above.) It can be seen that model fit depends on the the data. It is not clear a priori, which model would be better for a given series.

In general $\tilde{\mathbf{I}}_t$ is not invertible. We can write it as $\tilde{\mathbf{I}}_t = \mathbf{V}_t \mathbf{W}_t \mathbf{V}_t^\intercal$, where $\mathbf{W}_t$ is symmetric positive definite. Then

$$\bar{\mathbf{P}}_t = \tilde{\mathbf{P}}_t - \tilde{\mathbf{P}}_t \mathbf{V}_t (\mathbf{W}_t^{-1} + \mathbf{V}_t^\intercal \tilde{\mathbf{P}}_t \mathbf{V}_t)^{-1} \mathbf{V}_t^\intercal \tilde{\mathbf{P}}_t. \tag{3}$$

It can be seen that the formula is the same as that for the basic quasifilter (2) if we set

$$\mathbf{N}_t = \mathbf{V}_t (\mathbf{W}_t^{-1} + \mathbf{V}_t^\intercal \tilde{\mathbf{P}}_t \mathbf{V}_t)^{-1} \mathbf{V}_t^\intercal.$$

6.3 State-quadratic updating for Gaussian nonlinear measurement

Similar to subsection 4.4 we assume that the measurement distribution is given by

$$\mathbf{y}_t \mid \mathbf{a}_t, \mathbf{y}_{1:t-1} \sim \mathcal{N}(\mathbf{g}_{yt}(\mathbf{a}_t), \Omega_{yt}).$$

Thus the measurement log-density is

$$\lambda_t = \ln \varphi(\mathbf{y}_t - \mathbf{g}_{yt}(\mathbf{a}_t), \Omega_{yt}) = -\frac{1}{2} (\mathbf{y}_t - \mathbf{g}_{yt}(\mathbf{a}_t))^\intercal \Omega_{yt}^{-1} (\mathbf{y}_t - \mathbf{g}_{yt}(\mathbf{a}_t)) + const$$

and the corresponding measurement score at $\tilde{\mathbf{a}}_t$ is

$$\tilde{\mathbf{s}}_t = \nabla \lambda_t(\tilde{\mathbf{a}}_t) = \nabla \mathbf{g}_{yt} \Omega_{yt}^{-1} (\mathbf{y}_t - \mathbf{g}_{yt}),$$

where $\mathbf{g}_{yt} = \mathbf{g}_{yt}(\tilde{\mathbf{a}}_t)$, $\nabla \mathbf{g}_{yt} = \nabla \mathbf{g}_{yt}(\tilde{\mathbf{a}}_t) = \partial \mathbf{g}_{yt}^\intercal(\tilde{\mathbf{a}}_t) / \partial \tilde{\mathbf{a}}_t$. The measurement information matrix can be obtained as the conditional covariance matrix of the score vector:

$$\tilde{\mathbf{I}}_t = \nabla \mathbf{g}_{yt} \Omega_{yt}^{-1} \nabla \mathbf{g}_{yt}^\intercal,$$

Thus, the equations of the updating step are

$$\bar{\mathbf{P}}_t = (\tilde{\mathbf{P}}_t^{-1} + \tilde{\mathbf{I}}_t)^{-1} = (\tilde{\mathbf{P}}_t^{-1} + \nabla \mathbf{g}_{yt} \Omega_{yt}^{-1} \nabla \mathbf{g}_{yt}^\intercal)^{-1},$$
$$\bar{\mathbf{a}}_t = \tilde{\mathbf{a}}_t + \bar{\mathbf{P}}_t \tilde{\mathbf{s}}_t = \tilde{\mathbf{a}}_t + \bar{\mathbf{P}}_t \nabla \mathbf{g}_{yt} \Omega_{yt}^{-1} (\mathbf{y}_t - \mathbf{g}_{yt}).$$

Denoting $\mathbf{V}_t = \nabla \mathbf{g}_{yt}$ and $\mathbf{W}_t = \Omega_{yt}^{-1}$ we obtain from (3) that

$$\bar{\mathbf{P}}_t = \tilde{\mathbf{P}}_t - \tilde{\mathbf{P}}_t \nabla \mathbf{g}_{yt} (\nabla \mathbf{g}_{yt}^\intercal \tilde{\mathbf{P}}_t \nabla \mathbf{g}_{yt} + \Omega_{yt})^{-1} \nabla \mathbf{g}_{yt}^\intercal \tilde{\mathbf{P}}_t.$$

We can compare this result with the basic quasifilter in subsection 4.4. If we denote

$$\Sigma_{yt} = \nabla \mathbf{g}_{yt}^{\intercal} \tilde{\mathbf{P}}_t \nabla \mathbf{g}_{yt} + \Omega_{yt},$$

then

$$\bar{\mathbf{P}}_t = \tilde{\mathbf{P}}_t - \tilde{\mathbf{P}}_t \nabla \mathbf{g}_{yt} \Sigma_{yt}^{-1} \nabla \mathbf{g}_{yt}^{\intercal} \tilde{\mathbf{P}}_t.$$

The formula for $\bar{\mathbf{a}}_t$ can be rewritten in the same manner:

$$\bar{\mathbf{a}}_t = \tilde{\mathbf{a}}_t + \tilde{\mathbf{P}}_t \nabla \mathbf{g}_{yt} \Sigma_{yt}^{-1} (\mathbf{y}_t - \mathbf{g}_{yt}).$$

One can see from this that the state-quadratic approach leads to the extended Kalman filter updating formulas. This is unlike the basic quasifilter approach, which does not lead to the EKF updating formulas directly.

