TI 2023-016/III
Tinbergen Institute Discussion Paper

Extremum Monte Carlo Filters: Real-Time Signal Extraction via Simulation and Regression

Revision: December 2023

Francisco Blasques [1]
Siem Jan Koopman[2]
Karim Moussa[3]

1 Vrije Universiteit Amsterdam and Tinbergen Institute

2 Vrije Universiteit Amsterdam and Tinbergen Institute

3 Vrije Universiteit Amsterdam

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and Vrije Universiteit Amsterdam.

Contact: discussionpapers@tinbergen.nl

More TI discussion papers can be downloaded at https://www.tinbergen.nl

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 598 4580

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900

Extremum Monte Carlo Filters:

Real-Time Signal Extraction via Simulation and Regression

Francisco Blasques, Siem Jan Koopman[*], Karim Moussa

Vrije Universiteit Amsterdam and Tinbergen Institute, the Netherlands

First version: November 18, 2021
This version: December 20, 2023

Abstract

This paper introduces a novel simulation-based filtering method for general state space models. It can be used to compute time-varying conditional means, modes, and quantiles, and for predicting latent variables. The method consists of generating artificial data sets from the model and estimating quantities of interest via extremum estimation. We call this procedure *extremum Monte Carlo*. The approach is conceptually simple and easy to implement. It can be applied to any model from which samples of data can be simulated. Given that most of the computations can be performed in advance, the method is particularly suited for real-time applications. The filter is stable over time under mild assumptions, which remains valid under model misspecification. Conditions are provided for convergence to an optimal filter as the number of draws diverges. The linear version of the filter converges to the Kalman filter. Various other features of the filter are illustrated via examples related to nonlinearity, missing data, and intractable densities. An empirical application to exchange rates demonstrates that, despite a setting of limited tractability, the method is able to efficiently extract the time-varying volatility.

Keywords: Intractable densities, Least squares Monte Carlo, Nonlinear non-Gaussian state space models, Hidden Markov models, Real-time filtering.

[*]Corresponding author. E-mail: s.j.koopman@vu.nl. The authors are grateful for the comments from participants of the 2023 NESG seminar in Rotterdam. Blasques thanks the Dutch Research Council (VI.Vidi.195.099) for financial support. Koopman acknowledges support from Aarhus University, Denmark, and funding of the Danish National Research Foundation (DNRF78). Work performed in partial fulfilment of the third author's PhD requirements at the Vrije Universiteit Amsterdam.

1 Introduction

State space models (SSMs) decompose observed time series into two unobserved parts: the states (or signal) which are the true objects of interest, and the noise which complicates the extraction of the signal from the data. The state space modeling approach has become pervasive in both the scientific and industry domains, with applications in fields varying from financial econometrics and forecasting to robotics. Let $x_t \in \mathbb{R}^{N_x}$ denote the state vector at time t and let $y_t \in \mathbb{R}^{N_y}$ be the corresponding vector of measurements (i.e., the observed variables) for some $N_x, N_y \in \mathbb{N}$, with the related noise vectors denoted by ε_t^x and ε_t^y. We can then represent the SSM by

$$\begin{aligned} y_t &= m_t(x_t, \varepsilon_t^y), & (\varepsilon_t^x, \varepsilon_t^y) &\sim p(\varepsilon_t^x, \varepsilon_t^y), \\ x_{t+1} &= s_t(x_t, \varepsilon_t^x), & x_1 &\sim p(x_1), \end{aligned} \tag{1}$$

for $t = 1, \ldots, T$, where $T \in \mathbb{N}$ is the length of the time series, $m_t(\cdot)$ and $s_t(\cdot)$ are the (possibly non-linear) measurement and state transition functions, respectively, and we use $p(\cdot)$ to denote the probability density of the corresponding variables, which may be non-Gaussian. We shall assume that the SSM can be used to simulate paths of the states, $x_{1:T} = \{x_1, \ldots, x_T\}$, and observations, $y_{1:T}$, which holds when it is possible to draw from $p(x_1)$ and $p(\varepsilon_t^x, \varepsilon_t^y)$. The functions in (1) may depend on exogenous variables, lags of the states and observations, and on a vector of static parameters θ (also called the hyperparameters); these dependencies are suppressed in the notation for conciseness.

Once the static parameters θ have been provided or estimated, the interest is often shifted towards signal extraction, which may be performed via the conditional expectation of the states,

$$\mathbb{E}\left[x_t | Y_t\right], \tag{2}$$

for $t = 1, \ldots, T$, where Y_t denotes the conditioning set. Common choices are $Y_t = y_{1:t}$ for filtering, $Y_t = y_{1:t-k}$ for k-period forecasting, and $Y_t = y_{1:t+k}$ for smoothing, with $k \in \mathbb{N}$. If the SSM is linear and Gaussian, that is, m_t and s_t are linear functions, and all densities p are Gaussian, the conditional expectations in (2) can be computed recursively by the well-known Kalman filter (Kalman, 1960). A simple example is the following univariate Gaussian local level model,

$$\begin{aligned} y_t &= x_t + \varepsilon_t^y, & \varepsilon_t^y &\sim \mathrm{N}(0, \sigma_y^2), \\ x_{t+1} &= x_t + \varepsilon_t^x, & \varepsilon_t^x &\sim \mathrm{N}(0, \sigma_x^2), \end{aligned} \tag{3}$$

with $x_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$ for some $\mu_1 \in \mathbb{R}$ and $\sigma_1, \sigma_x, \sigma_y > 0$, and the scalar noise terms ε_t^x and ε_t^y are assumed to be mutually and serially independent, as well as independent from x_1. The local level model is a special case of the SSM in (1) with $m_t(x_t, \varepsilon_t^y) = x_t + \varepsilon_t^y$ and $s_t(x_t, \varepsilon_t^x) = x_t + \varepsilon_t^x$, normal density $p(\varepsilon_t^x, \varepsilon_t^y) = p_{\mathrm{N}}(\varepsilon_t^x) p_{\mathrm{N}}(\varepsilon_t^y)$, and $\theta = (\mu_1, \sigma_1, \sigma_x, \sigma_y)'$.

Figure 1 provides an illustration by applying the local level model to measurements of the annual flow volume of the Nile river taken at Aswan from 1871 to 1970 (Durbin & Koopman, 2012, Ch. 2). After setting the static parameters to the maximum likelihood estimates $\sigma_x = 38.329$ and $\sigma_y = 122.877$, with $\mu_1 = 0$ and $\sigma_1^2 = 10^7$ for approximate diffuse initialization, the expectations $\mathbb{E}[x_t | y_{1:t}]$ for $t = 1, \ldots, T$ can be obtained by the Kalman filter.

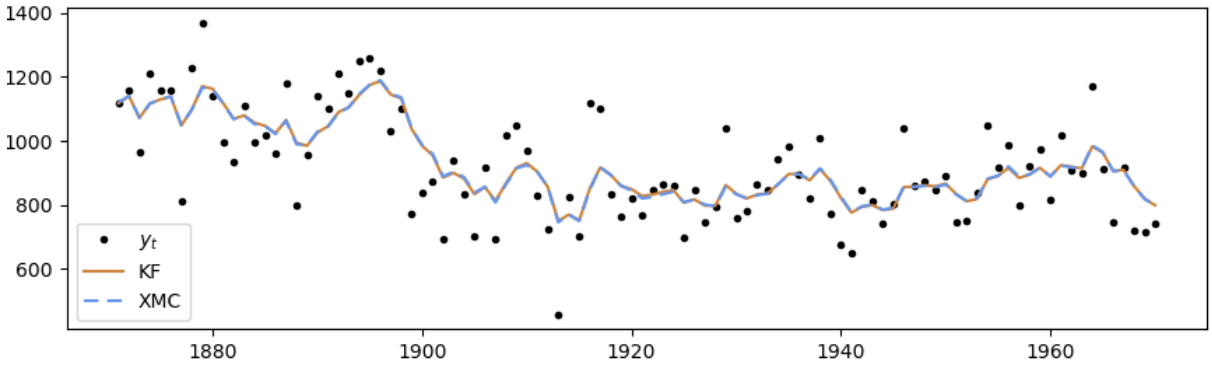

Figure 1: Analysis of the annual flow volume measurements y_t of the Nile river (discharge at Aswan in $10^8 m^3$) from 1871 to 1970 based on the local level model in (3): signals extracted via $\mathbb{E}[x_t|y_{1:t}]$ by the Kalman filter (KF) and linear extremum Monte Carlo filter (XMC) with $N = 5 \cdot 10^4$ paths and steady state reached at $t = 19$. The data are due to Cobb (1978).

In practice, however, the convenient linear Gaussian assumption appears to hold by exception, rather than the rule, so that the generalities of nonlinearity and non-Gaussian noise are often needed; see Creal (2012) and Durbin and Koopman (2012) for multiple examples in economics and finance. For many nonlinear and non-Gaussian SSMs, the computation of the conditional expectation in (2) is a challenging task and is often tackled by particle filtering methods (e.g., Gordon, Salmond, and Smith 1993; Pitt and Shephard 1999; Creal 2012). However, filtering remains challenging in cases where, for example, the SSM is charactarized by limited tractability, or when it is required to evaluate the expectation sequentially in real time.

In this study, we propose a novel simulation-based filtering method that relies on generating artificial samples of data from the SSM in (1) and estimating the conditional expectations in (2) via extremum estimation (e.g., Amemiya, 1985; Hayashi, 2000). We call this procedure *extremum Monte Carlo* (XMC) and use it to define a corresponding class of filters for signal extraction. The XMC method is mainly a filtering technique for SSMs and is therefore related to the Kalman filter and its nonlinear/non-Gaussian extensions. It is also related to the least squares Monte Carlo method (LSMC; Longstaff & Schwartz, 2001), which was developed for the valuation of American options in financial trading. A crucial step in the LSMC algorithm is the approximation of the conditional expectation function $\mathbb{E}[X|Y]$ by simulation of the random variables X and Y,

$$X^{(i)},\ Y^{(i)}, \qquad i = 1, \ldots, N.$$

The variates are then used as data in the following least squares regression,

$$\widehat{f}^N \in \underset{f \in \mathbb{F}_N}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} L\Big(X^{(i)} - f\big(Y^{(i)}\big)\Big),$$

with $L(u) = u^2$ the squared error loss, $f(\cdot)$ a prediction function, and $\mathbb{F}_N$ a suitable function space. Finally, the function estimate is used to predict X for any Y value of interest by

$$\widehat{f}^N(y) \approx \mathbb{E}[X|Y = y].$$

This general approach allows for estimating latent variables X based on observed data y.

The LSMC method is widely adopted for the valuation of derivatives with early-exercise features, as well as credit valuation adjustments (Green, 2015), and it has found many other applications. Examples range from solving backwards stochastic differential equations (Gobet, Lemor, and Warin 2005; Bender and Steiner 2012), to the estimation of complex *un*conditional moments for various dynamic volatility models (Engle, 2002). In addition, the method is increasingly being used in portfolio optimization (e.g., Denault & Simonato, 2017; R. Zhang et al., 2019), where it is called "simulation-and-regression."

In its simplest form, the XMC method consists of applying the above procedure repeatedly to perform signal extraction, by setting $X = x_t$ and $Y = \widetilde{Y}_t \subseteq Y_t$ for times $t = 1, \ldots, T$, where the covariates $\widetilde{Y}_t$ are an appropriate subset of the conditioning set. In essence, we first use the SSM in (1) to simulate paths of the states and observations, after which we regress the former onto subsets of the latter. The estimated regression functions are then evaluated at the observed data to predict the unobserved states. By allowing for loss functions $L(u)$ other than the squared error loss, the XMC method can be used to estimate various aspects of the conditional distributions of interest. Important examples are the tilted absolute error loss, $L_\tau(u) = u(\tau - 1_{\{u<0\}})$, with prediction error $u = X - f(Y)$ to estimate the conditional τ-quantile for $\tau \in (0, 1)$, and the all-or-nothing loss, $L_\delta(u) = 1_{\{|u|\geq\delta\}}$, with tolerance level $\delta > 0$ to approximate the conditional mode, which corresponds to the limit $\delta \to 0$. Each choice of function estimator and loss function yields a different filter, hence the method defines a class of *extremum Monte Carlo filters*.

While the above approach may appear computationally "naive" at first, it offers many opportunities for substantial computational savings. For example, in most cases it will not be necessary to estimate a separate regression function for each time t, so that the function estimates can simply be re-used. For illustration, consider again the local level model example. Since the Kalman filter is linear in the observations (Harvey, 1990, Ch. 3), we can attempt to mimic this filter by applying the XMC method with linear regression to minimize the squared error loss for a sample of N simulated paths. Figure 1 shows the filtered states based on the resulting linear XMC filter with $N = 5 \cdot 10^4$, which are seen to coincide with the Kalman filter. The function estimate at time $t = 19$ was re-used to filter the states for all subsequent times, hereby circumventing 81% of the regressions.

The proposed filtering method is conceptually simple and easy to implement. The combination of simulation and regression allows for a wide range of conditioning sets, including data sets with missing entries, unequal spacing, and measurements observed with mixed frequencies. The method can be applied to any model from which data can be simulated. In this way, it fills a gap in the simulation-based estimation literature by enabling signal extraction in complex models where static parameters are estimated by the method of simulated moments (McFadden, 1989) or indirect inference (Gourieroux, Monfort, & Renault, 1993). Furthermore, since most of the computations (simulation and estimation) can be performed in advance, the method is particularly suited for real-time applications, such as recommender systems in e-commerce (Schafer, Konstan, & Riedl, 1999) and algorithmic trading in finance (Kolm & Maclin, 2010).

The remainder of this paper is structured as follows. Section 2 presents the XMC method. Section 3 provides a stability and convergence analysis. Section 4 presents some illustrations to highlight and discuss the key properties of the method. Section 5 considers an empirical application to a daily time series of exchange rates. Section 6 concludes. The appendices contain proofs and other supplementary material.

2 The extremum Monte Carlo method

2.1 The basic filtering algorithm

Algorithm 1 presents the basic version of the XMC method, which consists of three fundamental steps: simulation, fitting, and prediction. For conciseness we assume that the state x_t is univariate; the vector case is handled by performing the last two steps separately for each element of x_t. The simulation step ensures that N paths are available for the states and observations. The generated data are then split in two parts. The training sample is directly used in the regressions, while the validation sample is used to regularize the tuning parameters of the chosen regression method.[1] After all regressions have been performed, the states are predicted by evaluating the estimated regression functions at the observed data. Of course, in many cases it will not be necessary to estimate a separate regression function for each time t. Substantial computational savings may therefore be obtained by re-using function estimates; see Section 2.2.

The regularization is performed by selecting the minimizer of the validation loss from

Algorithm 1 Extremum Monte Carlo filtering method.

1. **Simulate**: Use the SSM in (1) to simulate N paths of the states and observations,

$$x_{1:T}^{(i)},\ y_{1:T}^{(i)}, \qquad i = 1, \dots, N.$$

2. **Fit**:

   (a) *Split data*: Set $c_{\text{val}} \in (0, 1)$ and split the data into training and validation samples with sizes

$$N_{\text{tr}} = N - N_{\text{val}} \qquad \text{and} \qquad N_{\text{val}} = [c_{\text{val}} N].$$

   (b) *Regularization*: For a set of candidate tuning parameters, perform the following regression at time $t = t^*$:

$$\widehat{f}_t^N \in \underset{f \in \mathbb{F}_N}{\arg\min} \frac{1}{N_{\text{tr}}} \sum_{i=1}^{N_{\text{tr}}} L\left(x_t^{(i)} - f\left(\widetilde{Y}_t^{(i)}\right)\right), \tag{4}$$

   with function space $\mathbb{F}_N$ and covariates $\widetilde{Y}_t^{(i)} \subseteq Y_t^{(i)}$. Select the tuning parameters that minimize the corresponding loss for the validation sample.

