
Munich Personal RePEc Archive

A closed-form filter for binary time series

Fasano, Augusto and Rebaudo, Giovanni and Durante, Daniele and Petrone, Sonia

2021

Online at https://mpra.ub.uni-muenchen.de/122349/
MPRA Paper No. 122349, posted 05 Nov 2024 14:51 UTC

Noname manuscript No.
(will be inserted by the editor)

A closed-form filter for binary time series

Augusto Fasano[1] · Giovanni Rebaudo[2] · Daniele Durante[3] · Sonia Petrone[3]

Abstract Non-Gaussian state-space models arise in several applications, and within this framework the binary time series setting provides a relevant example. However, unlike for Gaussian state-space models — where filtering, predictive and smoothing distributions are available in closed form — binary state-space models require approximations or sequential Monte Carlo strategies for inference and prediction. This is due to the apparent absence of conjugacy between the Gaussian states and the likelihood induced by the observation equation for the binary data. In this article we prove that the filtering, predictive and smoothing distributions in dynamic probit models with Gaussian state variables are, in fact, available and belong to a class of unified skew-normals (SUN) whose parameters can be updated recursively in time via analytical expressions. Also the key functionals of these distributions are, in principle, available, but their calculation requires the evaluation of multivariate Gaussian cumulative distribution functions. Leveraging SUN properties, we address this issue via novel Monte Carlo methods based on independent samples from the smoothing distribution, that can easily be adapted to the filtering and predictive case, thus improving state-of-the-art approximate and sequential Monte Carlo inference in small-to-moderate dimensional studies. Novel sequential Monte Carlo procedures that exploit the SUN properties are also developed to deal with online inference in high dimensions. Performance gains over competitors are outlined in a financial application.

Keywords Dynamic probit model · Kalman filter · Particle filter · State-space model · SUN.

1 Introduction

Despite the availability of several alternative approaches for dynamic inference and prediction of binary time series (MacDonald and Zucchini, 1997), state-space models are a source of constant interest due to their flexibility in accommodating a variety of representations and dependence structures via an interpretable formulation (West and Harrison, 2006; Petris et al., 2009; Durbin and Koopman, 2012). Let $\mathbf{y}_t = (y_{1t}, \ldots, y_{mt})^\intercal \in \{0;1\}^m$ be a vector of binary event data observed at time t, and denote with $\boldsymbol{\theta}_t = (\theta_{1t}, \ldots, \theta_{pt})^\intercal \in \mathbb{R}^p$ the corresponding vector of state variables. Adapting the notation in, e.g., Petris et al. (2009) to our setting, we aim to provide closed-form expressions for the filtering, predictive and smoothing distributions in the general multivariate dynamic probit model

$$p(\mathbf{y}_t \mid \boldsymbol{\theta}_t) = \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t), \quad (1)$$

$$\boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathrm{N}_p(\mathbf{0}, \mathbf{W}_t), \; t = 1 \ldots, n, \quad (2)$$

with $\boldsymbol{\theta}_0 \sim \mathrm{N}_p(\mathbf{a}_0, \mathbf{P}_0)$, and dependence structure as defined by the directed acyclic graph displayed in Fig. 1. In (1), $\Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)$ is the cumulative distribution function of a $\mathrm{N}_m(\mathbf{0}, \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)$ evaluated at $\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t$, with $\mathbf{B}_t = \mathrm{diag}(2y_{1t} - 1, \ldots, 2y_{mt} - 1)$ denoting the $m \times m$ sign matrix associated with $\mathbf{y}_t$, which defines the multivariate probit likelihood in (1).

Model (1)–(2) generalizes univariate dynamic probit models to multivariate settings, as we will clarify in equations (3)–(5). The quantities $\mathbf{F}_t, \mathbf{V}_t, \mathbf{G}_t, \mathbf{W}_t, \mathbf{a}_0$ and $\mathbf{P}_0$ denote, instead, known matrices controlling the location, scale and dependence structure in the state-space model (1)–(2). Estimation and inference for these matrices is, itself, a relevant problem which can be addressed both from a frequentist and a Bayesian perspective. Yet our focus is on providing exact results for inference on state variables and prediction of future binary events under (1)–(2). Hence, consistent with the classical Kalman filter (Kalman, 1960), we rely on known