6.4 State-quadratic updating for time-varying regression

For the time-varying regression example of section 5 one can assume that $\epsilon_t = e^{-h_t/2}(y_t - \mathbf{x}_t^{\intercal} \boldsymbol{\beta}_t)$ has the Student's t distribution with κ degrees of freedom standardized to have unit variance, that is

$$\sqrt{\frac{\kappa}{\kappa - 2}}\, \epsilon_t \sim t_\kappa,$$

where t_κ is the ordinary t distribution. The corresponding measurement log-density is given by

$$\lambda_t = \ln \mathrm{tden}\left(\sqrt{\frac{\kappa}{\kappa - 2}}\, \epsilon_t, \kappa\right) - \frac{1}{2} \ln\left(\frac{\kappa}{\kappa - 2}\right) - \frac{h_t}{2}$$

or

$$\lambda_t = \ln \Gamma\left(\frac{\kappa + 1}{2}\right) - \ln \Gamma\left(\frac{\kappa}{2}\right) - \frac{1}{2} \ln(2\pi(\kappa - 2)) - \frac{\kappa + 1}{2} \ln\left(1 + \frac{1}{\kappa - 2} \epsilon_t^2\right) - \frac{h_t}{2},$$

with the measurement score

$$\tilde{\mathbf{s}}_t = \nabla \lambda_t = \begin{pmatrix} (\kappa + 1) \frac{\epsilon_t}{\kappa - 2 + \epsilon_t^2} e^{-h_t/2} \mathbf{x}_t \\ \frac{1}{2}(\kappa + 1) \frac{\epsilon_t^2}{\kappa - 2 + \epsilon_t^2} - \frac{1}{2} \end{pmatrix} = \begin{pmatrix} (\kappa + 1) \frac{\epsilon_t}{\kappa - 2 + \epsilon_t^2} e^{-h_t/2} \mathbf{x}_t \\ \frac{1}{2} \frac{\kappa \epsilon_t^2 - \kappa + 2}{\kappa - 2 + \epsilon_t^2} \end{pmatrix}$$

and the measurement information matrix

$$\tilde{\mathbf{I}}_t = \begin{pmatrix} \frac{\kappa(\kappa+1)}{(\kappa-2)(\kappa+3)} e^{-\tilde{h}_t} \mathbf{x}_t \mathbf{x}_t^{\intercal} & \mathbf{0} \\ \mathbf{0}^{\intercal} & \frac{1}{2} \frac{\kappa}{\kappa+3} \end{pmatrix}.$$

This produces

$$\mathbf{V}_t = \begin{pmatrix} \mathbf{x}_t & \mathbf{0} \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{W}_t = \begin{pmatrix} \frac{\kappa(\kappa+1)}{(\kappa-2)(\kappa+3)} e^{-\tilde{h}_t} & 0 \\ 0 & \frac{1}{2} \frac{\kappa}{\kappa+3} \end{pmatrix},$$

suitable for (3). Restricting scaling matrices to be block-diagonal, that is,

$$\tilde{\mathbf{P}}_t = \mathrm{diag}(\tilde{\mathbf{P}}_{\boldsymbol{\beta} t}, \tilde{p}_{ht})$$

we obtain

$$\mathbf{W}_t^{-1} + \mathbf{V}_t^{\intercal} \tilde{\mathbf{P}}_t \mathbf{V}_t = \begin{pmatrix} \frac{(\kappa-2)(\kappa+3)}{\kappa(\kappa+1)} e^{\tilde{h}_t} + \mathbf{x}_t^{\intercal} \tilde{\mathbf{P}}_{\boldsymbol{\beta} t} \mathbf{x}_t & 0 \\ 0 & 2\frac{\kappa+3}{\kappa} + \tilde{p}_{ht} \end{pmatrix},$$

$$\mathbf{N}_t = \mathbf{V}_t (\mathbf{W}_t^{-1} + \mathbf{V}_t^{\intercal} \tilde{\mathbf{P}}_t \mathbf{V}_t)^{-1} \mathbf{V}_t^{\intercal} = \begin{pmatrix} \left(\frac{(\kappa-2)(\kappa+3)}{\kappa(\kappa+1)} e^{\tilde{h}_t} + \mathbf{x}_t^{\intercal} \tilde{\mathbf{P}}_{\boldsymbol{\beta} t} \mathbf{x}_t\right)^{-1} \mathbf{x}_t \mathbf{x}_t^{\intercal} & \mathbf{0} \\ \mathbf{0}^{\intercal} & \left(2\frac{\kappa+3}{\kappa} + \tilde{p}_{ht}\right)^{-1} \end{pmatrix}.$$