   (c) *Regression*: Use the regularized tuning parameters to perform the regression in (4) at all times $t = 1, \dots, T$ to obtain the function estimates $\{\widehat{f}_t^N\}_{t=1}^T$.

3. **Predict**: Evaluate the estimated regression functions at the observed data $\widetilde{Y}_t$ for $t = 1, \dots, T$ to predict the states:

$$\widehat{x}_t = \widehat{f}_t^N(\widetilde{Y}_t).$$

[1] Given that the data can be generated at will, there is little benefit to optimizing the validation sample fraction c_{val} in Algorithm 1. It can therefore simply be set to a reasonable value, say, $c_{\text{val}} = 0.1$. The validation sample also remains useful after the regularization step because it can be used to monitor convergence, for example, by comparing the average losses incurred in the training and validation samples.

several candidate tuning parameters generated by a Bayesian optimization procedure (Bergstra, Yamins, & Cox, 2013). While this could be done for each time separately, in practice it is usually sufficient to determine the tuning parameters at some suitable time-point $t = t^*$. We consider the window size $W \in \{1, \ldots, T\}$ as a tuning parameter and define the covariate set $\widetilde{Y}_t$ to consist of the W observations from the conditioning set that are nearest to time t. For example, the conditioning set for filtering is $Y_t = y_{1:t}$, so we define the covariate set by

$$\widetilde{Y}_t = y_{\underset{\sim}{t}:t}, \qquad \text{with} \qquad \underset{\sim}{t} = \max\{t - W + 1, 1\}. \tag{5}$$

The autoregressive structure for the states in SSM (1) implies that these observations are generally the most informative on x_t. For the covariate set defined above, $t^* = T$ is a natural choice to prevent underestimation of the window size, since the validation loss is then computed where most observations are available.

By specifying the loss function and regression method, Algorithm 1 defines a corresponding XMC filter. In addition to the assumptions that are specific to the regression method (see Section 3), it is required for the loss function to have a bounded first moment. The latter is satisfied in most regression applications, and we note that the objective function in (4) can be generalized to include weights, so that one can always define a trimmed analogue of the loss function for which bounded moments are guaranteed. The XMC method further requires that the SSM in (1) can be used to simulate paths of the states and observations, which holds when it is possible to draw the initial states, $x_1 \sim p(x_1)$, and the noise terms, $(\varepsilon_t^x, \varepsilon_t^y) \sim p(\varepsilon_t^x, \varepsilon_t^y)$.

Remark 1. *The SSM is allowed to be non-stationary, a simple example of which is the local level model in (3). This non-stationarity does not pose a problem to the XMC filter because all regressions are cross-sectional, with the data $(x_{t:T}^{(i)}, y_{t:T}^{(i)})$ being IID in the index $i = 1, \ldots, N$. Figure 1 provides an illustration of this filter property.*

The optimal regression method will generally vary with the signal extraction problem. For the local level model example in the introduction, a linear regression was preferable, while in other cases a nonlinear function estimator may be more appropriate. We therefore consider several nonlinear regression methods: the tree-based gradient boosting (GB; Friedman, 2001) and the random forest (RF; Breiman, 2001) for estimating conditional means, and the generalized random forest (GRF; Athey, Tibshirani, & Wager, 2019) for estimating conditional quantiles. See also the corresponding chapters in Hastie, Tibshirani, and Friedman (2009) for a discussion of the GB and RF methods. The above methods have been chosen for their general applicability and wide use in practice, but we stress that the XMC method is not bound to any specific regression method. In principle, any function estimator can be used in Algorithm 1, from classic polynomial regression and generalized additive model regression (Hastie & Tibshirani, 1987) to deep neural network methods (LeCun, Bengio, & Hinton, 2015).

To analyze the computational complexity of the XMC method, we focus on the regression step in Algorithm 1 as it is generally the dominant runtime factor. For both the number of states N_x and the time series length T, the complexity is linear because each regression is performed separately. This separability also implies that the total runtime can be made to approximate that of the longest among the regressions by increasing the

Table 1: Computational complexity of the regression step for several XMC filters. The estimate for linear regression is based on the least squares method via the QR decomposition. For gradient boosting (GB), the estimate is based on the complexity of $O\big(CN\log(N)\big)$ for a single regression tree with $C = WN_y$ covariates. The estimates for the random forest (RF) and generalized random forest (GRF) is based on the common choice of $\sqrt{C}$ split variables for RFs (Hastie et al., 2009).

XMC filter	Linear	GB	RF and GRF
Complexity	$O\left(N_x TNC^2\right)$	$O\big(N_x TCN\log(N)\big)$	$O\big(N_x T\sqrt{C}N\log(N)\big)$

number of physical cores. On the other hand, the scaling in the number of paths N and covariates $C := WN_y$ depends on the chosen regression method and corresponding implementation. Table 1 shows current estimates of the computational complexity for several XMC filters.

2.2 Steady state filtering

A modification of Algorithm 1 is to re-use regression function estimates for prediction at other time points. We refer to this as the *steady state* (SS) XMC filter (by analogy to the Kalman filter; Harvey 1990, Ch. 3.3.3). The computational savings can be substantial for long time series, which are common in financial econometrics and with data on natural phenomena (e.g., astronomical and meteorological data).

The idea behind the SS approach is to stop performing regressions after some time t_{ss} and use the function estimate $\widehat{f}_{t_{\text{ss}}}^N$ for prediction at the remaining times $t > t_{\text{ss}}$. A minimal requirement for such approach to be sensible is that the covariate sets are translations with respect to the time index. More precisely, we require the existence of a time index $t_W \in \mathbb{N}$ such that

$$\widetilde{Y}_{t+1} = \left\{ y_{j+1} \middle| y_j \in \widetilde{Y}_t \right\} \qquad \forall\, t \geq t_W, \tag{6}$$

which means that the covariate set at time $t+1$ is the result of applying the lead operator to every observation in $\widetilde{Y}_t$. For filtering with covariate sets defined by (5), the above condition is satisfied with $t_W = W$, so that any time $t \geq t_W$ is a feasible choice for t_{ss}. We emphasize that for an SSM with time-varying locations and scales, the SS approach can still be adopted because they can be handled by standardizing the data (both simulated and observed) before performing Algorithm 1.

To determine whether the impact of the SS approach is acceptable, we propose an intuitive estimate of the largest increase in the loss. Given that the function estimates are expected to differ more from each other when they are further apart in time, we compare the estimates at times $t \geq t_W$ with the estimate at time T. For this purpose, we check for $t = t_W, \ldots, T-1$ whether the condition

$$\sum_{i=1}^{N_{\text{val}}} L\left(x_T^{\langle i\rangle} - \widehat{f}_t^N\left(\widetilde{Y}_T^{\langle i\rangle}\right)\right) \leq (1 + c_{\text{ss}}) \sum_{i=1}^{N_{\text{val}}} L\left(x_T^{\langle i\rangle} - \widehat{f}_T^N\left(\widetilde{Y}_T^{\langle i\rangle}\right)\right), \tag{7}$$

is satisfied, with chosen tolerance level $c_{\text{ss}} \geq 0$ and superscript $\langle i\rangle$ indicating a case from the validation sample $(x_{1:T}^{\langle i\rangle}, y_{1:T}^{\langle i\rangle})$, $i = 1, \ldots, N_{\text{val}}$. If the condition in (7) is satisfied, the

validation procedure is terminated, and we conclude that an SS has been reached. We then set $t_{\text{ss}} := t$ and use the estimate $\widehat{f}^N_{t_{\text{ss}}}$ to circumvent the remaining regressions.

The above SS approach was used to compute the filtered states shown in Figure 1. The condition in (7) was satisfied at time $t_{\text{ss}} = 19$ for $c_{\text{ss}} = 0$, after which the corresponding function estimate $\widehat{f}^N_{19}$ was re-used to predict the states at all subsequent times, hereby circumventing 81% of the regressions. The filtered states are seen to coincide with the Kalman filter. In Section 4.4 the impact of the SS approach on the accuracy of the XMC filter will be investigated via a simulation study.

2.3 Further extensions

The XMC method can be generalized in a number of ways. For example, it can include the prediction of functions of the states and the forecasting of observations. Both of these are established simply by changing the dependent variable in the regressions. In addition, the XMC method can be adjusted to allow for other conditioning sets. For example, with k-period forecasting, one can use the covariate set $\widetilde{Y}_t = y_{\underset{\sim}{t}:t-k}$ with $\underset{\sim}{t} = \max\{t-k-W+1, 1\}$, while $t^* = T$ is still a natural choice for performing the regularization. The method can also be extended to accommodate the various types of smoothing (Harvey, 1990, Ch. 3.6); this extension will be addressed in a follow-up study.

Besides the computation of point estimates such as $\mathbb{E}[x_t|y_{1:t}]$, one may consider using the simulated data to estimate conditional distributions. The XMC method could accommodate this by letting the loss function have the more general form $L(x_t, \widetilde{Y}_t, f)$, where $f(x_t|\widetilde{Y}_t)$ is a conditional density or probability mass function. This approach is related to the reprojection method of Gallant and Tauchen (1998), in which a long simulated path of the observations is used to perform maximum likelihood via estimates of the observation transition density, and the Bayesian amortized inference approach (e.g., Stuhlmüller, Taylor, and Goodman 2013; Cranmer, Brehmer, and Louppe 2020), in which draws from the prior are used to train a neural network approximation to the posterior density. Of course, it is also possible to extract point estimates from an estimated density, but it is generally more efficient—both statistically and computationally—to estimate the function of interest directly.

Lastly, we note that the common issue of missing data is handled naturally in the XMC method. This is done by omitting the corresponding covariates from the regressions, such that the XMC filter is fitted to the conditioning sets characterized by missing data. The same approach can be used to handle data with unequal spacing in time, or vector measurements of which the elements are observed with mixed frequencies.

3 Stability and convergence

This section considers the stability of the XMC method over time, as well as its convergence as the number of Monte Carlo draws N diverges to infinity. Our analysis takes as a given the instance of the SSM in (1), the loss function L, and the composition of the conditioning sets Y_t for $t = 1, \ldots, T$. The only restriction imposed on the conditioning set is that $Y_t \subseteq y_{1:T}$, so that in addition to filtering ($Y_t = y_{1:t}$), the results also apply to other sequential prediction problems (e.g., forecasting, smoothing). Exogenous regressors

may also be included in Y_t but will be omitted for conciseness. In addition, we assume x_t to be univariate, unless stated otherwise. This is without loss of generality due to the filter's separate treatment of the state elements.

We shall investigate the relationship between the XMC filter and an optimal filter, where the latter is made precise below. Historically, optimal filters have often been defined as the minimum mean square estimator of the states (e.g., Anderson and Moore 1979, Ch. 2; Harvey 1990, Ch. 3). The following definition generalizes this notion to other loss functions.

Definition 1 (Optimal filter)**.** For a given SSM and loss function L, an *optimal filter* $\{f_t^*(Y_t)\}_{t=1}^T$ is a set of functions such that

$$f_t^*(Y_t) \in \arg\min_{c\in\mathbb{R}} \mathbb{E}\left[L\left(x_t - c\right)\middle|Y_t\right] \tag{8}$$

holds for $t = 1, \ldots, T$ and all paths $y_{1:T}$ such that $p(y_{1:T}) > 0$.

An optimal filter is thus defined as a set of prediction functions $f_t^*(Y_t)$ that are pointwise minimizers of the expected loss, where the "points" are the conditioning sets Y_t. By the law of total expectations, the optimal filter also minimizes the unconditional mean loss. As the above definition indicates, we focus on those paths that are *realizable* in the sense that $p(y_{1:T}) > 0$ (Frühwirth-Schnatter, 1994). In the following, we use the concise notation

$$x_t^* \coloneqq f_t^*(Y_t)$$

to denote the filtered estimates.

Example (Optimal filter)**.** *For a filtering problem with the squared error loss, the objective function in* (8) *becomes*

$$\mathbb{E}\left[L\left(x_t - c\right)\middle|Y_t\right] = \mathbb{E}[(x_t - c)^2|y_{1:t}].$$

If the objective function exists, it is well known that the corresponding minimizer x_t^* is unique and given by the conditional expectation $\mathbb{E}[x_t|y_{1:t}]$. These expectations can be computed by the Kalman filter if the SSM is linear and Gaussian (Anderson & Moore, 1979). The model is then of the form

$$\begin{aligned} y_t &= H_t x_t + \varepsilon_t^y, & \varepsilon_t^y &\sim \mathrm{N}(0, R_t), \\ x_{t+1} &= F_t x_t + G_t \varepsilon_t^x, & \varepsilon_t^x &\sim \mathrm{N}(0, Q_t), \end{aligned} \tag{9}$$

for $t = 1, \ldots, T$, with initial state $x_1 \sim \mathrm{N}(0, \Sigma_1)$ that is independent of $\{\varepsilon_t^x\}$ and $\{\varepsilon_t^y\}$, the noise terms are serially and mutually independent, and with possibly time-varying system matrices F_t, G_t, H_t, R_t, and Q_t of appropriate dimensions. The above model reduces to the local level model in (3) *for $F_t = G_t = H_t = 1$, $\Sigma_1 = \sigma_1^2$, $R_t = \sigma_y^2$ and $Q_t = \sigma_x^2$. The Kalman recursions (Harvey, 1990, Ch. 3) for $t = 1, \ldots, T$ are*

$$\begin{aligned} \mathbb{E}[x_t|y_{1:t}] &= x_{t|t-1}^* + \Sigma_{t|t-1} H_t' (H_t \Sigma_{t|t-1} H_t' + R_t)^{-1} (y_t - H_t x_{t|t-1}^*), \\ \mathbb{V}\mathrm{ar}\left[x_t|y_{1:t}\right] &= \Sigma_{t|t-1} - \Sigma_{t|t-1} H_t' (H_t \Sigma_{t|t-1} H_t' + R_t)^{-1} H_t \Sigma_{t|t-1}, \\ K_t &= F_t \Sigma_{t|t-1} H_t' (H_t \Sigma_{t|t-1} H_t' + R_t)^{-1}, \\ x_{t+1|t}^* \coloneqq \mathbb{E}[x_{t+1}|y_{1:t}] &= (F_t - K_t H_t) x_{t|t-1}^* + K_t y_t, \\ \Sigma_{t+1|t} \coloneqq \mathbb{V}\mathrm{ar}\left[x_{t+1}|y_{1:t}\right] &= F_t \Big(\Sigma_{t|t-1} - \Sigma_{t|t-1} H_t' (H_t \Sigma_{t|t-1} H_t' + R_t)^{-1} H_t \Sigma_{t|t-1}\Big) F_t' \\ &\quad + G_t Q_t G_t', \end{aligned} \tag{10}$$

with initialization

$$x^*_{1|0} \coloneqq \mathbb{E}[x_1], \qquad \Sigma_{1|0} \coloneqq \Sigma_1.$$

By analogy to the optimal filter, the XMC filter is a set of function estimators that are used for prediction, which can be represented by

$$\{\widehat{f}^N_t(\widetilde{Y}_t)\}_{t=1}^T, \qquad \widetilde{Y}_t \subseteq Y_t, \tag{11}$$

where the *covariate set* $\widetilde{Y}_t$ is a subset of the conditioning set at time t, and we use a similar shorthand notation as before for the filtered estimates,

$$\widehat{x}^N_t \coloneqq \widehat{f}^N_t(\widetilde{Y}_t).$$

The power set $\mathcal{P}(Y_t)$ is the collection of all feasible covariate sets. The number of covariates used in the regressions represents a trade-off between the bias and variance of the filter, where the optimal number generally depends on the number of draws; see Appendix D.2 for an illustration.