Corresponding author: Augusto Fasano
augusto.fasano@unito.it
[1] ESOMAS Department, University of Turin, and Collegio Carlo Alberto, Turin, Italy
[2] Department of Statistics and Data Sciences, the University of Texas at Austin, Austin, United States of America
[3] Department of Decision Sciences and Institute for Data Science and Analytics, Bocconi University, Milan, Italy

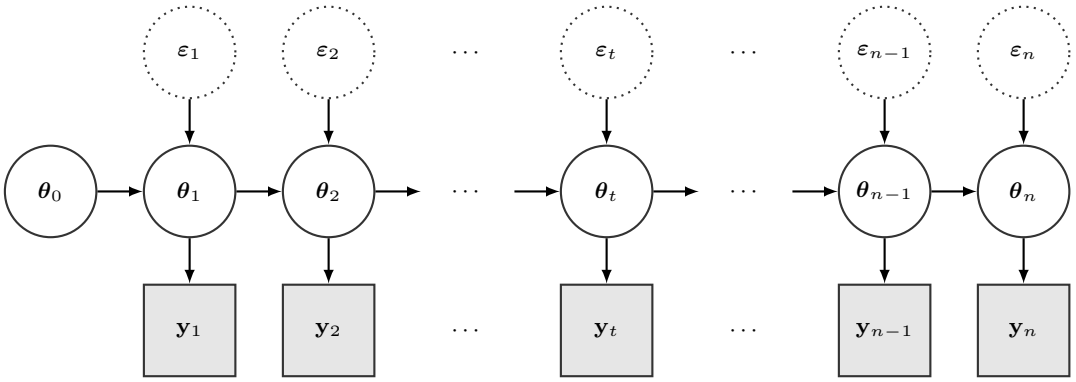

Fig. 1: Graphical representation of model (1)–(2). The dashed circles, solid circles and grey squares denote Gaussian errors, Gaussian states and observed binary data, respectively.

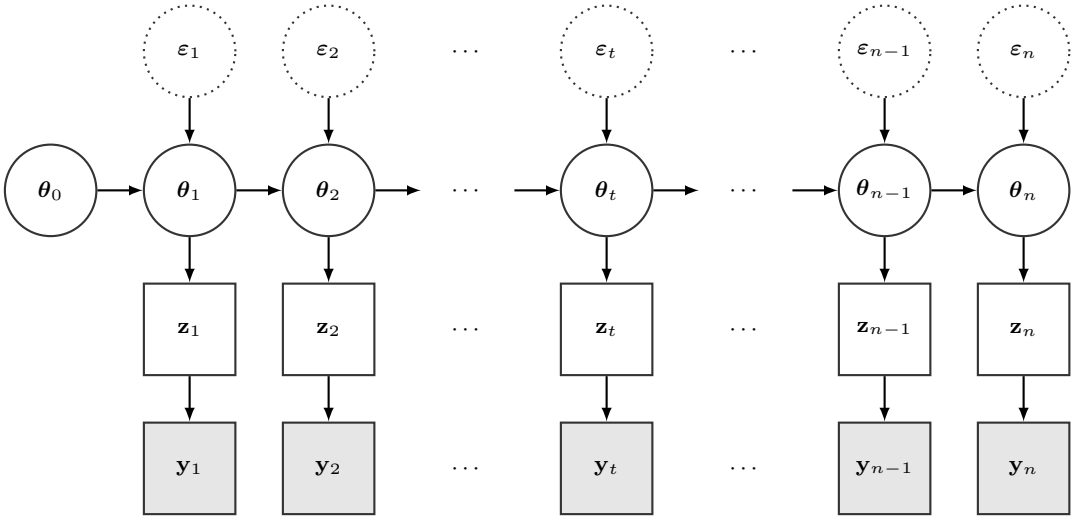

Fig. 2: Graphical representation of model (3)–(5). Dashed circles, solid circles, white squares and grey squares denote Gaussian errors, Gaussian states, latent Gaussian data and observed binary data, respectively.

system matrices $\mathbf{F}_t, \mathbf{V}_t, \mathbf{G}_t, \mathbf{W}_t, \mathbf{a}_0$ and $\mathbf{P}_0$. Nonetheless, results on marginal likelihoods, which can be used in parameter estimation, are provided in Sect. 3.2.