This gives the following quasifilter updating formulas:

$$\tilde{c}_t = \sqrt{\frac{(\kappa-2)(\kappa+3)}{\kappa(\kappa+1)} e^{\tilde{h}_t} + \mathbf{x}_t^\intercal \tilde{\mathbf{P}}_{\beta t} \mathbf{x}_t},$$

$$\bar{\mathbf{P}}_{\beta t} = \tilde{\mathbf{P}}_{\beta t} - \tilde{\mathbf{P}}_{\beta t} \mathbf{x}_t \tilde{c}_t^{-2} \mathbf{x}_t^\intercal \tilde{\mathbf{P}}_{\beta t},$$

$$\bar{p}_{ht} = \tilde{p}_{ht} - \left(2\frac{\kappa+3}{\kappa} + \tilde{p}_{ht}\right)^{-1} \tilde{p}_{ht}^2 = \tilde{p}_{ht} \left(1 + \frac{\kappa}{\kappa+3}\frac{\tilde{p}_{ht}}{2}\right)^{-1},$$

$$\tilde{\epsilon}_t = e^{-\tilde{h}_t/2}(y_t - \mathbf{x}_t^\intercal \tilde{\boldsymbol{\beta}}_t),$$

$$\bar{\boldsymbol{\beta}}_t = \tilde{\boldsymbol{\beta}}_t + (\kappa+1)\frac{\tilde{\epsilon}_t}{\kappa - 2 + \tilde{\epsilon}_t^2} e^{-\tilde{h}_t/2} \bar{\mathbf{P}}_{\beta t} \mathbf{x}_t,$$

which can be rewritten as

$$\bar{\boldsymbol{\beta}}_t = \tilde{\boldsymbol{\beta}}_t + \frac{(\kappa-2)(\kappa+3)}{\kappa} \frac{\tilde{\epsilon}_t}{\kappa - 2 + \tilde{\epsilon}_t^2} \frac{e^{\tilde{h}_t/2}}{\tilde{c}_t^2} \tilde{\mathbf{P}}_{\beta t} \mathbf{x}_t,$$

and

$$\bar{h}_t = \tilde{h}_t + \frac{\bar{p}_{ht}}{2} \frac{\kappa\tilde{\epsilon}_t^2 - \kappa + 2}{\kappa - 2 + \tilde{\epsilon}_t^2}.$$

Further, we can fix $\bar{p}_{ht} = 2\bar{\rho}$, which gives

$$\bar{h}_t = \tilde{h}_t + \bar{\rho} \frac{\kappa\tilde{\epsilon}_t^2 - \kappa + 2}{\kappa - 2 + \tilde{\epsilon}_t^2}.$$

State-quadratic approach does not provide a valid likelihood function. Thus, these formulas should be complemented by a formulation of the likelihood function. Here we can utilize the likelihood function from the basic quasifilter of section 5. The log-likelihood is

$$\ell_t = \ln \mathrm{tden}(T_t, \nu) - \ln c_t - \ln A,$$

where

$$c_t = \sqrt{e^{\tilde{h}_t + \tilde{p}_{ht}/2} + \mathbf{x}_t^\intercal \tilde{\mathbf{P}}_{\beta t} \mathbf{x}_t}$$

and

$$T_t = \frac{y_t - \mathbf{x}_t^\intercal \tilde{\boldsymbol{\beta}}_t}{A c_t}.$$

In section 5 the coefficient A was chosen in a way which preserves positive definiteness of the scaling matrices. Here we can instead set $A = \sqrt{\frac{\nu-2}{\nu}}$, to equate variances and provide a closer correspondence between ℓ_t and $\ell_{\sharp t}$.

Note that state-quadratic quasifilter introduces an additional parameter, κ, compared to the basic quasifilter. This makes the models obtained by state-quadratic approach less parsimonious in terms of the number of parameters. This drawback can be potentially offset by a better model fit. There is also a possibility to fix κ. A natural choice is $\kappa = +\infty$, which corresponds to the Gaussian underlying regression errors.

7 Example: A seasonal time-varying autoregression

Consider a first-order autoregressive model with deterministic seasonality:

$$y_t = \mu + \mathbf{d}_t^\intercal \boldsymbol{\gamma} + \varphi y_{t-1} + \sigma\epsilon_t, \quad \operatorname{var}(\epsilon_t) = 1. \tag{4}$$

If M is the number of seasons, then $\boldsymbol{\gamma}$ is a vector of seasonal coefficients, which is constrained to have zero mean, and $\mathbf{d}_t$ is a seasonal dummy vector, which is zero except for a unit corresponding to the current season. A time-varying modification of this model assumes that all parameters follow Gaussian random walks (with a special covariance matrix for seasonality).

$$y_t = \mu_t + \mathbf{d}_t^\intercal \boldsymbol{\gamma}_t + \varphi_t y_{t-1} + e^{h_t/2}\epsilon_t, \quad \operatorname{var}(\epsilon_t) = 1. \tag{5}$$

Time-varying parameter μ_t can be interpreted as the level correction for y_t, φ_t as the persistence of y_t, h_t as the short-run volatility, while $\boldsymbol{\gamma}_t$ are the seasonal factors, one for each season. Note that unlike the models of section 4.6 the seasonal factors are not circularly permuted. A specific feature of this extended model is that the persistence of the process is subject to variation, which introduces long-memory effects.