Example (Linear XMC filter)**.** *Consider the XMC filter defined by using a linear regression function with parameters estimated by the least squares method,*

$$\widehat{f}^N_t(\widetilde{Y}_t) = \sum_{y_j \in \widetilde{Y}_t} \widehat{\beta}_{j,t} y_j. \tag{12}$$

for $t = 1, \ldots, T$, where we omit the intercept term for conciseness. Each feasible covariate set $\widetilde{Y}_t \in \mathcal{P}(Y_t)$ defines a different function estimator $\widehat{f}^N_t(\widetilde{Y}_t)$. For instance, filtering at time $t = 2$ with $1 = 1$ has as conditioning set $Y_t = \{y_1, y_2\}$, which yields the following three possible estimators,

$$\widehat{f}^N_2(\{y_1\}) = \widehat{\beta}^1_{1,2} y_1, \qquad \widehat{f}^N_2(\{y_2\}) = \widehat{\beta}^2_{2,2} y_2, \qquad \widehat{f}^N_2(\{y_1, y_2\}) = \widehat{\beta}^3_{1,2} y_1 + \widehat{\beta}^3_{2,2} y_2,$$

in addition to the trivial estimator $\widehat{f}^N_2(\emptyset) = 0$, each of which could be used to predict the state x_2.

For any filtering method to be useful in real-time applications, it is required that both the statistical errors and the computational costs for any time t remain bounded as t increases. For simulation-based methods, this essentially means that the errors must remain bounded while the number of Monte Carlo draws N remains fixed. This important property is often difficult to establish for recursive methods, which pertains to the majority of filters used in practice. For example, the standard assumptions for particle filters to obtain stability require compactness of the state space, which is usually not met in practice; see Chopin and Papaspiliopoulos (2020, Ch. 11) for a discussion. However, the non-recursive nature of the XMC method makes it straightforward to derive this stability property under mild assumptions. The result below considers the general setting in which $\{(x_t, y_t)\}$ is generated by a strictly stationary process. Moreover, we emphasize that the result remains valid under model misspecification, in which case the SSM does not correspond to the process that has generated the data. In the following, $\|\cdot\|_p = (\mathbb{E}|\cdot|^p)^{1/p}$ denotes the $\mathcal{L}^p$-norm, where the expectation is with respect to the Monte Carlo draws and the actual observations.

Theorem 1 (Filter stability). *Suppose that* $\{(x_t, y_t)\}_{t\in\mathbb{N}}$ *is strictly stationary, and let the loss function be of the form* $L(u) = |u|^p$ *for some* $p \geq 1$ *such that* $\|x_1\|_p < \infty$. *Assume there exists an optimal filter* $\{x_t^*\}$ *such that the* $x_t^* = f_t^*(Y_t)$ *are measurable functions of the conditioning sets* $Y_t = y_{1:t}$. *For some* $N, W \in \mathbb{N}$, *consider an XMC filter* $\{\widehat{x}_t^N\}$ *defined via Algorithm 1 such that* $\|\widehat{x}_t^N\|_p < \infty$ *for* $t \leq W$, *and the* $\widehat{x}_t^N = \widehat{f}_t^N(\widetilde{Y}_t)$ *are measurable functions of the covariate sets* $\widetilde{Y}_t$ *given by (5). Then, the filter approximation errors remain bounded in* $\mathcal{L}^p$*-norm:*

$$\sup_{t\in\mathbb{N}} \|x_t^* - \widehat{x}_t^N\|_p < \infty.$$

Proof. See Appendix A.1. □

Theorem 1 establishes mild conditions that ensure stability of the XMC filter. The loss being of absolute power form means that the result applies to the popular squared and absolute error loss functions. The strict stationarity assumption on $\{(x_t, y_t)\}$ implies that the initial states are drawn from their long-run distribution, but this is only used to simplify the proof; similar results can be obtained by allowing for other initializations of the SSM. Notably, the SSM is allowed to be misspecified—as is typically the case in practice, in which case we require the assumed SSM to be strictly stationary. The assumption that the covariate sets are defined via (5) can be replaced by the more general requirement that they satisfy the translation condition in (6), which means that the result also applies to k-period forecasting and fixed-lag smoothing. The assumption that the conditioning sets are of the filtering type can be relaxed by the requirement that they are non-decreasing in time, $Y_t \subseteq Y_{t+1}$, $t \in \mathbb{N}$, which accommodates the other types of signal extraction mentioned above. Lastly, we note that the result also applies to the SS approach from Section 2.2, which plays an important role in real-time filtering with the XMC method.

The property of $\mathcal{L}^p$-bounded errors is reassuring, but as a consequence of the mild assumptions, the above result does not tell us whether and at which rate the XMC filter converges to an optimal filter as the number of draws N diverges. To this end, we consider the pointwise convergence

$$\widehat{x}_t^N \xrightarrow{P} x_t^* \qquad \text{as} \qquad N \to \infty,$$

for $t = 1, \ldots, T$ and all realizable paths $y_{1:T}$.[2] It will be assumed that the time series length $T \in \mathbb{N}$ is finite, though it can be arbitrarily large. The case in which both the time series length and the number of draws diverge simultaneously is explored in Appendix B. Contrary to Theorem 1, we will now allow for non-stationary processes (see Remark 1).

The filtered estimates $\widehat{x}_t^N$ are obtained by evaluating the XMC filter at suitably chosen covariate sets. To cover the out-of-sample optimization procedure from Section 2, we allow the covariate sets to depend on the number of draws N, which is made explicit by the notation $\widetilde{Y}_t^N$. We make the following assumption for these *regularized* covariate sets.

[2] The pointwise mode of convergence corresponds to the assumption that the actual observations $y_{1:T}$ are fixed, which is standard in particle filter convergence analyses (e.g., Crisan & Doucet, 2002, Sec. 4).

Assumption 1 (Convergence of regularized covariate sets)**.** *For $t = 1, \ldots, T$ the regularized covariate set converges in probability to the conditioning set,*

$$\lim_{N \to \infty} P(\widetilde{Y}_t^N = Y_t) = 1,$$

where the convergence is with respect to the Monte Carlo draws.

Apart from the above assumption, the way in which the regularized covariate sets depend on N will be left open. This allows for many other common methods of regularization, such as the use of a fixed growth rate for the window size, or adding a penalty term for the latter to the objective function. Sufficient conditions to guarantee that Assumption 1 holds for some common regularization methods are given in Appendix C.

The following result provides sufficient conditions under which the linear XMC filter converges to the Kalman filter in the important special case of linear Gaussian SSMs.

Theorem 2 (Convergence to Kalman filter)**.** *Let Assumption 1 hold, and suppose the following holds:*

A2.1 The SSM is linear and Gaussian as in (9), the initial variance Σ_1 of the states is bounded, and the observations in $y_{1:T}$ are linearly independent.

A2.2 L is the squared error loss.

A2.3 The XMC filter is linear as in (12).

Then, the XMC filter converges in probability to the Kalman filter $\{x_t^*\}$ at rate $\sqrt{N}$, such that for all realizable paths $y_{1:T}$ we have

$$\sup_t \sqrt{N} \left| x_t^* - \widehat{x}_t^N \right| = O_P(1).$$

Proof. See Appendix A.2. □

In Theorem 2, the assumption that Σ_1 is bounded ensures that the variance of the states and observations is bounded, which is needed for obtaining the $\sqrt{N}$ convergence rate. Strictly speaking, this assumption rules out diffuse initialization, but the latter can be approximated arbitrarily well by a sufficiently large but finite initial variance, as is often done in practice. The assumption that the observations are linearly independent holds for most SSMs of practical interest, with a sufficient condition being that the measurement noise is non-degenerate.

Theorem 2 can be generalized in a number of ways. First, the loss function may be chosen different from the squared error loss. For example, Assumption A2.2 can be relaxed to allow for the absolute error loss via Theorem 2 of Pollard (1991). Second, the SSM can be nonlinear and/or non-Gaussian, which is often needed in practice. And third, the XMC filter can rely on nonlinear regression methods, which are often needed once the linear Gaussian assumption is relaxed.

As a natural generalization, one could consider a parametric nonlinear XMC filter. Although this approach is legitimate, parametric estimators are most suitable when there are strong indications on the functional form of the optimal filter, and this knowledge

is typically unavailable with nonlinear non-Gaussian SSMs. In such settings, nonparametric estimators may offer a suitable solution, as this class contains a large number of alternatives that guarantee convergence under mild assumptions, with many results on convergence rates available in the literature. For example, T. Zhang and Yu (2005) discuss convergence rates for general boosting procedures with early stopping, Peng, Coleman, and Mentch (2022) provide results for the RF method, while overviews of convergence rates for (semi-)nonparametric methods are given by van de Geer (2000) and Chen (2007). In particular, the latter reference discusses several common variants of neural networks.

As it is beyond the scope of this paper to provide a separate analysis for all the alternatives mentioned above, we instead provide a general auxiliary result which substantially simplifies proving convergence for any specific regression method. The result has also been used to establish Theorem 2. In particular, where the latter result has demonstrated that the usual $\sqrt{N}$ convergence rate of the least squares estimator is preserved by the linear XMC filter, the following result shows that this property is neither specific to the rate $r_N = \sqrt{N}$, nor to the linear version of the XMC filter.

Lemma 1 (General filter convergence)**.** *Suppose that for the given SSM and loss function, there exists an optimal filter* $\{x_t^*\}$ *in accordance with Definition 1. Let Assumption 1 hold, and let* $r_N\big|x_t^* - \widehat{f}_t^N(Y_t)\big| = O_P(1)$ *for* $t = 1, \dots, T$ *with rate* $r_N > 0$ *that diverges as* $N \to \infty$*. Then, the XMC filter converges in probability to the optimal filter at rate* r_N*, such that for all realizable paths* $y_{1:T}$ *we have*

$$\sup_t r_N\big|x_t^* - \widehat{x}_t^N\big| = O_P(1).$$

Proof. See Appendix A.3. □

The above result establishes that under Assumption 1, neither the consistency property nor the convergence rate of the function estimators is impacted by the use of a regularized covariate set instead of the conditioning set. It follows that the regularized covariate set can be ignored in convergence analyses, hence we can conclude that, asymptotically, the XMC filter is as good as the regression method it uses.

4 Filter properties: illustrations and discussion

This section presents several illustrations to highlight and discuss the key properties of the XMC method. We focus on filtering via the conditional means of the states $\mathbb{E}[x_t|y_{1:t}]$ for $t = 1, \dots, T$; additional illustrations can be found in Appendix D. In all applications we set the validation sample fraction to $c_{\text{val}} = 0.1$, and the noise terms ε_t^x and ε_t^y are assumed to be mutually and serially independent, as well as independent of the initial state x_1.

4.1 Nonlinear filtering

Consider the following nonlinear model for a univariate time series y_t as given by

$$\begin{aligned} y_t &= \frac{x_t^2}{20} + \varepsilon_t^y, & \varepsilon_t^y &\sim \mathrm{N}(0, \sigma_y^2), \\ x_{t+1} &= \frac{1}{2}x_t + \frac{25x_t}{1+x_t^2} + 8\cos\big(1.2(t+1)\big) + \varepsilon_t^x, & \varepsilon_t^x &\sim \mathrm{N}(0, \sigma_x^2), \end{aligned} \tag{13}$$

with $x_1 \sim \mathrm{N}(0,1)$, and the static parameters are set to $\sigma_x^2 = 0.1$ and $\sigma_y^2 = 1$ as in Kitagawa (1996). This model is a special case of the SSM in (1) and is often used for illustrating the performance of nonlinear filters. We use it first to simulate a single path of the states $x_{1:T}$ and observations $y_{1:T}$ of length $T = 100$. The simulated states are then predicted using the corresponding observations by means of several XMC filters based on the above model with $N = 10^5$. The simulated observations are shown in Figure 2 (a), while Part (b) shows the simulated states and their estimates based on the GB-XMC filter. The estimates are typically very close to the true states.

For comparison purposes, we adopt the bootstrap filter (Gordon et al., 1993), which is a standard version of the particle filter. This method requires an importance sampler to draw N values of the states $x_t^{(i)}$, $i = 1, \ldots, N$, which are called the particles. In the bootstrap filter, these are drawn as $x_1^{(i)} \sim p(x_1)$ and $x_{t+1}^{(i)} = s_t\big(x_t^{(i)}, z^{(i)}\big)$, where $z^{(i)}$ is a draw of ε_t^x. The particles are then weighted to form a discrete approximation to the density $p(x_t|y_{1:t})$, which yields $\mathbb{E}\left[x_t|y_{1:t}\right] \approx \sum_{i=1}^N \omega_t^{(i)} x_t^{(i)}$, with convex weights $\omega_t^{(i)} \propto \omega_{t-1}^{(i)} p(y_t|x_t^{(i)})$ and $\omega_0^{(i)} = 1/N$. To prevent the weights from degenerating, the particles are resampled when the "effective sample size," defined as $\mathrm{ESS}_t = 1/\sum_{i=1}^N \big(\omega_t^{(i)}\big)^2 \in [1, N]$, drops below $N/2$ (Doucet, De Freitas, & Gordon, 2001, p. 333). We set the number of particles to $N = 10^7$ to ensure a highly accurate approximation to the filtering means. The resulting filtered states are shown in Figure 2 (b), while Part (c) shows the difference from the bootstrap filter for several XMC filters. By comparing the scales of Figures 2 (b) and (c) we find that the GB- and RF-XMC filters are generally adequate, while the linear filter (Lin-XMC) is not. The latter remains unchanged when the number of draws is increased, which indicates that the means $\mathbb{E}\left[x_t|y_{1:t}\right]$ are inherently nonlinear in the observations. This example demonstrates the need for general regression methods such as GB and RF in the XMC method.

4.2 Real-time speed and accuracy

This section discusses the relation between real-time speed and accuracy of the XMC method. In real-time applications, the observations $y_{1:t}$ are not known in advance, which means that the estimated regression functions must be accurate on most of their domain. It is therefore expected that a larger number of draws N is needed to achieve the same accuracy as other simulation-based methods that provide direct point estimates (e.g., particle filters). However, an important property of the XMC method is that most of the computations take place in the simulation and fitting steps, which can be performed off-line. The on-line (or real-time) phase then only consists of the prediction step, which is computationally light. Furthermore, it is expected that the impact of increasing N on computing the predictions is small, so that the desired level of accuracy may be achieved without compromising the real-time speed.