Model (1)–(2) provides a general representation encompassing a variety of formulations. For example, setting $\mathbf{V}_t = \mathbf{I}_m$ in (1) for each t yields a set of standard dynamic probit regressions, which include the classical univariate dynamic probit model when $m = 1$. These representations have appeared in several applications, especially within the econometrics literature, due to a direct connection between (1)–(2) and dynamic discrete choice models (Keane and Wolpin, 2009). This is due to the fact that representation (1)–(2) can be alternatively obtained via the dichotomization of an underlying state-space model for the m-variate Gaussian time series $\mathbf{z}_t = (z_{1t}, \ldots, z_{mt})^\intercal \in \mathbb{R}^m$, $t = 1, \ldots, n$, which is regarded, in econometric applications, as a set of time-varying utilities. Indeed, adapting classical results from static probit regression (Albert and Chib, 1993; Chib and Greenberg, 1998), model (1)–(2) is equivalent to

$$\begin{aligned}\mathbf{y}_t &= (y_{1t}, \ldots, y_{mt})^\intercal = \mathbb{1}(\mathbf{z}_t > \mathbf{0}) \\ &= [\mathbb{1}(z_{1t} > 0), \ldots, \mathbb{1}(z_{mt} > 0)]^\intercal, \quad t = 1, \ldots, n,\end{aligned} \tag{3}$$

with $\mathbf{z}_1, \ldots, \mathbf{z}_n$ evolving in time according to the Gaussian state-space model

$$p(\mathbf{z}_t \mid \boldsymbol{\theta}_t) = \phi_m(\mathbf{z}_t - \mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{V}_t), \tag{4}$$

$$\boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathrm{N}_p(\mathbf{0}, \mathbf{W}_t), \; t = 1 \ldots, n, \tag{5}$$

having $\boldsymbol{\theta}_0 \sim \mathrm{N}_p(\mathbf{a}_0, \mathbf{P}_0)$ and dependence structure as defined by the directed acyclic graph displayed in Fig. 2. In (4), $\phi_m(\mathbf{z}_t - \mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{V}_t)$ denotes the density function of the Gaussian $\mathrm{N}_m(\mathbf{F}_t\boldsymbol{\theta}_t, \mathbf{V}_t)$ evaluated at $\mathbf{z}_t \in \mathbb{R}^m$. To clarify the connection between (1)–(2) and (3)–(5), note that if $\tilde{\mathbf{z}}_t$ is a generic Gaussian random variable with density (4), then it holds $p(\mathbf{y}_t \mid \boldsymbol{\theta}_t) = \mathrm{pr}(\mathbf{B}_t\tilde{\mathbf{z}}_t > \mathbf{0}) = \mathrm{pr}[-\mathbf{B}_t(\tilde{\mathbf{z}}_t - \mathbf{F}_t\boldsymbol{\theta}_t) < \mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t] = \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)$, given that $-\mathbf{B}_t(\tilde{\mathbf{z}}_t - \mathbf{F}_t\boldsymbol{\theta}_t) \sim \mathrm{N}_m(\mathbf{0}, \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)$ under (4).

As is clear from model (4)–(5), if $\mathbf{z}_{1:t} = (\mathbf{z}_1^\intercal, \ldots, \mathbf{z}_t^\intercal)^\intercal$ were observed, dynamic inference on the states $\boldsymbol{\theta}_t$, for $t = 1, \ldots, n$, would be possible via direct application of the Kalman filter (Kalman, 1960). Indeed, exploiting Gaussian-Gaussian conjugacy and the conditional independence properties that are represented in Fig. 2, the filtering $p(\boldsymbol{\theta}_t \mid \mathbf{z}_{1:t})$ and predictive $p(\boldsymbol{\theta}_t \mid \mathbf{z}_{1:t-1})$ densities are also Gaussian and have parameters which can be computed recursively via simple expressions relying on the previous updates. Moreover, the smoothing density $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{z}_{1:n})$ and its marginals $p(\boldsymbol{\theta}_t \mid \mathbf{z}_{1:n})$, $t \leq n$, can also be obtained in closed form leveraging Gaussian-Gaussian conjugacy. However, in (3)–(5) only a dichotomized version $\mathbf{y}_t$ of $\mathbf{z}_t$ is available. Therefore, the filtering, predictive and smoothing densities of interest are $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ and $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$, respectively. Recalling model (1)–(2) and Bayes' rule, the calculation of these quantities proceeds by updating the Gaussian distribution for the states in (2) with the probit likelihood in (1), thereby providing conditional distributions which do not have an obvious closed form (Albert and Chib, 1993; Chib and Greenberg, 1998).