It can be seen that the extended model is a time-varying regression model with

$$\mathbf{x}_t = \begin{pmatrix} 1 \\ \mathbf{d}_t \\ y_{t-1} \end{pmatrix}, \quad \boldsymbol{\beta}_t = \begin{pmatrix} \mu_t \\ \boldsymbol{\gamma}_t \\ \varphi_t \end{pmatrix}.$$

We assume that the covariance matrix of the disturbances corresponding to the transition equation for the regression coefficients is block-diagonal:

$$\boldsymbol{\Omega}_{\boldsymbol{\beta}t} = \boldsymbol{\Omega}_{\boldsymbol{\beta}} = \operatorname{diag}\left(\sigma_\mu^2, \sigma_\gamma^2(\mathbf{I}_M - \frac{1}{M}\mathbf{1}_M\mathbf{1}_M^\intercal), \sigma_\varphi^2\right).$$

The block for seasonality ensures that seasonality remains centered. The transition distribution for the volatility variable has constant variance $\omega_{ht}^2 = \sigma_h^2$. The coefficients of the transition equation are

$$\mathbf{R}_{at} = \mathbf{0}, \qquad \mathbf{R}_{aat} = \mathbf{I}.$$

Other specifications follow section 5 and subsection 6.4. .

The empirical example is based on the U.S. monthly CPI inflation series for the period from 1913-01 to 2014-11. The series is rather long. It covers periods with very different macroeconomic conditions and the initial fixed coefficients model (4) demonstrates poor fit to the data. The residuals are characterized by large autocorrelation and changing volatility. Experiments with rolling estimation show that the estimated parameters vary widely. This can be explained by time variation of the coefficients, which suggests using the time-varying model (5). A similar time-varying AR model of inflation was suggested in Evans (1991).[3] A more general model (time-varying VAR) was used in Cogley and Sargent (2005) to describe the joint dynamics of inflation, unemployment, and interest rates.

Here we replace the initial unobserved components model by the quasifilters considered in section 5 and subsection 6.4. The quasifilter approach greatly simplifies computation of the likelihood function compared to the parent nonlinear non-Gaussian state-space model. Table 5 shows the results. Although the state-quadratic approach produces a model with an additional parameter, it outperforms the model obtained from the basic quasifilter in terms of information criteria.

[3] In Evans (1991) Kalman filter innovations were used inside ARCH.

Table 5: Seasonal time-varying autoregression for the U.S. inflation.

	basic	state-quad.	hybrid
σ_μ	0.015 (0.0059)	0.016 (0.0054)	0.018 (0.0067)
σ_γ	0.0073 (0.0011)	0.0077 (0.0012)	0.0087 (0.0019)
σ_φ	0.041 (0.0082)	0.043 (0.0088)	0.050 (0.0140)
σ_h	0.158 (0.0282)	0.131 (0.0265)	0.153 (0.0288)
ν	6.66 (1.0)	4.06 (0.3)	6.60 (1.0)
κ	—	18.62 (7.9)	3.29 (1.4)
Max LL	−542.1	−537.0	−542.5
AIC	0.895	0.889	0.898
BIC	0.916	0.914	0.923

Note: The data is Consumer Price Index for All Urban Consumers, all items, not seasonally adjusted for the period from 1913-01 to 2014-11 (1222 monthly observations). The inflation series was obtained as $y_t = \Delta \ln(\mathrm{CPI}_t) \cdot 100$. Standard errors in brackets

8 Mildly nonlinear and/or non-Gaussian transition

8.1 The general case

We impose a requirement that $f_\circ(\mathbf{a}_t \mid \mathbf{a}_{t-1}, \mathbf{y}_{1:t-1})$ is nonlinear and/or non-Gaussian so that

$$f_\circ(\mathbf{a}_t \mid \mathbf{a}_{t-1}, \mathbf{y}_{1:t-1}) q(\mathbf{a}_{t-1})$$

is a unimodal (strictly quasiconcave) density of $\mathbf{a}_t$ and $\mathbf{a}_{t-1}$ for any multivariate Gaussian density $q(\mathbf{a}_{t-1})$. Moreover, we assume that a closed-form formula is available for the mode.

Suppose that we have $f(\mathbf{a}_{t-1} \mid \mathbf{y}_{1:t-1})$ which is an approximation of $f_\circ(\mathbf{a}_{t-1} \mid \mathbf{y}_{1:t-1})$ from the previous period in the form of a Gaussian density corresponding to $\mathcal{N}(\bar{\mathbf{a}}_{t-1}, \bar{\mathbf{P}}_{t-1})$. We want to find an approximation to $f_\circ(\mathbf{a}_t \mid \mathbf{y}_{1:t-1})$ as a density corresponding to $\mathcal{N}(\tilde{\mathbf{a}}_t, \tilde{\mathbf{P}}_t)$.

Assume that $(\tilde{\mathbf{a}}_t, \breve{\mathbf{a}}_{t-1})$ is the mode of $f_\circ(\mathbf{a}_t \mid \mathbf{a}_{t-1}, \mathbf{y}_{1:t-1}) \varphi(\mathbf{a}_{t-1} - \bar{\mathbf{a}}_{t-1}, \bar{\mathbf{P}}_{t-1})$. For a typical transition density (for example, $\mathbf{a}_t \mid \mathbf{a}_{t-1}, \mathbf{y}_{1:t-1} \sim \mathcal{N}(\mathbf{g}_{at}(\mathbf{a}_{t-1}), \Omega_{at})$) we have that $\max_{\mathbf{a}_t} f_\circ(\mathbf{a}_t \mid \mathbf{a}_{t-1}, \mathbf{y}_{1:t-1})$ does not depend on $\mathbf{a}_{t-1}$ and thus $\breve{\mathbf{a}}_{t-1}$ maximizes $f(\mathbf{a}_{t-1} \mid \mathbf{y}_{1:t-1})$, which gives just $\breve{\mathbf{a}}_{t-1} = \bar{\mathbf{a}}_{t-1}$.