To illustrate this last point, we performed a simulation study using the nonlinear model in (13). The root mean squared error (RMSE) and runtimes of the GB-XMC filter are compared with those of the bootstrap filter for various values of N. Table 2 shows the results from the simulation study, in which the path length was set to $T = 100$, and $N_{\text{test}} = 10^4$ simulated test paths were used to estimate the performance. As expected, the bootstrap filter is more accurate than the XMC filter for an equal number of draws. However, the time spent in the on-line phase by the XMC filter is several orders of

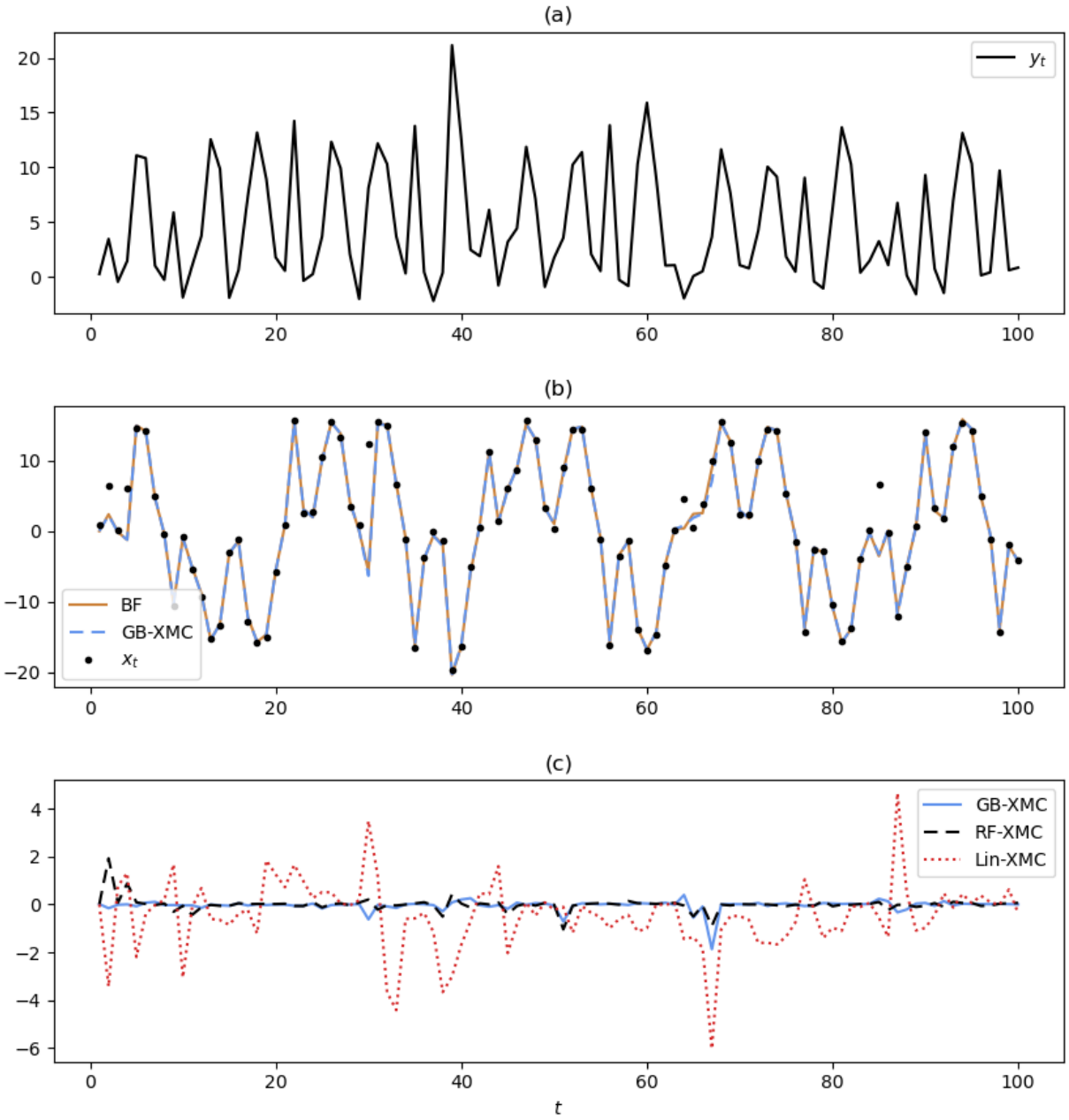

Figure 2: Analysis of a simulated path from the nonlinear model in (13): (a) observations; (b) true and filtered states by the bootstrap filter (BF) and gradient boosting (GB) XMC filter; (c) differences with BF for the GB, random forest (RF), and linear (Lin) XMC filters. The BF is based on 10^7 particles; the XMC filters are based on 10^5 simulated paths.

magnitude smaller. These times are impacted by the number of draws only indirectly—via the estimated regression functions, with runtimes that are not necessarily increasing in N. In this case, the runtime is larger for $N = 10^3$ than for $N = 10^4$ because the GB method has a tendency to overfit the training data for small samples, which results in function estimates that are more complex and, therefore, more expensive to evaluate. By contrast, the bootstrap filter incurs all its computing costs in real time, and the runtimes are directly impacted by the number of draws. The results illustrate that particle filters

Table 2: Results from simulation study based on the nonlinear model in (13): overall root mean squared error (RMSE) and runtimes (computer execution time in seconds) based on $N_{\text{test}} = 10^4$ test paths for the bootstrap filter (BF) and gradient boosting XMC filter methods with various number of draws N. The results are generated by a computer with an Intel i5 quad-core processor having 3.3 GHz clock frequency. The software is written in `Python` and is optimized by calling various functions from pre-compiled `C/C++` code.

$\log_{10}(N)$	3		4		5	
Method	BF	XMC	BF	XMC	BF	XMC
RMSE	1.688	1.858	1.664	1.709	1.662	1.674
Runtime off-line	-	72.1	-	437.2	-	4240.6
Runtime on-line	361.3	**3.0**	1295.7	**2.4**	13723.2	**3.4**

have an inherent trade-off between real-time speed and accuracy, whereas XMC filters do not. The computational "bottleneck" in Algorithm 1 can be executed off-line, which makes the XMC method particularly suited for real-time applications.

4.3 Missing data

In practice it often occurs that some of the data are missing. The XMC method handles this issue naturally by omitting the corresponding covariates from the regressions. As illustration, we consider the LL model example from the introduction and treat the Nile measurements at times $t = 21, \ldots, 40$ and $t = 61, \ldots, 80$ as missing (Durbin & Koopman, 2012, Ch. 2). The resulting data set is shown in Figure 3. To deal with these longer sequences of missing data, the window size was set to 40. Figure 3 shows the filtered states from the linear XMC filter with $N = 10^5$ paths. The predictions are seen to coincide with those of the Kalman filter, which has an exact treatment of missing data (Durbin & Koopman, 2012, Ch. 4.10).

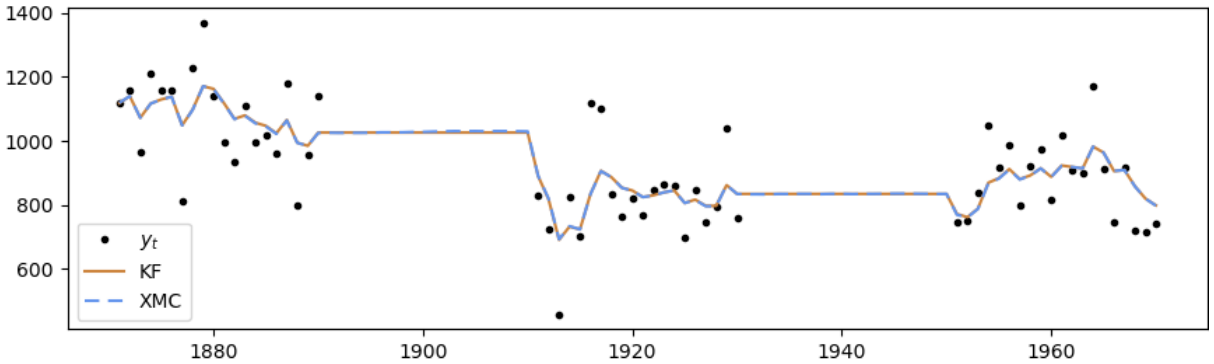

Figure 3: Filtering analysis based on the local level model in (3) and the partial Nile data set, in which the observations at time points $21, \ldots, 40$ and $61, \ldots, 80$ are treated as missing: filtered states from the Kalman filter (KF) and linear XMC filter with $N = 10^5$.

4.4 Intractable model densities

The stochastic volatility (SV) model is often used for the modeling of time series of daily financial returns. We consider the SV model with stable measurement noise (e.g., Vankov, Guindani, & Ensor, 2019) given by

$$\begin{aligned} y_t &= \exp(x_t/2)\varepsilon_t^y, & \varepsilon_t^y &\sim \mathrm{S}(\alpha,\beta), \\ x_{t+1} &= \mu + \phi(x_t-\mu) + \sigma_x\varepsilon_t^x, & \varepsilon_t^x &\sim \mathrm{N}(0,1), \end{aligned} \tag{14}$$

where x_t represents the unobserved log volatility, with $x_1 \sim \mathrm{N}\big(\mu, \sigma_x^2/(1-\phi^2)\big)$ and static parameters $\mu \in \mathbb{R}$, $|\phi| < 1$, and $\sigma_x > 0$. Furthermore, $\mathrm{S}(\alpha,\beta)$ denotes the first parametrization of the standard univariate stable distribution as in Nolan (2009), with tail index parameter $\alpha \in (0,2]$ and asymmetry parameter $\beta \in [-1,1]$. Except for a few specific choices of the parameters, the density is not available in closed form, so that in general, the characteristic function is used to describe the distribution:

$$\mathbb{E}\{\exp(iu\varepsilon_t^y)\} = \begin{cases} \exp\Big(-|u|\big(1 + i\beta\tfrac{2}{\pi}(\operatorname{sgn} u)\log|u|\big)\Big) & \text{if } \alpha = 1, \\ \exp\Big(-|u|^\alpha\big(1 - i\beta\tan(\tfrac{\pi\alpha}{2})\operatorname{sgn} u\big)\Big) & \text{otherwise.} \end{cases}$$

Simulation from the stable distribution can be performed using the method of Chambers, Mallows, and Stuck (1976).

An important property of the XMC method is that it circumvents most issues related to limited tractability because, in principle, the only required model-specific knowledge is a sample of draws of the states and observations. To illustrate this, we performed a simulation study using the SV model in (14) with the parameter choice from Vankov et al. (2019), that is, $\mu = -0.2$, $\phi = 0.95$, $\sigma_x = 0.2$, $\alpha = 1.75$ and $\beta = 0.1$. The path length is set to $T = 100$ and the number of simulated test paths for evaluating the filter performance to $N_{\text{test}} = 10^5$, as is the number of draws N for the XMC method.

As a simple benchmark, we consider the quasi-maximum likelihood (QML) filter of Harvey, Ruiz, and Shephard (1994), which remains valid without a tractable observation density. The method is based on transforming the observations by $\widetilde{y}_t = \log y_t^2$ to cast the SV model into the linear state space form given by

$$\begin{aligned} \widetilde{y}_t &= x_t + 2\widetilde{\varepsilon}_t^y, \\ x_{t+1} &= (1-\phi)\mu + \phi x_t + \sigma_x\varepsilon_t^x, \end{aligned} \tag{15}$$

with $\widetilde{\varepsilon}_t^y = \log|\varepsilon_t^y|$. Although $\widetilde{\varepsilon}_t^y$ is not normally distributed, one can assume it is, so that the Kalman filter can be used to act as an approximate filter for x_t. This normal approximation matches the first two moments of $\widetilde{\varepsilon}_t^y$, which are given in Lemma 3.19 of Nolan (2009).

Figure 4 shows the RMSE at different time points for the QML filter (brown) and the basic (blue, dashed) and SS (red, dotted) XMC filter. The SS estimate $\widehat{f}_{t_{\text{ss}}}^N$ corresponds to $t_{\text{ss}} = 25$ ($c_{\text{ss}} = 0; W = 21$), such that only a small part of the regressions had to be performed. The SS approach is seen to have no material impact on the accuracy of the XMC filter. This result is not surprising, since the process in (14) is strictly stationary, so that the limit estimators $\lim_{N\to\infty}\widehat{f}_t^N$ for covariate sets satisfying the condition in (6)

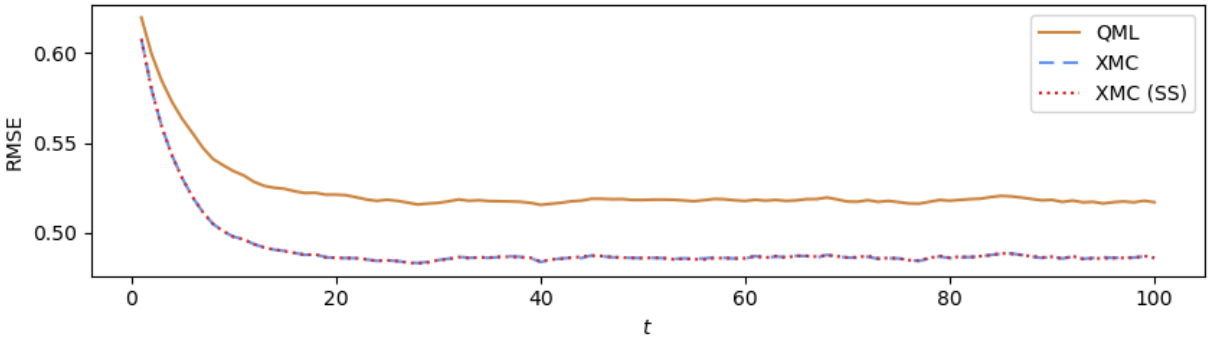

Figure 4: Results from simulation study based on the SV model in (14): the RMSEs over time for the quasi-maximum likelihood (QML) filter and the gradient boosting XMC filter ($N = 10^5$), which is applied with and without the steady state (SS) modification of Section 2.2.

are the same for all $t \geq t_W$. Shortly after $t = 20$, the RMSE stops decreasing for all filters, which suggests a relatively small contribution from further lags. The XMC filter is seen to outperform the QML filter at all time points.

Comparison with recent approximate filtering methods

In a simulation study based on the same static parameters and $T = 350$, Vankov et al. (2019, Fig. 2, p. 38) report that the *minimum* RMSE out of a 100 runs of their approximate Bayesian computation (ABC) filter with $5 \cdot 10^3$ particles exceeds 0.82, while that attained by the ABC filter of Jasra, Singh, Martin, and McCoy (2012) exceeds 0.92. The benchmark QML filter outperforms the ABC filters with an overall RMSE of 0.524, and although the accuracy of the ABC filters could be improved by increasing the number of particles, the difference with the benchmark is of such a magnitude that further consideration does not seem worthwhile. The overall RMSE for both versions of the XMC filter is 0.492, which corresponds to a substantial improvement in accuracy. The above illustrates that the XMC filter provides an accurate alternative to ABC filters in settings characterized by limited tractability.