When the focus is on online inference for filtering and prediction, one solution to the above issue is to rely on approximations of model (1)–(2) which allow the implementation of standard Kalman filter updates, thus leading to approximate dynamic inference on the states via extended (Uhlmann, 1992) or unscented (Julier and Uhlmann, 1997) Kalman filters, among others. However, these approximations may lead to unreliable inference in various settings (Andrieu and Doucet, 2002). Markov chain Monte Carlo (MCMC) strategies (e.g., Carlin et al., 1992; Shephard, 1994; Soyer and Sung, 2013) address this problem but, unlike the Kalman filter, these methods are only suitable for batch learning of smoothing distributions, and tend to face mixing or scalability issues in binary settings (Johndrow et al., 2019).

Sequential Monte Carlo methods (e.g., Doucet et al., 2001) partially solve MCMC issues, and are specifically developed for online inference via particle-based representations of the states' conditional distributions, which are then propagated in time for dynamic filtering and prediction (Gordon et al., 1993; Kitagawa, 1996; Liu and Chen, 1998; Pitt and Shephard, 1999; Doucet et al., 2000; Andrieu and Doucet, 2002). These strategies provide state-of-the-art solutions in non-Gaussian state-space models, and can be also adapted to perform batch learning of the smoothing distribution; see Doucet and Johansen (2009) for a discussion on particles' degeneracy issues that may arise in such a setting. Nonetheless, sequential Monte Carlo is clearly still sub-optimal compared to the case in which $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ and $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$ are available in closed form and belong to a tractable class of known densities whose parameters can be sequentially updated via analytical expressions.

In Sect. 3, we prove that, for the dynamic multivariate probit model in (1)–(2), the quantities $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ and $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$ are unified skew-normal (SUN) densities (Arellano-Valle and Azzalini, 2006) having tractable expressions for the recursive computation of the corresponding parameters. To the best of our knowledge, such a result provides the first closed-form filter and smoother for binary time series, and facilitates improvements both in online and batch inference. As we will highlight in Sect. 2, the SUN distribution has several closure properties (Arellano-Valle and Azzalini, 2006; Azzalini and Capitanio, 2014) in addition to explicit formulas — involving the cumulative distribution function of multivariate Gaussians — for the moments (Azzalini and Bacchieri, 2010; Gupta et al., 2013) and the normalizing constant (Arellano-Valle and Azzalini, 2006). In Sect. 3, we exploit these properties to derive closed-form expressions for functionals of $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ and $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$, including, in particular, the observations' predictive density $p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$ and the marginal likelihood $p(\mathbf{y}_{1:n})$. In Sect. 4.1, we also derive an exact Monte Carlo scheme to compute generic functionals of the smoothing distribution. This routine relies on a generative representation of the SUN via linear combinations of multivariate Gaussians and truncated normals (Arellano-Valle and Azzalini, 2006), and can be also applied effectively to evaluate the functionals of filtering and predictive densities in small-to-moderate dimensions where mt is of the order of few hundreds, a common situation in routine applications.

As clarified in Sect. 4.2, the above strategies face computational bottlenecks in higher dimensions (Botev, 2017), due to challenges in computing cumulative distribution functions of multivariate Gaussians, and in sampling from multivariate truncated normals. In these contexts, we develop new sequential Monte Carlo methods that exploit SUN properties. In particular, we first prove in Sect. 4.2.1 that an optimal particle filter, in the sense of Doucet et al. (2000), can be derived analytically, thus covering a gap in the literature. This strategy is further improved in Sect. 4.2.2 via a class of partially collapsed sequential Monte Carlo methods that recursively update via lookahead strategies (Lin et al., 2013) the multivariate truncated normal component in the SUN generative additive representation, while keeping the Gaussian part exact. As outlined in an illustrative financial application in Sect. 5, this class improves approximation accuracy relative to competing methods, and includes, as a special case, the Rao–Blackwellized particle filter proposed by Andrieu and Doucet (2002). Concluding remarks can be found in Sect. 6.