Denote

$$\begin{aligned}\gamma_{at}(\mathbf{a}_t, \mathbf{a}_{t-1}) &= \ln f_\circ(\mathbf{a}_t \mid \mathbf{a}_{t-1}, \mathbf{y}_{1:t-1}) + \ln \varphi(\mathbf{a}_{t-1} - \bar{\mathbf{a}}_{t-1}, \bar{\mathbf{P}}_{t-1}) \\ &= \ln f_\circ(\mathbf{a}_t \mid \mathbf{a}_{t-1}, \mathbf{y}_{1:t-1}) - \frac{m_{t-1}}{2} \ln(2\pi) - \frac{1}{2} \ln |\bar{\mathbf{P}}_{t-1}| - \frac{1}{2} (\mathbf{a}_{t-1} - \bar{\mathbf{a}}_{t-1})^\intercal \bar{\mathbf{P}}_{t-1}^{-1} (\mathbf{a}_{t-1} - \bar{\mathbf{a}}_{t-1})\end{aligned}$$

Then $(\tilde{\mathbf{a}}_t, \breve{\mathbf{a}}_{t-1})$ maximizes $\gamma_{at}(\mathbf{a}_t, \mathbf{a}_{t-1})$. In the spirit of the Laplace's method of integration γ_{at} can be approximated up to a constant term by a log-density of the Gaussian distribution with the mean $(\tilde{\mathbf{a}}_t, \breve{\mathbf{a}}_{t-1})$ and the covariance matrix equal to the negated inverted Hessian matrix $\nabla^2 \gamma_{at}(\tilde{\mathbf{a}}_t, \breve{\mathbf{a}}_{t-1})$. Denote

$$\mathbf{M}_{ijt} = -\frac{d^2 \ln f_\circ(\mathbf{a}_t \mid \mathbf{a}_{t-1}, \mathbf{y}_{1:t-1})}{d\mathbf{a}_{t-i} d\mathbf{a}_{t-j}^\intercal}\bigg|_{\mathbf{a}_t = \tilde{\mathbf{a}}_t, \mathbf{a}_{t-1} = \breve{\mathbf{a}}_{t-1}}.$$

Then

$$-\nabla^2 \gamma_{at}(\tilde{\mathbf{a}}_t, \breve{\mathbf{a}}_{t-1}) = \begin{pmatrix} \mathbf{M}_{00t} & \mathbf{M}_{01t} \\ \mathbf{M}_{01t}^\intercal & \mathbf{M}_{11t} + \bar{\mathbf{P}}_{t-1}^{-1} \end{pmatrix}.$$

The upper left block of its inverse is given by

$$\tilde{\mathbf{P}}_t = (\mathbf{M}_{00t} - \mathbf{M}_{01t} (\mathbf{M}_{11t} + \bar{\mathbf{P}}_{t-1}^{-1})^{-1} \mathbf{M}_{01t}^\intercal)^{-1},$$

which provides the covariance matrix for $f(\mathbf{a}_t \mid \mathbf{y}_{1:t-1}) = \varphi(\mathbf{a}_t - \tilde{\mathbf{a}}_t, \tilde{\mathbf{P}}_t)$.

8.2 Prediction step for the case of a Gaussian transition equation

Suppose that

$$\mathbf{a}_t \mid \mathbf{a}_{t-1}, \mathbf{y}_{1:t-1} \sim \mathcal{N}(\mathbf{g}_{at}(\mathbf{a}_{t-1}), \Omega_{at}).$$

Then

$$\ln f_\circ(\mathbf{a}_t \mid \mathbf{a}_{t-1}, \mathbf{y}_{1:t-1}) = -\frac{1}{2}(\mathbf{a}_t - \mathbf{g}_{at}(\mathbf{a}_{t-1}))^\top \Omega_{at}^{-1}(\mathbf{a}_t - \mathbf{g}_{at}(\mathbf{a}_{t-1})) + \text{const}.$$

For this transition density we have $\tilde{\mathbf{a}}_t = \mathbf{g}_{at}(\bar{\mathbf{a}}_{t-1})$.

For $\nabla\tilde{\mathbf{g}}_{at} = \nabla\mathbf{g}_{at}(\tilde{\mathbf{a}}_{t-1})$, $\breve{\mathbf{a}}_{t-1} = \bar{\mathbf{a}}_{t-1}$

$$\mathbf{M}_{00t} = -\frac{d^2 \ln f_\circ(\mathbf{a}_t \mid \mathbf{a}_{t-1}, \mathbf{y}_{1:t-1})}{d\mathbf{a}_t d\mathbf{a}_t^\top}\bigg|_{\mathbf{a}_t=\tilde{\mathbf{a}}_t, \mathbf{a}_{t-1}=\breve{\mathbf{a}}_{t-1}} = \Omega_{at}^{-1},$$

$$\mathbf{M}_{01t} = -\frac{d^2 \ln f_\circ(\mathbf{a}_t \mid \mathbf{a}_{t-1}, \mathbf{y}_{1:t-1})}{d\mathbf{a}_t d\mathbf{a}_{t-1}^\top}\bigg|_{\mathbf{a}_t=\tilde{\mathbf{a}}_t, \mathbf{a}_{t-1}=\breve{\mathbf{a}}_{t-1}} = -\Omega_{at}^{-1}\nabla\tilde{\mathbf{g}}_{at},$$

$$\mathbf{M}_{11t} = -\frac{d^2 \ln f_\circ(\mathbf{a}_t \mid \mathbf{a}_{t-1}, \mathbf{y}_{1:t-1})}{d\mathbf{a}_{t-1} d\mathbf{a}_{t-1}^\top}\bigg|_{\mathbf{a}_t=\tilde{\mathbf{a}}_t, \mathbf{a}_{t-1}=\breve{\mathbf{a}}_{t-1}} = \nabla\tilde{\mathbf{g}}_{at}^\top \Omega_{at}^{-1}\nabla\tilde{\mathbf{g}}_{at}.$$