5 Empirical application

As empirical application, we provide an analysis of the log returns of the British Pound against the Deutsche Mark from January 1, 1987 to December 31, 1995, which is shown in Figure 5 (a). The time series consists of $T = 2347$ observations, which means that substantial computational savings may be obtained via the SS approach from Section 2.2. The data set is characterized by volatility clustering and contains several extreme observations. The largest negative return corresponds to the speculative attack on September 16, 1992, which led the British monetary authorities to abandon their shadowing policy of the Deutsche Mark. This data set is considered by Lombardi and Calzolari (2009), who have used the indirect inference method (Gourieroux et al., 1993) to estimate the static parameters of several stable SV models, including a symmetric version of (14). The indirect inference estimator can be applied to complex models, but it does not allow for the estimation of latent variables. In the application of Lombardi and Calzolari (2009),

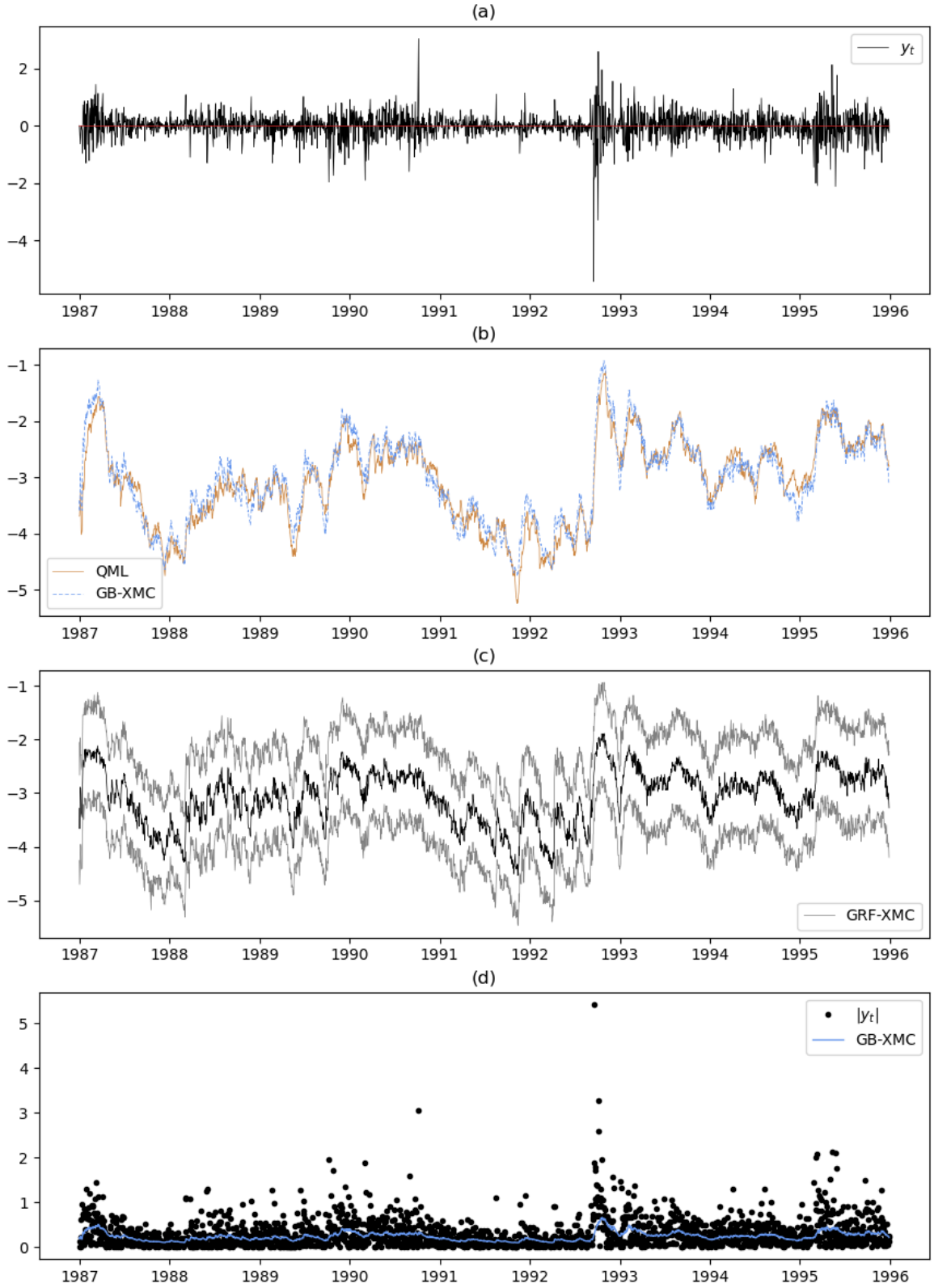

Figure 5: Filtering analysis of the daily log returns of the British Pound against the Deutsche Mark from January 1, 1987 to December 31, 1995: (a) log returns times 100; (b) state means, $\mathbb{E}[x_t|y_{1:t}]$, by the quasi-maximum likelihood (QML) and gradient boosting (GB) XMC filters; (c) $10\%, 50\%$, and 90% quantiles of the states by the generalized random forest (GRF) XMC filter; (d) volatility estimates, $\mathbb{E}[\exp(x_t/2)|y_{1:t}]$, by the GB-XMC filter. The XMC estimates were obtained using the steady state approach ($c_{\text{ss}} = 0, N = 10^5$). Data source: `https://fxtop.com/`.

the estimation of the states x_t or the volatility $\exp(x_t/2)$ is omitted; this estimation forms an important part of the analysis below.

To analyze the time-varying volatility of the exchange rate, we filter the states and volatility with the XMC method based on the stable SV model in (14). We set the static parameters to their estimates from Lombardi and Calzolari (2009): $\phi = 0.994$, $\sigma_x = 0.094$, $\alpha = 1.796$, $\beta = 0$, and with $\mu = -3.069$ the method of moments estimate based on the QML transformation in (15).

Figure 5 (b) shows the filtered states based on their estimated means, $\mathbb{E}[x_t|y_{1:t}]$, by the QML and GB-XMC filters for $N = 10^5$. The estimates from both filters are similar and a comparison with Figure 5 (a) shows that higher estimates correspond to periods with larger movements of the exchange rates, as expected. The largest estimate of the state coincides with the crash on September 16, 1992. Part (c) shows the $10\%, 50\%$, and 90% quantiles of the filtering density, $p(x_t|y_{1:t})$, by the GRF-XMC filter. The quantiles show similar movements as the mean estimates, but the median is typically somewhat lower, which indicates that the filtering density is at times skewed to the right. In Part (d), the absolute values of the log returns are shown, as well as the GB-XMC filtered volatility based on $\mathbb{E}[\exp(x_t/2)|y_{1:t}]$. The estimated volatility is higher in periods with larger movements of the exchange rates, as was the case for the filtered states.

For the above applications, the SS based on $c_{\text{ss}} = 0$ is reached either immediately or almost immediately, at $t = 89$ ($W = 89$, Part b), $t = 50$ ($W = 42$, Part c), and $t = 75$ ($W = 75$, Part d). Less than 4% of the $T = 2347$ maximum possible regressions are performed in each case, which illustrates the computational efficiency of the SS approach.

6 Conclusion

This paper introduces a novel simulation-based filtering method for general state space models. It can be used to compute time-varying conditional means, modes, and quantiles, and for predicting latent variables. The XMC method consists of generating artificial samples of data from the model and estimating quantities of interest via an extremum estimation method. The approach is conceptually simple and easy to implement. It can be applied to any model from which data can be simulated. Since most computations can be performed in advance, the method is particularly suited for real-time applications. The XMC filter is shown to be stable over time under mild assumptions, a result that remains valid under model misspecification. Conditions are provided for convergence to an optimal filter as the number of draws diverges. In particular, the linear version of the XMC filter converges to the Kalman filter in the linear Gaussian setting. Illustrations are presented for problems characterized by nonlinearity, missing data, and intractable density functions. The empirical application to a long time series of exchange rates demonstrates that, despite a setting of limited tractability, the method is able to efficiently extract the time-varying volatility.

References

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press.
Anderson, B., & Moore, J. B. (1979). Optimal filtering. *Prentice-Hall*.

Andrews, D. W. (1988). Laws of large numbers for dependent non-identically distributed random variables. *Econometric Theory*, *4*(3), 458–467.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests.

Bender, C., & Steiner, J. (2012). Least-squares Monte Carlo for backward SDEs. In *Numerical methods in finance* (pp. 257–289). Springer.

Bergstra, J., Yamins, D., & Cox, D. (2013). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning* (pp. 115–123).

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Chambers, J. M., Mallows, C. L., & Stuck, B. (1976). A method for simulating stable random variables. *Journal of the American Statistical Association*, *71*(354), 340–344.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, *6*, 5549–5632.

Chopin, N., & Papaspiliopoulos, O. (2020). *An introduction to sequential monte carlo*. Springer.

Cobb, G. W. (1978). The problem of the nile: Conditional solution to a changepoint problem. *Biometrika*, *65*(2), 243–251.

Cranmer, K., Brehmer, J., & Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, *117*(48), 30055–30062.

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

Crisan, D., & Doucet, A. (2002). A survey of convergence results on particle filtering methods for practitioners. *IEEE Transactions on signal processing*, *50*(3), 736–746.

Denault, M., & Simonato, J.-G. (2017). Dynamic portfolio choices by simulation-and-regression: Revisiting the issue of value function vs portfolio weight recursions. *Computers & Operations Research*, *79*, 174–189.

Doucet, A., De Freitas, N., & Gordon, N. J. (2001). *Sequential Monte Carlo methods in practice* (Vol. 1) (No. 2). Springer.

Durbin, J., & Koopman, S. J. (2012). *Time series analysis by state space methods*. Oxford University Press.

Engle, R. F. (2002). New frontiers for ARCH models. *Journal of Applied Econometrics*, *17*(5), 425–446.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189–1232.

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, *15*(2), 183–202.

Gallant, A. R., & Tauchen, G. (1998). Reprojecting partially observed systems with application to interest rate diffusions. *Journal of the American Statistical Association*, *93*(441), 10–24.

Gobet, E., Lemor, J.-P., & Warin, X. (2005). A regression-based monte carlo method to solve backward stochastic differential equations. *The Annals of Applied Probability*, *15*(3), 2172–2202.

Gordon, N. J., Salmond, D. J., & Smith, A. F. (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. In *IEEE Proceedings F (radar and signal processing)* (Vol. 140, pp. 107–113).

Gourieroux, C., Monfort, A., & Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics*, *8*(S1), S85–S118.

Green, A. (2015). *XVA: credit, funding and capital valuation adjustments*. John Wiley & Sons.

Harvey, A. C. (1990). Forecasting, structural time series models and the kalman filter.

Harvey, A. C., Ruiz, E., & Shephard, N. (1994). Multivariate stochastic variance models. *The Review of Economic Studies*, *61*(2), 247–264.

Hastie, T., & Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, *82*(398), 371–386.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer.

Hayashi, F. (2000). *Econometrics*. Princeton University Press.

Jasra, A., Singh, S. S., Martin, J. S., & McCoy, E. (2012). Filtering via approximate bayesian computation. *Statistics and Computing*, *22*(6), 1223–1237.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, *82*(Series D), 35–45.

Kitagawa, G. (1996). Monte carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, *5*(1), 1–25.

Kolm, P. N., & Maclin, L. (2010). Algorithmic trading. *Encyclopedia of Quantitative Finance*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Lombardi, M. J., & Calzolari, G. (2009). Indirect estimation of α-stable stochastic volatility models. *Computational Statistics & Data Analysis*, *53*(6), 2298–2308.

Longstaff, F. A., & Schwartz, E. S. (2001). Valuing American options by simulation: a simple least-squares approach. *The Review of Financial Studies*, *14*(1), 113–147.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 995–1026.

Nolan, J. P. (2009). Univariate stable distributions. *Stable Distributions: Models for Heavy Tailed Data*, *22*(1), 79–86.

Peng, W., Coleman, T., & Mentch, L. (2022). Rates of convergence for random forests via generalized u-statistics. *Electronic Journal of Statistics*, *16*(1), 232–292.

Pitt, M. K., & Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, *94*(446), 590–599.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, *7*(2), 186–199.

Pötscher, B. M., & Prucha, I. (1997). *Dynamic nonlinear econometric models: Asymptotic theory*. Springer.

Schafer, J. B., Konstan, J., & Riedl, J. (1999). Recommender systems in e-commerce. In *Proceedings of the 1st acm conference on electronic commerce* (pp. 158–166).

Stuhlmüller, A., Taylor, J., & Goodman, N. (2013). Learning stochastic inverses. *Advances in neural information processing systems*, *26*.

van de Geer, S. A. (2000). *Empirical processes in M-estimation* (Vol. 6). Cambridge University Press.

van der Vaart, A. W. (2000). *Asymptotic Statistics* (Vol. 3). Cambridge University Press.

Vankov, E. R., Guindani, M., & Ensor, K. B. (2019). Filtering and estimation for a class of

stochastic volatility models with intractable likelihoods. *Bayesian Analysis*, *14*(1), 29–52.

Zhang, R., Langrené, N., Tian, Y., Zhu, Z., Klebaner, F., & Hamza, K. (2019). Dynamic portfolio optimization with liquidity cost and market impact: a simulation-and-regression approach. *Quantitative Finance*, *19*(3), 519–532.

Zhang, T., & Yu, B. (2005). Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, *33*(4), 1538–1579.

Appendix A Proofs

A.1 Proof of Theorem 1

By the triangle inequality,

$$\sup_{t\in\mathbb{N}} \left\|x_t^* - \widehat{x}_t^N\right\|_p \leq \sup_{t\in\mathbb{N}} \left\|x_t^* - x_t\right\|_p + \sup_{t\in\mathbb{N}} \left\|x_t - \widehat{x}_t^N\right\|_p, \tag{16}$$

so that boundedness of the left-hand side can be established by considering the two terms on the right-hand side separately. We start with the second term. Since $\{y_t\}_{t\in\mathbb{N}}$ is assumed to be strictly stationary, the same holds for $\widetilde{Y}_t$ for $t \geq W$ because for these indices the covariate sets satisfy the translation condition in (6). As the Monte Carlo draws are based on a strictly stationary process, it follows that for $t \geq W$ the regression functions $\widehat{f}_t^N(\cdot)$ are identically distributed (ID) with respect to the time index t. In addition, since they are assumed to be measurable functions of $\widetilde{Y}_t$, the filtered estimates $\widehat{x}_t^N = \widehat{f}_t^N(\widetilde{Y}_t)$ are ID over time (i.e., $\{\widehat{x}_t^N\}_{t=W}^{\infty}$). By norm subadditivity,

$$\sup_{t\geq W} \left\|x_t - \widehat{x}_t^N\right\|_p \leq \sup_{t\geq W}(\left\|x_t\right\|_p + \left\|\widehat{x}_t^N\right\|_p) = \left\|x_W\right\|_p + \left\|\widehat{x}_W^N\right\|_p < \infty,$$

where the equality holds by the fact that the filtered and true states are ID over time as $\{x_t\}_{t\in\mathbb{N}}$ is assumed to be strictly stationary, which implies that $\left\|x_W\right\|_p = \left\|x_1\right\|_p < \infty$, while the boundedness of $\left\|\widehat{x}_W^N\right\|_p$ holds by assumption. By similar reasoning,

$$\sup_{1\leq t\leq W-1} \left\|x_t - \widehat{x}_t^N\right\|_p \leq \sup_{1\leq t\leq W-1} (\left\|x_t\right\|_p + \left\|\widehat{x}_t^N\right\|_p) = \left\|x_1\right\|_p + \sup_{1\leq t\leq W-1} \left\|\widehat{x}_t^N\right\|_p < \infty,$$

which establishes that

$$\sup_{t\in\mathbb{N}} \left\|x_t - \widehat{x}_t^N\right\|_p < \infty.$$

We now consider the first term on the right-hand side of (16). By the choice of loss function, it follows that for $Y_t = y_{1:t}$ the optimal filter component, $x_t^* = f_t^*(y_{1:t})$, minimizes the $\mathcal{L}^p$-norm of the error over all functions of $y_{1:t}$,

$$f_t^*(y_{1:t}) \in \arg\min_f \left\|f(y_{1:t}) - x_t\right\|_p.$$

We will show that the error

$$\left\|f_t^*(y_{1:t}) - x_t\right\|_p \tag{17}$$

is non-increasing in t. Suppose by contradiction that for some $t \in \mathbb{N}$

$$\left\|f_t^*(y_{1:t}) - x_t\right\|_p < \left\|f_{t+1}^*(y_{1:t+1}) - x_{t+1}\right\|_p. \tag{18}$$