2 The unified skew-normal distribution

Before deriving filtering, predictive and smoothing distributions under model (1)–(2), let us first briefly review the SUN family. Arellano-Valle and Azzalini (2006) proposed this broad class to unify different extensions (e.g., Arnold and Beaver, 2000; Arnold et al., 2002; Gupta et al., 2004; González-Farías et al., 2004) of the original multivariate skew-normal (Azzalini and Dalla Valle, 1996), whose density is obtained as the product between a multivariate Gaussian density and the cumulative distribution function of a standard normal evaluated at a value which depends on a skewness-inducing vector of parameters. Motivated by the success of this formulation and of its generalizations (Azzalini and Capitanio, 1999), Arellano-Valle and Azzalini (2006) developed a unifying representation, namely the SUN distribution. A random vector $\boldsymbol{\theta} \in \mathbb{R}^q$ has unified skew-normal distribution, $\boldsymbol{\theta} \sim \text{SUN}_{q,h}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$, if its density function $p(\boldsymbol{\theta})$ can be expressed as

$$\phi_q(\boldsymbol{\theta}-\boldsymbol{\xi}; \boldsymbol{\Omega}) \frac{\Phi_h[\boldsymbol{\gamma} + \boldsymbol{\Delta}^\intercal \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\omega}^{-1}(\boldsymbol{\theta} - \boldsymbol{\xi}); \boldsymbol{\Gamma} - \boldsymbol{\Delta}^\intercal \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\Delta}]}{\Phi_h(\boldsymbol{\gamma}; \boldsymbol{\Gamma})}, \tag{6}$$

where the covariance matrix $\boldsymbol{\Omega}$ of the Gaussian density $\phi_q(\boldsymbol{\theta} - \boldsymbol{\xi}; \boldsymbol{\Omega})$ can be decomposed as $\boldsymbol{\Omega} = \boldsymbol{\omega}\bar{\boldsymbol{\Omega}}\boldsymbol{\omega}$, that is by re-scaling the $q \times q$ correlation matrix $\bar{\boldsymbol{\Omega}}$ via the positive diagonal scale matrix $\boldsymbol{\omega} = (\boldsymbol{\Omega} \odot \mathbf{I}_q)^{1/2}$, with $\odot$ denoting the element-wise Hadamard product. In (6), the skewness-inducing mechanism is driven by the cumulative distribution function of the $\text{N}_h(\mathbf{0}, \boldsymbol{\Gamma} - \boldsymbol{\Delta}^\intercal \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\Delta})$ computed at $\boldsymbol{\gamma} + \boldsymbol{\Delta}^\intercal \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\omega}^{-1}(\boldsymbol{\theta} - \boldsymbol{\xi})$, whereas $\Phi_h(\boldsymbol{\gamma}; \boldsymbol{\Gamma})$ denotes the normalizing constant obtained by evaluating the cumulative distribution function of a $\text{N}_h(\mathbf{0}, \boldsymbol{\Gamma})$ at $\boldsymbol{\gamma}$. Arellano-Valle and Azzalini (2006) added a further identifiability condition which restricts the matrix $\boldsymbol{\Omega}^*$, with blocks $\boldsymbol{\Omega}^*_{[11]} = \boldsymbol{\Gamma}$, $\boldsymbol{\Omega}^*_{[22]} = \bar{\boldsymbol{\Omega}}$ and $\boldsymbol{\Omega}^*_{[21]} = \boldsymbol{\Omega}^{*\intercal}_{[12]} = \boldsymbol{\Delta}$, to be a full–rank correlation matrix. Note that in (6) the quantities q and h define the dimensions of the Gaussian density and cumulative distribution function, respectively. As clarified by our closed-form SUN results in Sect. 3, q defines the dimension of the states' vector, and coincides with p in the SUN filtering and predictive distributions, while it is equal to pn in the SUN smoothing distribution. On the other hand, h increases linearly with time in all the distributions of interest.