This produces

$$\tilde{\mathbf{P}}_t = (\Omega_{at}^{-1} - \Omega_{at}^{-1}\nabla\tilde{\mathbf{g}}_{at}(\nabla\tilde{\mathbf{g}}_{at}^\top \Omega_{at}^{-1}\nabla\tilde{\mathbf{g}}_{at} + \bar{\mathbf{P}}_{t-1}^{-1})^{-1}\nabla\tilde{\mathbf{g}}_{at}^\top \Omega_{at}^{-1})^{-1}$$

or

$$\tilde{\mathbf{P}}_t = \nabla\tilde{\mathbf{g}}_{at}\bar{\mathbf{P}}_{t-1}\nabla\tilde{\mathbf{g}}_{at}^\top + \Omega_{at},$$

which together with $\tilde{\mathbf{a}}_t = \mathbf{g}_{at}(\bar{\mathbf{a}}_{t-1})$ comprise the well-known prediction step of the extended Kalman filter.

9 Conclusions and discussion

Quasifilter can be parsimonious in terms of the number of parameters, but the recursions for the underlying components are usually not so simple. Various devices can be employed to make the dynamics of the underlying components simpler. If this leads to a noticeable deterioration of model fit, then one has to make a choice between simplicity and empirical performance.

In general the estimates of the parameters of the derived model would be inconsistent for the parameters of processes described by the parent model. However, there is no problem in this observation. After derivation of a quasifilter one can forget about the parent model and go on with the result. However, even though there is no necessity in this, potentially it could be interesting and illuminating to demonstrate that a quasifilter model is a close approximation to the parent unobserved component model.

There is also a related aspect, that a quasifilter model it is not meant to produce estimates of latent processes. Exact or approximate filters and smoothers corresponding to unobserved component models do produce such estimates. For example, in technical applications filtering can be used to predict a position of some moving object given some noisy and/or indirect information on this position. In economic applications smoothing can be used to extract some imaginary component of a time series such as seasonality or trend. By definition quasifilter is not an approximate filter, but a stand-alone model. Thus, its observable underlying variables are part of model description rather than estimates of something latent.

The quasifilter techniques are of heuristic nature and are not based on a comprehensive theory. Nevertheless, they are useful for formulation of new kinds of dynamic models. The historical development of some well-known time series models demonstrate this. In particular, the history of ARCH model and its numerous extensions suggests that adoption of a new model can be stimulated by good empirical properties, while firm theoretical foundations can be provided by further research. This second stage of model exploration can include proving consistency and efficiency of the MLE estimates, etc., but the most important substantiation of a model is provided by its good empirical performance.

References

Archibald, B. C. and Koehler, A. B. (2003). Normalization of seasonal factors in Winters' methods, *International Journal of Forecasting* **19**(1): 143–148.

Cogley, T. and Sargent, T. J. (2005). Drifts and volatilities: Monetary policies and outcomes in the post WWII US, *Review of Economic Dynamics* **8**(2): 262–302.

Cox, D. R. (1981). Statistical analysis of time series: Some recent developments, *Scandinavian Journal of Statistics* **8**(2): 93–115.

Creal, D. (2012). A survey of sequential monte carlo methods for economics and finance, *Econometric Reviews* **31**(3): 245–296.

Creal, D. D., Koopman, S. J. and Lucas, A. (2013). Generalized autoregressive score models with applications, *Journal of Applied Econometrics* **28**(5): 777–795.

Creal, D., Koopman, S. J. and Lucas, A. (2008). A general framework for observation driven time-varying parameter models, *Discussion Paper TI 2008-108/4*, Tinbergen Institute.

Evans, M. (1991). Discovering the link between inflation rates and inflation uncertainty, *Journal of Money, Credit and Banking* **23**(2): 169–184.

Harvey, A. (2006). Forecasting with unobserved components time series models, *in* G. Elliott, C. Granger and A. Timmermann (eds), *Handbook of Economic Forecasting*, Vol. 1, Elsevier, chapter 7, pp. 327–412.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press.

Harvey, A. C. (2013). *Dynamic Models for Volatility and Heavy Tails: with Applications to Financial and Economic Time Series*, Econometric society monographs, Cambridge University Press.

Harvey, A. and Chakravarty, T. (2008). Beta-t-(e)garch, *Cambridge working papers in economics*, University of Cambridge, Faculty of Economics.

Hyndman, R. J., Koehler, A. B., Ord, J. K. and Snyder, R. D. (2008). *Forecasting with Exponential Smoothing: The State Space Approach*, Springer.

Kitagawa, G. (1987). Non-Gaussian State-Space modeling of nonstationary time series, *Journal of the American Statistical Association* **82**(400): 1032–1041.