For any $t \in \mathbb{N}$, the process $\{x_{t+k}, y_{1+k:t+k}\}_{k\in\mathbb{N}_0}$ is strictly stationary. As the f_t^* are measurable functions, the same therefore holds for the process $\{\widetilde{x}_k\}_{k\in\mathbb{N}_0}$ defined via

$$\widetilde{x}_k = f_t^*(y_{1+k:t+k}).$$

It follows that

$$\begin{aligned}\left\|f_t^*(y_{2:t+1}) - x_{t+1}\right\|_p = \left\|\widetilde{x}_1 - x_{t+1}\right\|_p = \left\|\widetilde{x}_0 - x_t\right\|_p = \left\|f_t^*(y_{1:t}) - x_t\right\|_p \\ < \left\|f_{t+1}^*(y_{1:t+1}) - x_{t+1}\right\|_p,\end{aligned}$$

where the second equality follows by the strict stationarity of $\{\widetilde{x}_k, x_{t+k}\}_{k\in\mathbb{N}_0}$, and the inequality follows from (18). Since $f_t^*(y_{2:t+1})$ is also a function of $y_{1:t+1}$, the above inequality contradicts the optimality of $f_{t+1}^*(y_{1:t+1})$ as a predictor of x_{t+1}, which shows that (18) cannot hold. We thus have that the error in (17) is non-increasing in t, hence

$$\sup_{t\in\mathbb{N}} \left\|x_t^* - x_t\right\|_p = \left\|x_1^* - x_1\right\|_p \leq \left\|x_1^*\right\|_p + \left\|x_1\right\|_p < \infty,$$

where the inequality follows from norm-subadditivity, $\left\|x_1^*\right\|_p = \left\|f_1^*(y_1)\right\|_p < \infty$ as f_1^* minimizes of the corresponding norm over all functions of y_1, and $\left\|x_1\right\|_p < \infty$ by assumption. It then follows from the triangle inequality in (16) that

$$\sup_{t\in\mathbb{N}} \left\|x_t^* - \widehat{x}_t^N\right\|_p < \infty.$$

□

A.2 Proof of Theorem 2

By the triangle inequality,

$$|x_t^* - \widehat{x}_t^N| = |f_t^*(Y_t) - \widehat{f}_t^N(\widetilde{Y}_t^N)| \leq |f_t^*(Y_t) - \widehat{f}_t^N(Y_t)| + |\widehat{f}_t^N(Y_t) - \widehat{f}_t^N(\widetilde{Y}_t^N)|,$$

where, by Assumptions A2.1 and A2.2, the optimal filtered estimates $x_t^* = \mathbb{E}[x_t|Y_t]$ correspond to the Kalman filter. We focus on the first term, $|f_t^*(Y_t) - \widehat{f}_t^N(Y_t)|$, in which the conditioning set is used as covariate set. The linear regression model is correctly specified since it follows from the joint normality of x_t and Y_t that

$$x_t = \mathbb{E}[x_t|Y_t] + v_t = \sum_{y_j\in Y_t} \beta_{j,t} y_j + v_t, \qquad v_t \sim \mathrm{N}(0, \mathbb{V}\mathrm{ar}\,[x_t|Y_t]), \tag{19}$$

for coefficients $\beta_{j,t} \in \mathbb{R}^{N_x\times N_y}$ (Anderson & Moore, 1979, Sec. 3.1). The errors v_t are independent of Y_t because they are jointly normal and uncorrelated, where uncorrelatedness follows from the mean independence

$$\mathbb{E}[v_t|Y_t] = \mathbb{E}\Big[x_t - \mathbb{E}[x_t|Y_t] \;\Big|\; Y_t\Big] = 0 = \mathbb{E}[v_t].$$

Let z_i denote the vector of vertically stacked observations $y_j^{(i)}$ from the conditioning set, so that z_i' is the i-th row of the design matrix in the least squares regression. Then, since the data used in the regressions are IID with respect to the index $i = 1, \ldots, N$, all standard assumptions for consistency and $\sqrt{N}$-convergence of the least squares estimator are satisfied if the matrix $\mathbb{E}[z_i z_i']$ is non-singular (e.g., Hayashi, 2000, Proposition 2.1).[3] If $\mathbb{E}[z_i z_i']$

[3] As the errors are independent of the covariates, the asymptotic variance matrix of the least squares estimator reduces to $\mathbb{V}\mathrm{ar}\,[v_t] \cdot \mathbb{E}[z_i z_i']^{-1}$.

exists, then $\mathbb{E}[z_i z_i'] = \mathbb{V}\mathrm{ar}\,[z_i]$ given that $\mathbb{E}[z_i] = 0$, hence the required non-singularity follows from the assumption that the observations in $y_{1:T}$ are linearly independent. To establish the existence of $\mathbb{E}[z_i z_i']$, we consider the diagonal and off-diagonal elements separately. Assumption A2.1 implies that the second moments of the observations y_t (the diagonal elements of $\mathbb{E}[z_i z_i']$) are finite for $t = 1, \ldots, T$. For the off-diagonal elements of $\mathbb{E}[z_i z_i']$, assume without loss of generality that the observations y_j are univariate. Then, finiteness follows from Hölder's inequality as $\mathbb{E}\,|y_j y_k| \leq \sqrt{\mathbb{E}\,|y_j|^2} \cdot \sqrt{\mathbb{E}\,|y_k|^2}$ for any j, k. It follows that the least squares estimator is consistent for the true parameters $\beta_{j,t}$, and it is normal with convergence rate $\sqrt{N}$,

$$\sup_t \sup_j \left| \beta_{j,t} - \widehat{\beta}_{j,t} \right| = O_P(N^{-1/2}).$$

We therefore have that as $N \to \infty$,

$$\begin{aligned} \sup_t \left| x_t^* - \widehat{f}_t^N(Y_t) \right| &= \sup_t \left| \sum_{y_j \in Y_t} \left(\beta_{j,t} y_j - \widehat{\beta}_{j,t} y_j \right) \right| \leq \sup_t \sum_{y_j \in Y_t} \left| \beta_{j,t} - \widehat{\beta}_{j,t} \right| |y_j| \\ &= O_P(N^{-1/2}) \cdot O_P(1) = O_P(N^{-1/2}). \end{aligned}$$

Since the above holds for $t = 1, \ldots, T$, Lemma 1 can be applied with rate $r_N = \sqrt{N}$ to establish the desired filter convergence.

□

A.3 Proof of Lemma 1

By the triangle inequality, it holds for $t = 1, \ldots, T$ that

$$|x_t^* - \widehat{x}_t^N| = |f_t^*(Y_t) - \widehat{f}_t^N(\widetilde{Y}_t^N)| \leq |f_t^*(Y_t) - \widehat{f}_t^N(Y_t)| + |\widehat{f}_t^N(Y_t) - \widehat{f}_t^N(\widetilde{Y}_t^N)|. \tag{20}$$

The first term on the right-hand side represents the error based on an XMC filter that uses the conditioning set as covariate set. By assumption, $|f_t^*(Y_t) - \widehat{f}_t^N(Y_t)| = O_P(r_N^{-1})$. For the second term, which represents the error from use of a covariate set instead of the conditioning set, we will show that $|\widehat{f}_t^N(Y_t) - \widehat{f}_t^N(\widetilde{Y}_t^N)| = o_P(r_N^{-1})$. For $t = 1, \ldots, T$,

$$\left| \widehat{f}_t^N(Y_t) - \widehat{f}_t^N(\widetilde{Y}_t^N) \right| \leq \sup_{\widetilde{Y}_t \in \mathcal{P}(Y_t)} \left| \widehat{f}_t^N(Y_t) - \widehat{f}_t^N(\widetilde{Y}_t) \right| \cdot 1_{\{\widetilde{Y}_t^N \neq Y_t\}}.$$

Then, for any $\epsilon > 0$

$$\begin{aligned} & P\left(r_N \cdot \left| \widehat{f}_t^N(Y_t) - \widehat{f}_t^N(\widetilde{Y}_t^N) \right| > \epsilon \right) \\ \leq & P\left(r_N \cdot \sup_{\widetilde{Y}_t \in \mathcal{P}(Y_t)} \left| \widehat{f}_t^N(Y_t) - \widehat{f}_t^N(\widetilde{Y}_t) \right| \cdot 1_{\{\widetilde{Y}_t^N \neq Y_t\}} > \epsilon \right) \\ \leq & P\left(r_N \cdot \sup_{\widetilde{Y}_t \in \mathcal{P}(Y_t)} \left| \widehat{f}_t^N(Y_t) - \widehat{f}_t^N(\widetilde{Y}_t) \right| \cdot 1_{\{\widetilde{Y}_t^N \neq Y_t\}} > 0 \right) \\ \leq & P(\widetilde{Y}_t^N \neq Y_t) \to 0 \quad \text{as} \quad N \to \infty, \end{aligned}$$

where the convergence step follows from Assumption 1. It follows that the second term on the right-hand side of (20) is $o_P(r_N^{-1})$, so that by the triangle inequality,

$$|x_t^* - \widehat{x}_t^N| \leq O_P(r_N^{-1}) + o_P(r_N^{-1}) = O_P(r_N^{-1})$$

for $t = 1, \ldots, T$.

□

Appendix B Simultaneous convergence analysis

In this section we consider the case where the time series length and the number of Monte Carlo draws N simultaneously diverge. This pertains to the practical scenario in which an XMC filter based on some number of draws is applied in real time, while in parallel, another XMC filter is being estimated based on a larger number of draws. Once estimation of the latter is complete, it would replace the initial XMC filter, a process that may be repeated multiple times. In this light, note that it is often possible to use a function estimate as starting "value" in the optimization of a related estimation problem.

It will be assumed that the SSM is initialized at some time t_0 in the infinite past, $t_0 \to -\infty$, where we focus on filtering with the infinite conditioning set $Y_t = y_{t_0:t}$. In this setting, Assumption 1 appears to be less reasonable, and the same therefore holds for the application of Lemma 1. In order to obtain filter convergence, we shall employ a notion similar to that of $\mathcal{L}^p$-approximability (Pötscher & Prucha, 1997, Ch. 6.2), which generalizes the $\mathcal{L}^p$-near epoch dependence concept from Andrews (1988). These concepts are both particularly well suited to describe how the filtered estimates x_t^* are related to the near epoch of the observations $y_{t_0:t}$. In the following, $\widetilde{Y}_{t,m}$ is used to denote a covariate set containing the $m \in \mathbb{N}_0$ nearest lags of the observations y_t,

$$\widetilde{Y}_{t,m} = \{y_t, y_{t-1}, \ldots, y_{t-m}\},$$

with corresponding minimizer of the expected loss

$$x_{t,m}^* \coloneqq f_t^*(\widetilde{Y}_{t,m}) \in \arg\min_{c \in \mathbb{R}} \mathbb{E}[L(x_t - c)|\widetilde{Y}_{t,m}]. \tag{21}$$

Definition 2 (Filter $\mathcal{L}^p$-approximability)**.** The sequence $\{x_t^*\}_{t\in\mathbb{Z}}$ is *$\mathcal{L}^p$-approximable* by $\{x_{t,m}^*\}_{t\in\mathbb{Z}}$ if for some $p \geq 1$ there exists a sequence of constants $\{v_{t,m}\}$ such that

$$\left\|x_t^* - x_{t,m}^*\right\|_p = \left(\mathbb{E}\left|x_t^* - x_{t,m}^*\right|^p\right)^{1/p} \leq v_{t,m},$$

with $\sup_t v_{t,m} \to 0$ as $m \to \infty$

The above definition is essentially a special case of $\mathcal{L}^p$-approximability that is adapted to filtering, where the observations y_t are used as the basis process and the conditioning is limited to contemporaneous and lagged values of y_t (Pötscher & Prucha, 1997, Ch. 6.2). Similar to $\mathcal{L}^p$-near epoch dependence (Andrews, 1988), we explicitly consider the convergence rate $v_{t,m}$ because this will impact the convergence rate of the XMC filter. In fact, Definition 2 reduces to $\mathcal{L}^p$-near epoch dependence for $p \leq 2$ when $v_{t,m} = v_m d_t$, with

$\{d_t\}$ a summable sequence of non-negative constants, and the squared error loss is used, in which case $x^*_{t,m} = \mathbb{E}[x_t|\widetilde{Y}_{t,m}]$.

We let the XMC filter $\{\widehat{x}^N_t\}$ use $\widetilde{Y}_{t,m}$ as covariate set and consider filter convergence in $\mathcal{L}^p$, so that $\left\|x^*_t - \widehat{x}^N_t\right\|_p \to 0$ as m and N diverge. The concept of $\mathcal{L}^p$-approximability allows us to split the problem of convergence into two parts via the triangle inequality

$$\left\|x^*_t - \widehat{x}^N_t\right\|_p \leq \left\|x^*_t - x^*_{t,m}\right\|_p + \left\|x^*_{t,m} - \widehat{x}^N_t\right\|_p,$$

where the first part represents the error from using a finite covariate set $\widetilde{Y}_{t,m}$ to approximate the conditioning set, while the second part is an estimation error based on a finite number of covariates.

Below we illustrate our approach by considering convergence of the linear XMC filter to the *asymptotic* (or steady state) Kalman filter, which is essentially the limit of the Kalman filter for $t_0 \to -\infty$ or $t \to \infty$ (Anderson & Moore, 1979, Ch. 4.4). This concept applies to linear SSMs of the form (9) with time-invariant system matrices,

$$(F_t, G_t, H_t, Q_t, R_t) = (F, G, H, Q, R) \qquad \forall\ t.$$

If there exists a constant solution $\bar{\Sigma}$ to the recursion for $\Sigma_{t+1|t}$ in (10), it satisfies the following Riccati equation (Anderson & Moore, 1979, Eq. (4.4), p. 77),

$$\bar{\Sigma} = F[\bar{\Sigma} - \bar{\Sigma}H'(H\bar{\Sigma}H' + R)^{-1}H\bar{\Sigma}]F' GQG'. \tag{22}$$

In this case, the Kalman gain $K_t = K$ is time-invariant and is given by

$$K = F\bar{\Sigma}H'(H\bar{\Sigma}H' + R)^{-1}. \tag{23}$$

Substitution of the above expressions into (10) yields the asymptotic Kalman filter. We have the following convergence result.