To clarify the role of the parameters in (6), we first discuss a stochastic representation of the SUN. Let $\tilde{\mathbf{z}} \in \mathbb{R}^h$ and $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^q$ characterize two random vectors jointly distributed as a $\text{N}_{h+q}(\mathbf{0}, \boldsymbol{\Omega}^*)$, then $(\boldsymbol{\xi} + \boldsymbol{\omega}\tilde{\boldsymbol{\theta}} \mid \tilde{\mathbf{z}} + \boldsymbol{\gamma} > \mathbf{0}) \sim \text{SUN}_{q,h}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$ (Arellano-Valle and Azzalini, 2006). Hence, $\boldsymbol{\xi}$ and $\boldsymbol{\omega}$ control location and scale, respectively, while $\boldsymbol{\Gamma}$, $\bar{\boldsymbol{\Omega}}$ and $\boldsymbol{\Delta}$ define the dependence within $\tilde{\mathbf{z}} \in \mathbb{R}^h$, $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^q$ and between these two vectors, respectively. Finally, $\boldsymbol{\gamma}$ controls the truncation in the partially observed Gaussian vector $\tilde{\mathbf{z}} \in \mathbb{R}^h$. The above result provides also relevant insights on our closed-form filter for the dynamic probit model (1)–(2), which will be further clarified in Sect. 3. Indeed, according to (3)–(5), the filtering, predictive and smoothing densities induced by model (1)–(2) can be also defined as $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}) = p[\boldsymbol{\theta}_t \mid \mathbb{1}(\mathbf{z}_{1:t} > \mathbf{0})]$, $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1}) = p[\boldsymbol{\theta}_t \mid \mathbb{1}(\mathbf{z}_{1:t-1} > \mathbf{0})]$ and $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n}) = p[\boldsymbol{\theta}_{1:n} \mid \mathbb{1}(\mathbf{z}_{1:n} > \mathbf{0})]$, respectively, with $(\mathbf{z}_t, \boldsymbol{\theta}_t)$ from the Gaussian state-space model (4)–(5) for $t = 1, \ldots, n$, thus highlighting the direct connection between these densities and the stochastic representation of the SUN.

An additional generative additive representation of the SUN relies on linear combinations of Gaussian and truncated normal random variables, thereby facilitating sampling from the SUN. In particular, recalling Azzalini and Capitanio (2014, Sect. 7.1.2) and Arellano-Valle and Azzalini (2006), if $\boldsymbol{\theta} \sim \text{SUN}_{q,h}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$, then

$$\boldsymbol{\theta} \overset{\text{d}}{=} \boldsymbol{\xi} + \boldsymbol{\omega}(\mathbf{U}_0 + \boldsymbol{\Delta}\boldsymbol{\Gamma}^{-1}\mathbf{U}_1), \quad \mathbf{U}_0 \perp \mathbf{U}_1, \tag{7}$$

with $\mathbf{U}_0 \sim \text{N}_q(\mathbf{0}, \bar{\boldsymbol{\Omega}} - \boldsymbol{\Delta}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}^\intercal)$ and $\mathbf{U}_1$ from a $\text{N}_h(\mathbf{0}, \boldsymbol{\Gamma})$ truncated below $-\boldsymbol{\gamma}$. As clarified in Sect. 4, this result can facilitate efficient Monte Carlo inference on complex functionals of SUN filtering, predictive and smoothing distributions under model (1)–(2), leveraging independent and identically distributed samples from such variables. Indeed, although key moments can be explicitly derived via the differentiation of the SUN moment generating function (Gupta et al., 2013; Arellano-Valle and Azzalini, 2006), such a strategy requires tedious calculations when the focus is on complex functionals. Moreover, recalling Azzalini and Bacchieri (2010) and Gupta et al. (2013), the first and second order moments further require the evaluation of h-variate Gaussian cumulative distribution functions $\Phi_h(\cdot)$, thus affecting computational tractability in large h settings (e.g., Botev, 2017). In these situations, Monte Carlo integration provides an effective solution, especially when independent samples can be generated efficiently. Therefore, we mostly focus on improved Monte Carlo inference under model (1)–(2) exploiting the SUN properties, and refer to Azzalini and Bacchieri (2010) and Gupta et al. (2013) for a closed-form expression of the expectation, variance and cumulative distribution function of SUN variables.

Before concluding this general overview, we emphasize that SUN variables are also closed under marginalization, linear combinations and conditioning (Azzalini and Capitanio, 2014). These properties facilitate the derivation of the SUN filtering, predictive and smoothing distributions under model (1)–(2).

3 Filtering, prediction and smoothing

In Sects. 3.1 and 3.2, we prove that all the distributions of direct interest admit a closed-form SUN representation. Specifically, in Sect. 3.1 we show that closed-form filters — meant here as exact updating schemes for predictive and filtering distributions based on simple recursive expressions for the associated parameters — can be obtained under model (1)–(2). Similarly, in Sect. 3.2 we derive the form of the SUN smoothing distribution and present important consequences. The associated computational methods are then discussed in Sect. 4.