Masreliez, C. (1975). Approximate non-Gaussian filtering with linear state and observation relations, *IEEE Transactions on Automatic Control* **20**(1): 107–110.

Simon, D. (2006). *Optimal state estimation: Kalman, H∞ and nonlinear approaches*, Wiley-Interscience.

Appendix

Proof of Proposition 1. Under regularity conditions allowing interchanging the order of integration and differentiation the gradient of $\ell_{\sharp t} = \ln f_\sharp(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$ (the score vector) at $\tilde{\mathbf{a}}_t$ is given by

$$\begin{aligned}\mathbf{s}_{\sharp t} = \nabla \ell_{\sharp t}(\tilde{\mathbf{a}}_t) = \frac{\partial \ell_{\sharp t}(\tilde{\mathbf{a}}_t)}{\partial \tilde{\mathbf{a}}_t} &= \frac{1}{f_\sharp(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})} \frac{d f_\sharp(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})}{d \tilde{\mathbf{a}}_t} \\ &= \frac{1}{f_\sharp(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})} \int f_\circ(\mathbf{y}_t \mid \mathbf{a}_t, \mathbf{y}_{1:t-1}) \frac{d\varphi(\mathbf{a}_t - \tilde{\mathbf{a}}_t, \tilde{\mathbf{P}}_t)}{d \tilde{\mathbf{a}}_t} d\mathbf{a}_t.\end{aligned}$$

The derivative of the Gaussian density can be written as

$$\frac{d\varphi(\mathbf{a}_t - \tilde{\mathbf{a}}_t, \tilde{\mathbf{P}}_t)}{d \tilde{\mathbf{a}}_t} = \varphi(\mathbf{a}_t - \tilde{\mathbf{a}}_t, \tilde{\mathbf{P}}_t) \tilde{\mathbf{P}}_t^{-1} (\mathbf{a}_t - \tilde{\mathbf{a}}_t).$$

Consequently,

$$\begin{aligned}\mathbf{s}_{\sharp t} &= \frac{1}{f_\sharp(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})} \tilde{\mathbf{P}}_t^{-1} \int f_\circ(\mathbf{y}_t \mid \mathbf{a}_t, \mathbf{y}_{1:t-1}) \varphi(\mathbf{a}_t - \tilde{\mathbf{a}}_t, \tilde{\mathbf{P}}_t)(\mathbf{a}_t - \tilde{\mathbf{a}}_t) d\mathbf{a}_t \\ &= \tilde{\mathbf{P}}_t^{-1} \int f_\sharp(\mathbf{a}_t \mid \mathbf{y}_{1:t})(\mathbf{a}_t - \tilde{\mathbf{a}}_t) d\mathbf{a}_t = \tilde{\mathbf{P}}_t^{-1} \Big(\int f_\sharp(\mathbf{a}_t \mid \mathbf{y}_{1:t}) \mathbf{a}_t d\mathbf{a}_t - \tilde{\mathbf{a}}_t \Big)\end{aligned}$$

or

$$\mathbf{s}_{\sharp t} = \nabla \ell_{\sharp t}(\tilde{\mathbf{a}}_t) = \tilde{\mathbf{P}}_t^{-1} (\mathsf{E}_{\sharp t}\, \mathbf{a}_t - \tilde{\mathbf{a}}_t),$$

which gives

$$\mathsf{E}_{\sharp t}\, \mathbf{a}_t - \tilde{\mathbf{a}}_t = \tilde{\mathbf{P}}_t \mathbf{s}_{\sharp t}.$$

Similarly consider the negated Hessian matrix of $\ell_{\sharp t}$ at $\tilde{\mathbf{a}}_t$

$$\begin{aligned}\mathbf{N}_{\sharp t} = -\nabla^2 \ell_{\sharp t}(\tilde{\mathbf{a}}_t) = -\frac{\partial^2 \ell_{\sharp t}(\tilde{\mathbf{a}}_t)}{\partial \tilde{\mathbf{a}}_t \partial \tilde{\mathbf{a}}_t^\intercal} &= -\frac{d}{d \tilde{\mathbf{a}}_t^\intercal} \left(\frac{1}{f_\sharp(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})} \frac{d f_\sharp(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})}{d \tilde{\mathbf{a}}_t} \right) \\ &= \frac{1}{f_\sharp(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})^2} \frac{d f_\sharp(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})}{d \tilde{\mathbf{a}}_t} \frac{d f_\sharp(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})}{d \tilde{\mathbf{a}}_t^\intercal} - \frac{1}{f_\sharp(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})} \frac{d^2 f_\sharp(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})}{d \tilde{\mathbf{a}}_t d \tilde{\mathbf{a}}_t^\intercal} \\ &= \mathbf{s}_{\sharp t} \mathbf{s}_{\sharp t}^\intercal - \frac{1}{f_\sharp(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})} \int f_\circ(\mathbf{y}_t \mid \mathbf{a}_t, \mathbf{y}_{1:t-1}) \frac{d^2 \varphi(\mathbf{a}_t - \tilde{\mathbf{a}}_t, \tilde{\mathbf{P}}_t)}{d \tilde{\mathbf{a}}_t d \tilde{\mathbf{a}}_t^\intercal} d\mathbf{a}_t.\end{aligned}$$