Theorem 3 (Convergence to asymptotic Kalman filter)**.** *Let the SSM be time-invariant and linear Gaussian as in (9), and let L be the squared error loss. Suppose the states x_t are univariate with autoregressive coefficient $F \in (-1, 1)$ and assume the SSM is initialized in the infinite past ($t_0 \to -\infty$) with $\mathbb{V}\mathrm{ar}\,[x_{t_0}] < \infty$. Then the following holds:*

(a) The asymptotic Kalman filter is $\mathcal{L}^2$-approximable by $\{x^*_{t,m}\}$ with rate

$$v_{t,m} = |C_x|\frac{|A|^m}{1 - |A|}\left\|Ky_1\right\|_2, \tag{24}$$

where $A = F - KH \in (-1, 1)$ and $C_x = 1 - \bar{\Sigma}H'(H\bar{\Sigma}H' + R)^{-1}H$ are scalars with $\bar{\Sigma}$ and K given by (22) and (23), respectively,

(b) Suppose the XMC filter is linear as in (12). Then the XMC filter converges to the asymptotic Kalman filter as m and N diverge such that $m/\sqrt{N} \to 0$. Furthermore, convergence occurs at rate $\min\{\sqrt{N}/m, v^{-1}_{t,m}\}$, with $v_{t,m}$ given in (24), so that for $1 \leq p \leq 2$,

$$\sup_t \min\{\sqrt{N}/m, v^{-1}_{t,m}\}\left\|x^*_t - \widehat{x}^N_t\right\|_p = O(1).$$

Proof of Part (a). Substituting the expressions in (22) and (23) into (10) yields

$$x^*_{t+1|t} = Ax^*_{t|t-1} + Ky_t, \qquad A = (F - KH),$$

hence the asymptotic Kalman filter is given by

$$x^*_t = C_x x^*_{t|t-1} + C_y y_t,$$

with $1 \times N_y$ row vector $C_y = \bar{\Sigma}H'(H\bar{\Sigma}H' + R)^{-1}$ and scalar $C_x = 1 - C_y H$. Because the autoregressive coefficient satisfies $F \in (-1, 1)$ by Assumption A3.2, it follows that $A \in (-1, 1)$; see Anderson and Moore (1979, p. 77). The filter therefore admits the following representation,

$$x^*_t = C_x \sum_{j=0}^{\infty} A^j K y_{t-1-j} + C_y y_t = C_x \sum_{j=1}^{\infty} A^{j-1} K y_{t-j} + C_y y_t. \tag{25}$$

We have

$$\begin{aligned}
\left\|x^*_t - x^*_{t,m}\right\|_2 &= \left\|x^*_t - \mathbb{E}[x_t|\widetilde{Y}_{t,m}]\right\|_2 = \left\|x^*_t - \mathbb{E}\Big[\mathbb{E}[x_t|Y_t]\,\Big|\,\widetilde{Y}_{t,m}\Big]\right\|_2 \\
&= \left\|x^*_t - \mathbb{E}[x^*_t|\widetilde{Y}_{t,m}]\right\|_2 \leq \left\|x^*_t - \left(C_x \sum_{j=1}^{m} A^{j-1} K y_{t-j} + C_y y_t\right)\right\|_2 \\
&\leq \left\|C_x \sum_{j=m+1}^{\infty} A^{j-1} K y_{t-j}\right\|_2 \leq |C_x| \cdot \left\|\sum_{j=m+1}^{\infty} A^{j-1} K y_{t-j}\right\|_2,
\end{aligned}$$

where the first equality holds by the assumption of the squared error loss, the second by the tower property since $\widetilde{Y}_{t,m} \subset Y_t$, the third by noting that the asymptotic Kalman filter is optimal such that $x^*_t = \mathbb{E}[x_t|Y_t]$, the subsequent inequality follows by optimality of the conditional expectation as predictor in $\mathcal{L}^2$, the second inequality follows from (25), and the third from absolute homogeneity of the norm. In addition,

$$\begin{aligned}
\left\|\sum_{j=m+1}^{\infty} A^{j-1} K y_{t-j}\right\|_2^2 &= \mathbb{E}\left(\sum_{j=m+1}^{\infty} A^{j-1} K y_{t-j}\right)^2 \\
&= \mathbb{E}\left|\sum_{j=m+1}^{\infty} A^{j-1} K y_{t-j} \sum_{k=m+1}^{\infty} A^{k-1} K y_{t-k}\right| \\
&\leq \mathbb{E} \sum_{j=m+1}^{\infty} |A|^{j-1} \sum_{k=m+1}^{\infty} |A|^{k-1} |K y_{t-j} K y_{t-k}| \\
&= \sum_{j=m+1}^{\infty} |A|^{j-1} \sum_{k=m+1}^{\infty} |A|^{k-1}\, \mathbb{E}\,|K y_{t-j} K y_{t-k}| \\
&\leq \sum_{j=m+1}^{\infty} |A|^{j-1} \sum_{k=m+1}^{\infty} |A|^{k-1} \sqrt{\mathbb{E}\,|K y_{t-j}|^2\, \mathbb{E}\,|K y_{t-k}|^2} \\
&= \sum_{j=m+1}^{\infty} |A|^{j-1} \sum_{k=m+1}^{\infty} |A|^{k-1}\, \mathbb{E}\,|K y_1|^2 \\
&= \left(\frac{|A|^m}{1 - |A|}\right)^2 \mathbb{E}\,|K y_1|^2,
\end{aligned}$$

where the first inequality follows by subadditivity of the absolute value, the subsequent equality is established by Tonelli's theorem, the second inequality follows by the Cauchy-Schwarz inequality, the next equality holds by strict stationarity of $\{y_t\}$ (since $|F| < 1$), and the final expression follows from the partial geometric series with $|A| < 1$. We therefore have that

$$\left\|x_t^* - x_{t,m}^*\right\|_2 \leq |C_x| \frac{|A|^m}{1-|A|} \left\|Ky_1\right\|_2 .$$

The above expression is finite because $\left\|Ky_1\right\|_2 < \infty$ by strict stationarity of $\{y_t\}$ combined with the assumption that $\mathbb{V}\mathrm{ar}\left[x_{t_0}\right] < \infty$.

□

Proof of Part (b). By the triangle inequality and Part (a),

$$\left\|x_t^* - \widehat{x}_t^N\right\|_2 \leq \left\|x_t^* - x_{t,m}^*\right\|_2 + \left\|x_{t,m}^* - \widehat{x}_t^N\right\|_2 \leq v_{t,m} + \left\|x_{t,m}^* - \widehat{x}_t^N\right\|_2 , \tag{26}$$

where we shall focus on the last term. Because x_t and $\widetilde{Y}_{t,m}$ are jointly normal, it follows that the conditional expectation $\mathbb{E}[x_t|\widetilde{Y}_{t,m}]$ is linear so that $x_{t,m}^* = \mathbb{E}[x_t|\widetilde{Y}_{t,m}] = \sum_{j=t-m}^{t} \beta_{j,t} y_j$ for $1 \times N_y$ coefficient vectors $\beta_{j,t}$. We therefore have that

$$\left\|x_{t,m}^* - \widehat{x}_t^N\right\|_2 = \left\| \sum_{j=t-m}^{t} (\beta_{j,t} - \widehat{\beta}_{j,t}) y_j \right\|_2 . \tag{27}$$

Letting $u_{j,t} = \beta_{j,t} - \widehat{\beta}_{j,t} \in \mathbb{R}^{1 \times N_y}$, it follows that

$$\left\|(\beta_{j,t} - \widehat{\beta}_{j,t}) y_j\right\|_2^2 = \left\|u_{j,t} y_j\right\|_2^2 = \mathbb{E}\left|u_{j,t} y_j\right|^2 \leq \mathbb{E}\left(\left|u_{j,t} u_{j,t}'\right| \cdot \left|y_j' y_j\right|\right) = \mathbb{E}\left|u_{j,t} u_{j,t}'\right| \cdot \mathbb{E}\left|y_j' y_j\right|,$$

where the first inequality follows from the Cauchy-Schwarz inequality for the Euclidian inner product, while the final equality follows because $u_{j,t}$ is independent from y_j, since $\widehat{\beta}_{j,t}$ is estimated using a training sample that is independent from the prediction data, $\{y_j\}$. Furthermore, letting $u_{j,t} = (u_{1,j,t}, \ldots, u_{N_y,j,t})$ and $y_j = (y_{1,j}, \ldots, y_{N_y,j})'$, it follows that

$$\begin{aligned} &\mathbb{E}\left|u_{j,t} u_{j,t}'\right| \cdot \mathbb{E}\left|y_j' y_j\right| = \mathbb{E} \sum_{k=1}^{N_y} u_{k,j,t}^2 \, \mathbb{E} \sum_{l=1}^{N_y} y_{l,j}^2 = \sum_{k=1}^{N_y} \mathbb{E}\, u_{k,j,t}^2 \sum_{l=1}^{N_y} \mathbb{E}\, y_{l,j}^2 = \\ &\sum_{k=1}^{N_y} \mathbb{V}\mathrm{ar}\left[u_{k,j,t}\right] \sum_{l=1}^{N_y} \mathbb{E}\, y_{l,j}^2 = N_y \cdot O(N^{-1}) \cdot N_y \cdot O(1) = O(N^{-1}), \end{aligned} \tag{28}$$

where $\mathbb{E}\, u_{k,j,t}^2 = \mathbb{V}\mathrm{ar}\left[u_{k,j,t}\right]$ because $\mathbb{E}[u_{j,t}] = 0$ by unbiasedness of the least squares estimator and the variance rate $\mathbb{V}\mathrm{ar}\left[u_{k,j,t}\right] = O(N^{-1})$ is a standard result in least squares estimation (e.g., Hayashi, 2000, Proposition 2.1). In addition, $\mathbb{E}\, y_{l,j}^2 = O(1)$ because $\{y_t\}$ is strictly stationary (since $|F| < 1$) with bounded variance (as $\mathbb{V}\mathrm{ar}\left[x_{t_0}\right] < \infty$). Taking the square root to reverse squaring of the norm therefore yields

$$\left\|(\beta_{j,t} - \widehat{\beta}_{j,t}) y_j\right\|_2 = O\left(N^{-1/2}\right),$$

so that (27) combined with norm-subadditvity yields

$$\left\|x^*_{t,m} - \widehat{x}^N_t(Y_t)\right\|_2 \leq \sum_{j=t-m}^{t} \left\|(\beta_{j,t} - \widehat{\beta}_{j,t})y_j\right\|_2 = (m+1)\cdot O\left(N^{-1/2}\right) = O\left(m/\sqrt{N}\right).$$

Monotonicity of the $\mathcal{L}^p$-norm for $1 \leq p < \infty$ ensures that the rate $v_{t,m}$ and the above identity apply to all $1 \leq p \leq 2$. Because the observations and the optimal filter are both strictly stationary, the terms in the triangle inequality in (26) are the same for all t, hence

$$\sup_t \left\|x^*_t - \widehat{x}^N_t\right\|_p \to 0$$

as m and N diverge such that $m/\sqrt{N} \to 0$, and

$$\sup_t \min\{\sqrt{N}/m, v^{-1}_{t,m}\} \left\|x^*_t - \widehat{x}^N_t\right\|_p = O(1).$$

□

Remark 2. *In Theorem 3 (b) it is necessary that N diverges faster than m to guarantee that $m/\sqrt{N} \to 0$. As the third equality in (28) shows, this requirement is because $\left\|x^*_{t,m} - \widehat{x}^N_t\right\|_2^2$ is bounded by a term proportional to the sum over the variances of the individual elements from the least squares estimator, and the number of elements increases with m. Including more observations increases the number of coefficients to be estimated, which increases the variance of the predictions $\widehat{x}^N_t$. On the other hand, there is also an offsetting effect: the variance of the least squares estimator is proportional to the error variance which is given by $\mathbb{V}\mathrm{ar}[x_t|\widetilde{Y}_{t,m}]$; see (19) in which Y_t is finite. It can be shown that this variance is non-increasing in m (Anderson & Moore, 1979, p. 261). However, it follows from the Kalman recursion for $\mathbb{V}\mathrm{ar}\,[x_t|Y_t]$ in (10) that*

$$\lim_{m\to\infty} \mathbb{V}\mathrm{ar}[x_t|\widetilde{Y}_{t,m}] = \mathbb{V}\mathrm{ar}\,[x_t|Y_t] = \bar{\Sigma} - \bar{\Sigma}H'(H\bar{\Sigma}H' + R)^{-1}H\bar{\Sigma},$$

which does not depend on m. The offsetting effect is thus limited and therefore does not improve the convergence rate.

In Theorem 3, the assumption of univariate states is used because the Kalman filter does not have a separate treatment of the states. However, the assumption can be dropped by introducing additional assumptions on F and A. Moreover, the results remain valid without the assumption of normality, in which case the asymptotic Kalman filter is no longer optimal but is the best linear filter instead. This result requires the conditional expectation $x^*_{t,m} = \mathbb{E}[x_t|\widetilde{Y}_{t,m}]$ to be replaced by the corresponding linear projection as an $\mathcal{L}^2$-approximator. As with Theorem 2, the squared error loss assumption can be relaxed to allow for the absolute error loss.

In contrast to Theorem 2, the convergence rate in the above result is impacted by the use of covariate sets. In particular, the division by m in the rate $\sqrt{N}/m$ reflects the increased number of parameters that need to be estimated as m increases, which can now become arbitrarily large. The requirement that $m/\sqrt{N} \to 0$ shows that m and N should be chosen by taking the convergence rate into account. The following result provides guidelines.

Proposition 1 (Optimal m-N relation)**.** *Suppose that the XMC filter converges at rate* $\min\{\sqrt{N}/m, v_{t,m}^{-1}\}$*, with* $v_{t,m}^{-1} = O(q^m)$ *for some* $q > 1$*. Then the optimal convergence rate,* $v_{t,m}^{-1}$*, is obtained by letting*

$$N = O(m^2 q^{2m}),$$

or equivalently, by letting

$$m = O\left(W_0\big(C\sqrt{N}\big)\right)$$

for any $C > 0$*, where* W_0 *denotes the principle branch of the Lambert W function.*

Proof. The filter's convergence rate implies the optimal relation $\sqrt{N}/m \propto v_{t,m}^{-1}$, from which the optimal growth rate

$$N = O(m^2 q^{2m})$$

is immediate. For the optimal growth rate of m in terms of N, use the above relationship to define N by $\sqrt{N} \coloneqq kmq^m$, for some arbitrary constant $k > 0$. This gives

$$\sqrt{N} = km \exp\big(m \log(q)\big) = kz \exp(z)/\log(q),$$

with $z = m \log(q)$. Letting $C = \log(q)/k$ gives the equality

$$C\sqrt{N} = z \exp(z),$$

which can be solved to give $z = W_0\left(C\sqrt{N}\right)$, or equivalently,

$$m = W_0\left(C\sqrt{N}\right)/\log(q),$$

where we note that because $k > 0$ was arbitary and $q > 0$, the constant $C > 0$ is also arbitrary. □

The above result applies, for example, when the assumptions of Theorem 3 are satisfied to imply the rate $v_{t,m}$ in (24), so that $q = |A|^{-1} > 1$. Note that in this case, $m/\sqrt{N} = O(q^{-m})$, which goes to zero when $m \to \infty$, as required. It follows that by choosing m and N appropriately, the XMC filter attains the optimal convergence rate.

Appendix C Regularized covariate set convergence

In this section we provide sufficient conditions to guarantee that Assumption 1 holds when determining the regularized covariate set via out-of-sample optimization or penalization of the objective function. Throughout, we assume that $T \in \mathbb{N}$, such that the time series length is finite.