3.1 Filtering and predictive distributions

To obtain the exact form of the filtering and predictive distributions under (1)–(2), let us start from $p(\boldsymbol{\theta}_1 \mid \mathbf{y}_1)$. This first quantity characterizes the initial step of the filter recursion, and its derivation within Lemma 1 provides the key intuitions to obtain the state predictive $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ and filtering $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$ densities, for any $t \geq 2$. Lemma 1 states that $p(\boldsymbol{\theta}_1 \mid \mathbf{y}_1)$ is a SUN density. In the following, consistent with the notation of Sect. 2, whenever $\boldsymbol{\Omega}$ is a $q \times q$ covariance matrix, the associated matrices $\boldsymbol{\omega}$ and $\bar{\boldsymbol{\Omega}}$ are defined as $\boldsymbol{\omega} = (\boldsymbol{\Omega} \odot \mathbf{I}_q)^{1/2}$ and $\bar{\boldsymbol{\Omega}} = \boldsymbol{\omega}^{-1}\boldsymbol{\Omega}\boldsymbol{\omega}^{-1}$, respectively. All the proofs can be found in Appendix A, and leverage conjugacy properties of the SUN in probit models. The first result on this property has been derived by Durante (2019) for static univariate Bayesian probit regression. Here, we take a substantially different perspective by focusing on online inference in both multivariate and time-varying probit models that require novel and non-straightforward extensions. As seen in Soyer and Sung (2013) and Chib and Greenberg (1998), the increased complexity of this endeavor typically motivates a separate treatment relative to the static univariate case.

Lemma 1 *Under the dynamic probit model in (1)–(2), the first-step filtering distribution is*

$$(\boldsymbol{\theta}_1 \mid \mathbf{y}_1) \sim \text{SUN}_{p,m}(\boldsymbol{\xi}_{1|1}, \boldsymbol{\Omega}_{1|1}, \boldsymbol{\Delta}_{1|1}, \boldsymbol{\gamma}_{1|1}, \boldsymbol{\Gamma}_{1|1}), \tag{8}$$

with parameters defined by the recursive equations

$$\begin{aligned}
&\boldsymbol{\xi}_{1|1} = \mathbf{G}_1\mathbf{a}_0, \quad \boldsymbol{\Omega}_{1|1} = \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1,\\
&\boldsymbol{\Delta}_{1|1} = \bar{\boldsymbol{\Omega}}_{1|1}\boldsymbol{\omega}_{1|1}\mathbf{F}_1^\intercal\mathbf{B}_1\mathbf{s}_1^{-1}, \quad \boldsymbol{\gamma}_{1|1} = \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{F}_1\boldsymbol{\xi}_{1|1},\\
&\boldsymbol{\Gamma}_{1|1} = \mathbf{s}_1^{-1}\mathbf{B}_1(\mathbf{F}_1\boldsymbol{\Omega}_{1|1}\mathbf{F}_1^\intercal + \mathbf{V}_1)\mathbf{B}_1\mathbf{s}_1^{-1},
\end{aligned}$$

where $\mathbf{s}_1 = [(\mathbf{F}_1\boldsymbol{\Omega}_{1|1}\mathbf{F}_1^\intercal + \mathbf{V}_1) \odot \mathbf{I}_m]^{1/2}$.

Hence $p(\boldsymbol{\theta}_1 \mid \mathbf{y}_1)$ is a SUN density with parameters that can be obtained via tractable arithmetic expressions applied to the quantities defining model (1)–(2). Exploiting the results in Lemma 1, the general filter updates for the multivariate dynamic probit model can be obtained by induction for $t \geq 2$ and are presented in Theorem 1.