Here

$$\frac{d^2 \varphi(\mathbf{a}_t - \tilde{\mathbf{a}}_t, \tilde{\mathbf{P}}_t)}{d \tilde{\mathbf{a}}_t d \tilde{\mathbf{a}}_t^\intercal} = -\varphi(\mathbf{a}_t - \tilde{\mathbf{a}}_t, \tilde{\mathbf{P}}_t) \tilde{\mathbf{P}}_t^{-1} + \varphi(\mathbf{a}_t - \tilde{\mathbf{a}}_t, \tilde{\mathbf{P}}_t) \tilde{\mathbf{P}}_t^{-1} (\mathbf{a}_t - \tilde{\mathbf{a}}_t)(\mathbf{a}_t - \tilde{\mathbf{a}}_t)^\intercal \tilde{\mathbf{P}}_t^{-1}.$$

Thus,

$$\begin{aligned}\mathbf{N}_{\sharp t} &= \mathbf{s}_{\sharp t} \mathbf{s}_{\sharp t}^\intercal - \frac{1}{f_\sharp(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})} \int f_\circ(\mathbf{y}_t \mid \mathbf{a}_t, \mathbf{y}_{1:t-1}) \frac{d^2 \varphi(\mathbf{a}_t - \tilde{\mathbf{a}}_t, \tilde{\mathbf{P}}_t)}{d \tilde{\mathbf{a}}_t d \tilde{\mathbf{a}}_t^\intercal}\, d\mathbf{a}_t \\ &= \mathbf{s}_{\sharp t} \mathbf{s}_{\sharp t}^\intercal - \int [-\tilde{\mathbf{P}}_t^{-1} + \tilde{\mathbf{P}}_t^{-1} (\mathbf{a}_t - \tilde{\mathbf{a}}_t)(\mathbf{a}_t - \tilde{\mathbf{a}}_t)^\intercal \tilde{\mathbf{P}}_t^{-1}] f_\sharp(\mathbf{a}_t \mid \mathbf{y}_{1:t}) d\mathbf{a}_t \\ &= \mathbf{s}_{\sharp t} \mathbf{s}_{\sharp t}^\intercal + \tilde{\mathbf{P}}_t^{-1} - \tilde{\mathbf{P}}_t^{-1}\, \mathsf{E}_{\sharp t}[(\mathbf{a}_t - \tilde{\mathbf{a}}_t)(\mathbf{a}_t - \tilde{\mathbf{a}}_t)^\intercal] \tilde{\mathbf{P}}_t^{-1}.\end{aligned}$$

Note that

$$\begin{aligned}
\mathsf{var}_{\sharp t}\,\mathbf{a}_t &= \mathsf{E}_{\sharp t}[(\mathbf{a}_t - \mathsf{E}_{\sharp t}\,\mathbf{a}_t)(\mathbf{a}_t - \mathsf{E}_{\sharp t}\,\mathbf{a}_t)^\intercal] \\
&= \mathsf{E}_{\sharp t}[(\mathbf{a}_t - \tilde{\mathbf{a}}_t)(\mathbf{a}_t - \tilde{\mathbf{a}}_t)^\intercal] - 2\,\mathsf{E}_{\sharp t}[(\mathbf{a}_t - \tilde{\mathbf{a}}_t)(\mathsf{E}_{\sharp t}\,\mathbf{a}_t - \tilde{\mathbf{a}}_t)^\intercal] + \mathsf{E}_{\sharp t}[(\mathsf{E}_{\sharp t}\,\mathbf{a}_t - \tilde{\mathbf{a}}_t)(\mathsf{E}_{\sharp t}\,\mathbf{a}_t - \tilde{\mathbf{a}}_t)^\intercal] \\
&= \mathsf{E}_{\sharp t}[(\mathbf{a}_t - \tilde{\mathbf{a}}_t)(\mathbf{a}_t - \tilde{\mathbf{a}}_t)^\intercal] - (\mathsf{E}_{\sharp t}\,\mathbf{a}_t - \tilde{\mathbf{a}}_t)(\mathsf{E}_{\sharp t}\,\mathbf{a}_t - \tilde{\mathbf{a}}_t)^\intercal \\
&= \mathsf{E}_{\sharp t}[(\mathbf{a}_t - \tilde{\mathbf{a}}_t)(\mathbf{a}_t - \tilde{\mathbf{a}}_t)^\intercal] - \tilde{\mathbf{P}}_t \mathbf{s}_{\sharp t} \mathbf{s}_{\sharp t}^\intercal \tilde{\mathbf{P}}_t,
\end{aligned}$$

and thus

$$\mathbf{N}_{\sharp t} = \mathbf{s}_{\sharp t}\mathbf{s}_{\sharp t}^\intercal + \tilde{\mathbf{P}}_t^{-1} - \tilde{\mathbf{P}}_t^{-1}\left(\mathsf{var}_{\sharp t}\,\mathbf{a}_t + \tilde{\mathbf{P}}_t \mathbf{s}_{\sharp t}\mathbf{s}_{\sharp t}^\intercal \tilde{\mathbf{P}}_t\right)\tilde{\mathbf{P}}_t^{-1}$$

or

$$\mathbf{N}_{\sharp t} = \tilde{\mathbf{P}}_t^{-1} - \tilde{\mathbf{P}}_t^{-1}\,\mathsf{var}_{\sharp t}\,\mathbf{a}_t \tilde{\mathbf{P}}_t^{-1}.$$

□