Consider the average validation loss

$$M_N^{\text{val}}(\widetilde{Y}_t) = \frac{1}{N_{\text{val}}} \sum_{i=1}^{N_{\text{val}}} L\left(x_t^{\langle i \rangle} - \widehat{f}_t^N\big(\widetilde{Y}_t^{\langle i \rangle}\big)\right), \tag{29}$$

where the superscript $i = 1, \ldots, N_{\text{val}}$ indicates cases from a separate validation sample, and $N_{\text{val}} = [c_{\text{val}} N]$ for some $c_{\text{val}} \in (0, 1)$, say, $c_{\text{val}} = 0.1$. The out-of-sample optimization procedure is then defined by

$$\widetilde{Y}_t^N \in \underset{\widetilde{Y}_t \in \mathcal{P}(Y_t)}{\arg\min}\, M_N^{\text{val}}(\widetilde{Y}_t). \tag{30}$$

For the penalized estimator, let $M_N^{\text{tr}}(\widetilde{Y}_t)$ denote the average training loss defined by analogy to (29). Then the penalization procedure we consider is given by

$$\widetilde{Y}_t^N \in \underset{\widetilde{Y}_t \in \mathcal{P}(Y_t)}{\arg\min}\, M_N^{\text{tr}}(\widetilde{Y}_t) + \pi_N(|\widetilde{Y}_t|), \tag{31}$$

with penalty term $\pi_N(\cdot)$ that is increasing in the size of the covariate set, $|\widetilde{Y}_t|$,

$$\pi_N(k+1) \geq \pi_N(k) \geq 0, \qquad k = 0, \ldots, T-1,$$

and vanishing with N,

$$\lim_{N \to \infty} \sup_k \pi_N(k) = 0.$$

Examples of such penalization functions are $\pi_N(|\widetilde{Y}_t|) = c|\widetilde{Y}_t|/\sqrt{N}$ for $c > 0$, as well as the unregularized case $\pi_N = 0$.

The regularized covariate sets defined by (30) and (31) are M-estimators, which means that consistency can be established by showing that the average loss converges uniformly over the feasible covariate sets to the deterministic limit function

$$M_\infty(\widetilde{Y}_t) = \text{plim}_{N \to \infty}\, M_N^{\text{val}}(\widetilde{Y}_t) \qquad \forall\ \widetilde{Y}_t \in \mathcal{P}(Y_t), \tag{32}$$

and that Y_t is "well separated" from the other feasible covariate sets (e.g., van der Vaart, 2000, Sec. 5.2). For both conditions it is helpful to note that as the time series length is assumed to be finite, the same holds for the number of feasible covariate sets. Well-separatedness then reduces to the condition that the conditioning set Y_t is the unique minimizer of the limit objective function M_∞. Moreover, it can be shown that *pointwise* convergence of the average loss to $M_\infty(\widetilde{Y}_t)$ over the covariate sets $\widetilde{Y}_t \in \mathcal{P}(Y_t)$ implies the required *uniform* convergence.

The above ideas are used to establish the following result on the convergence of the regularized covariate sets. Below, $\mathcal{Y} = \{y_{1:T} \,|\, p(y_{1:T}) > 0\}$ will denote the set of all realizable paths (Frühwirth-Schnatter, 1994).

Proposition 2 (Convergence of regularized covariate set)**.** *Suppose the time series $y_{1:T}$ is of finite length, and let Y_t be the unique minimizer of the limit objective function M_∞ defined in (32),*

$$M_\infty(Y_t) < M_\infty(\widetilde{Y}_t) \qquad \forall\ \widetilde{Y}_t \in \mathcal{P}(Y_t) \setminus \{Y_t\}. \tag{33}$$

Then

$$\text{plim}_{N \to \infty}\, \widetilde{Y}_t^N = Y_t \tag{34}$$

is implied by either of the following conditions.

(a) The regularized covariate set is defined by the out-of-sample optimization procedure in (30) and it holds that as $N \to \infty$,

$$\sup_{y_{1:T} \in \mathcal{Y}} |M_N^{\text{val}}(\widetilde{Y}_t) - M_\infty(\widetilde{Y}_t)| \xrightarrow{P} 0 \qquad \forall\ \widetilde{Y}_t \in \mathcal{P}(Y_t). \tag{35}$$

(b) The regularized covariate set is defined by the penalization procedure in (31) and it holds that as $N \to \infty$,

$$\sup_{y_{1:T} \in \mathcal{Y}} |M_N^{\text{tr}}(\widetilde{Y}_t) - M_\infty(\widetilde{Y}_t)| \xrightarrow{P} 0 \qquad \forall\ \widetilde{Y}_t \in \mathcal{P}(Y_t).$$

Proof. For Part (a) we note that since the time series is assumed to be finite, it also holds that $|\mathcal{P}(Y_t)| < \infty$. Suppose then that there are K covariate sets in $\mathcal{P}(Y_t)$, that is, $\widetilde{Y}_t^{[k]}$, $k = 1, \ldots, K$, and let $d_k = \sup_{y_{1:T} \in \mathcal{Y}} |M_N^{\text{val}}(\widetilde{Y}_t^{[k]}) - M_\infty(\widetilde{Y}_t^{[k]})|$. Then by (35) it holds that for any two $\epsilon, \epsilon_P > 0$, there exists $N_k \in \mathbb{N}$ such that

$$P\left(d_k > \epsilon\right) < \epsilon_P \qquad \forall\ N \geq N_k.$$

It therefore holds for all $N \geq \max_k\{N_k\}$ that

$$\begin{aligned} P\left(\sup_{\widetilde{Y}_t \in \mathcal{P}(Y_t)} \sup_{y_{1:T} \in \mathcal{Y}} |M_N^{\text{val}}(\widetilde{Y}_t) - M_\infty(\widetilde{Y}_t)| > \epsilon\right) &= P\left(\sup_k d_k > \epsilon\right) = P\left(\bigcup_{k=1}^K \{d_k > \epsilon\}\right) \\ &\leq \sum_{k=1}^K P\left(d_k > \epsilon\right) < K\epsilon_P, \end{aligned}$$

where the first inequality follows by σ-subadditivity of P. The above result shows that $M_N^{\text{val}}(\widetilde{Y}_t)$ converges uniformly in probability to $M_\infty(\widetilde{Y}_t)$ over $\mathcal{P}(Y_t)$, and the same holds for the objective function in Part (b) because $\pi_N(|\widetilde{Y}_t|)$ was assumed to converge to zero uniformly over $\mathcal{P}(Y_t)$ as $N \to \infty$. Since the minimizer Y_t is well-separated because $|\mathcal{P}(Y_t)| < \infty$, Theorem 5.7 in van der Vaart (2000) applies, which guarantees the convergence in (34) for the M-estimators defined by (30) and (31).

□

The assumption in (33) states essentially that none of the observations can be omitted without negatively impacting the predictive performance as measured by the mean loss. This will generally be the case when the states $\{x_t\}$ follow an autoregressive process, which holds for many, if not most SSMs used in practice. With this in mind, the above result implies that convergence of the regularized covariate sets can often be formally shown by demonstrating that (35) holds for $t = 1, \ldots, T$, that is, if the average loss for specific covariate sets converges uniformly over the realizable paths to the mean loss. The following corollary establishes this result for the linear XMC filter.

Corollary 1 (Regularized covariate sets convergence for linear XMC filter)**.** *Suppose the assumptions from Theorem 2 apply (apart from Assumption 1) and that the least squares*

estimates $\widehat{\beta}_{j,t}$ *from the linear XMC filter in* (12) *take values in a bounded parameter space* Ψ_t. *Assume the condition in* (33) *holds for* $t = 1, \ldots, T$ *and the limit objective function*

$$M_\infty(\widetilde{Y}_t) = \mathbb{E}\left(x_t - \sum_{y_j \in \widetilde{Y}_t} \beta_{j,t} y_j\right)^2,$$

with minimizing coefficients $\beta_{j,t} = \beta_{j,t}(\widetilde{Y}_t) \in \Psi_t$. *Then the convergence in* (34) *holds for* $t = 1, \ldots, T$ *when* $\widetilde{Y}_t^N$ *is defined via* (30) *or* (31).

Proof. The proof is immediate once we establish that (35) holds for $t = 1, \ldots, T$. This can be done by showing the summands of $M_N^{\text{val}}(\widetilde{Y}_t)$ are continuous in the parameters and bounded by an integrable function, which implies that they belong to the class of Glivenko-Cantelli functions (e.g., van der Vaart, 2000, p. 46). The summands are given by

$$\begin{aligned}
L\Big(x_t - \widehat{f}_t^N\big(\widetilde{Y}_t\big)\Big) &= \Big(x_t - \sum_{y_j \in \widetilde{Y}_t} \widehat{\beta}_{j,t} y_j\Big)^2 \\
&= x_t^2 - 2x_t \sum_{y_j \in \widetilde{Y}_t} \widehat{\beta}_{j,t} y_j + \Big(\sum_{y_j \in \widetilde{Y}_t} \widehat{\beta}_{j,t} y_j\Big)^2 \\
&\leq x_t^2 + 2|x_t| \sum_{y_j \in \widetilde{Y}_t} |\widehat{\beta}_{j,t}|\ |y_j| + \Big(\sum_{y_j \in \widetilde{Y}_t} |\widehat{\beta}_{j,t}|\ |y_j|\Big)^2 \\
&\leq x_t^2 + 2|x_t| \sum_{y_j \in \widetilde{Y}_t} |\beta_{j,t}^*|\ |y_j| + \Big(\sum_{y_j \in \widetilde{Y}_t} |\beta_{j,t}^*|\ |y_j|\Big)^2,
\end{aligned}$$

where $\beta_{j,t}^*$ denotes the vector of elementwise maxima of the absolute values that $\beta_{j,t}$ can take on in the bounded parameter space Ψ_t. Furthermore, it follows from Assumption A 2.1 that the final upper bound is integrable, which implies that (35) holds for $t = 1, \ldots, T$, and the same argument applies to $M_N^{\text{tr}}(\widetilde{Y}_t)$.

□

In the out-of-sample optimization procedure from Section 2, the optimization in (30) is in terms of a window size W, which determines the covariate sets (e.g., via (5) for filtering). In addition, the optimization is performed only at a single time point, t^*, for computational considerations. By choosing t^* to be the index of the largest conditioning set, the convergence in (34) at $t = t^*$ implies that this convergence holds for all times $t = 1, \ldots, T$, which provides a justification for the choice $t^* = T$ for filtering and forecasting.

Appendix D Additional illustrations

This section contains several additional illustrations regarding covariate set regularization and filter convergence. The illustrations are based on the Gaussian local level model in (3) applied to measurements of the annual flow volume of the Nile river taken at Aswan from 1871 to 1970. The static parameters were set to the maximum likelihood estimates $\sigma_x = 38.329$ and $\sigma_y = 122.877$, with $\mu_1 = 0$ and $\sigma_1^2 = 10^7$ to approximate diffuse initialization. More information on this application can be found in Durbin and Koopman (2012, Ch. 2).

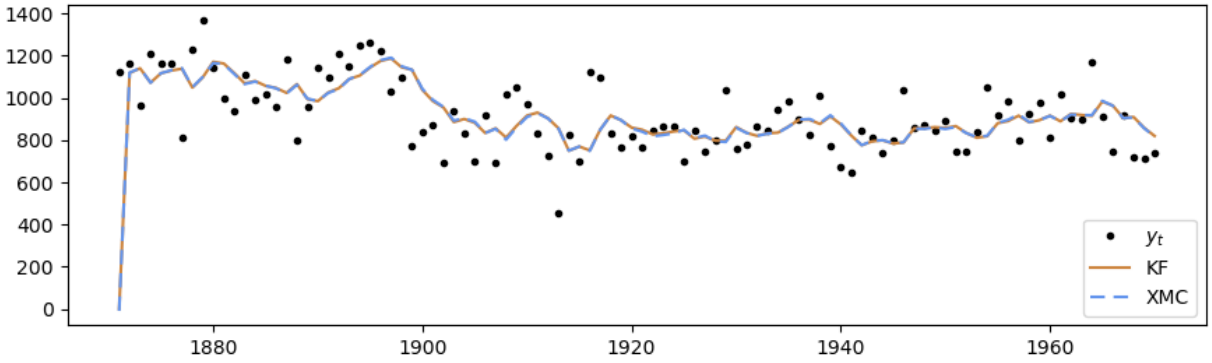

Figure 6: Forecasting analysis of the Nile data based on the local level model in (3): 1-period forecasts of the state from the Kalman filter (KF) and linear XMC filter with $N = 10^4$.

D.1 Filter convergence

The use of regression easily accommodates prediction based on other conditioning sets than the one used for filtering. For example, Figure 6 shows the 1-period forecasts of the linear XMC filter ($N = 10^4$), which coincide with the ones based on the Kalman filter. At $t = 1$, the prediction is unconditional, resulting in the value $\mu_1 = 0$, while for $t > 1$ the forecasts equal the lagged predictions from filtering, $\mathbb{E}[x_{t-1}|y_{1:t-1}]$.

D.2 Covariate set regularization

To investigate how the filter's performance is impacted by the window size, we performed a simulation study using the local level model in (3) with $T = 100$. We focus on the accuracy of the filtered state at the last time point, $\widehat{x}_T$, as a function of the window size, or equivalently, of the lower covariate set endpoint $\underset{\sim}{T} = T - W + 1$ with $\widetilde{Y}_t$ defined by (5). In particular, the RMSE of $\widehat{x}_T$ was computed for the linear XMC filter with $N \in \{10^3, 10^4\}$ based on a test sample of 10^5 paths and ten repetitions of Algorithm 1 for different seeds.

Figure 7 shows the results of the simulation study. As expected, the RMSE decreases

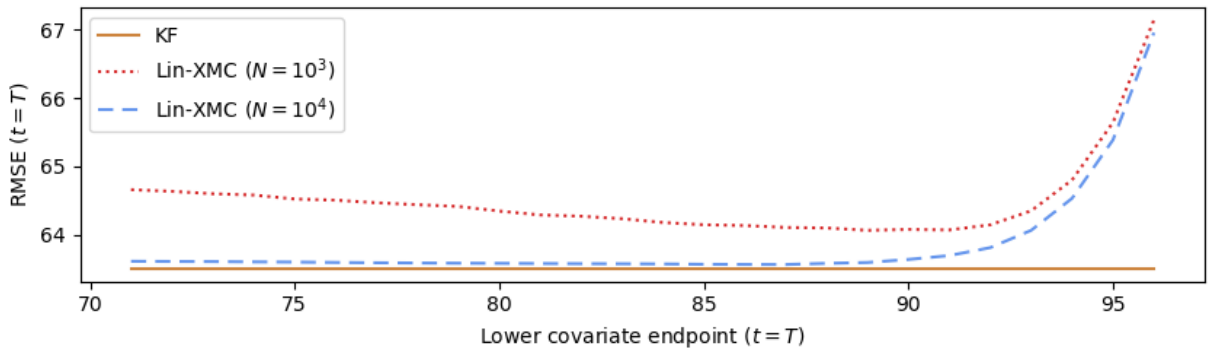

Figure 7: RMSE of $\widehat{x}_T$ in the Gaussian local level model based on the predictions for $N_{\text{test}} = 10^5$ simulated test paths. The results are shown for the Kalman filter (KF) and linear XMC filter with $N \in \{10^3, 10^4\}$ for various values of the lower covariate set endpoint, and the upper endpoint set to $T = 100$.

with N, and it is seen to be non-monotonic in the lower endpoint. Adding recent observations as covariates initially improves the performance, but after some point the increase in variance from having to estimate more parameters outweighs the decrease in the bias with respect to $\mathbb{E}[x_T|y_{1:T}]$. Regarding the bias, we note that there are clear diminishing returns to adding covariates because the observations are dependent and decreasingly informative the more remote they are from the state. The optimal covariate window is seen to vary with N, which indicates that an increase in the complexity of the regression method is warranted once more data are available. For comparison, the RMSE is also shown for the Kalman filter, which computes $\mathbb{E}[x_T|y_{1:T}]$ exactly using the recursion in (10). For $N = 10^4$, the performance of the linear XMC filter with $\underset{\sim}{T} = 87$ ($W = 14$) is almost indistinguishable from that of the optimal filter. The RMSE increases if the lower endpoint is altered, which underlines the importance of covariate set regularization.