Theorem 1 *Let* $(\boldsymbol{\theta}_{t-1}|\mathbf{y}_{1:t-1}) \sim \text{SUN}_{p,m(t-1)}(\boldsymbol{\xi}_{t-1|t-1}, \boldsymbol{\Omega}_{t-1|t-1}, \boldsymbol{\Delta}_{t-1|t-1}, \boldsymbol{\gamma}_{t-1|t-1}, \boldsymbol{\Gamma}_{t-1|t-1})$ *be the filtering distribution at time* $t-1$ *under model (1)–(2). Then, the one-step-ahead state predictive distribution at* t *is*

$$\begin{aligned}
&(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1}) \qquad\qquad (9)\\
&\sim \text{SUN}_{p,m(t-1)}(\boldsymbol{\xi}_{t|t-1}, \boldsymbol{\Omega}_{t|t-1}, \boldsymbol{\Delta}_{t|t-1}, \boldsymbol{\gamma}_{t|t-1}, \boldsymbol{\Gamma}_{t|t-1}),
\end{aligned}$$

with parameters defined by the recursive equations

$$\begin{aligned}
&\boldsymbol{\xi}_{t|t-1} = \mathbf{G}_t\boldsymbol{\xi}_{t-1|t-1}, \quad \boldsymbol{\Omega}_{t|t-1} = \mathbf{G}_t\boldsymbol{\Omega}_{t-1|t-1}\mathbf{G}_t^\intercal + \mathbf{W}_t,\\
&\boldsymbol{\Delta}_{t|t-1} = \boldsymbol{\omega}_{t|t-1}^{-1}\mathbf{G}_t\boldsymbol{\omega}_{t-1|t-1}\boldsymbol{\Delta}_{t-1|t-1},\\
&\boldsymbol{\gamma}_{t|t-1} = \boldsymbol{\gamma}_{t-1|t-1}, \qquad \boldsymbol{\Gamma}_{t|t-1} = \boldsymbol{\Gamma}_{t-1|t-1}.
\end{aligned}$$

Moreover, the filtering distribution at time t *is*

$$(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}) \sim \text{SUN}_{p,mt}(\boldsymbol{\xi}_{t|t}, \boldsymbol{\Omega}_{t|t}, \boldsymbol{\Delta}_{t|t}, \boldsymbol{\gamma}_{t|t}, \boldsymbol{\Gamma}_{t|t}), \tag{10}$$

with parameters defined by the recursive equations

$$\begin{aligned}
&\boldsymbol{\xi}_{t|t} = \boldsymbol{\xi}_{t|t-1}, \quad \boldsymbol{\Omega}_{t|t} = \boldsymbol{\Omega}_{t|t-1},\\
&\boldsymbol{\Delta}_{t|t} = [\boldsymbol{\Delta}_{t|t-1}, \bar{\boldsymbol{\Omega}}_{t|t}\boldsymbol{\omega}_{t|t}\mathbf{F}_t^\intercal\mathbf{B}_t\mathbf{s}_t^{-1}],\\
&\boldsymbol{\gamma}_{t|t} = [\boldsymbol{\gamma}_{t|t-1}^\intercal, \boldsymbol{\xi}_{t|t}^\intercal\mathbf{F}_t^\intercal\mathbf{B}_t\mathbf{s}_t^{-1}]^\intercal,
\end{aligned}$$

and $\boldsymbol{\Gamma}_{t|t}$ *is a full-rank correlation matrix having blocks* $\boldsymbol{\Gamma}_{t|t[11]} = \boldsymbol{\Gamma}_{t|t-1}$, $\boldsymbol{\Gamma}_{t|t[22]} = \mathbf{s}_t^{-1}\mathbf{B}_t(\mathbf{F}_t\boldsymbol{\Omega}_{t|t}\mathbf{F}_t^\intercal + \mathbf{V}_t)\mathbf{B}_t\mathbf{s}_t^{-1}$ *and* $\boldsymbol{\Gamma}_{t|t[21]} = \boldsymbol{\Gamma}_{t|t[12]}^\intercal = \mathbf{s}_t^{-1}\mathbf{B}_t\mathbf{F}_t\boldsymbol{\omega}_{t|t}\boldsymbol{\Delta}_{t|t-1}$, *where* $\mathbf{s}_t$ *is defined as* $\mathbf{s}_t = [(\mathbf{F}_t\boldsymbol{\Omega}_{t|t}\mathbf{F}_t^\intercal + \mathbf{V}_t) \odot \mathbf{I}_m]^{1/2}$.

As shown in Theorem 1, online prediction and filtering in the multivariate dynamic probit model (1)–(2) proceeds by iterating between equations (9) and (10) as new observations stream in with time t. Both steps are based on closed-form distributions and rely on analytical expressions for recursive updating of the corresponding parameters as a function of the previous ones, thus providing an analog of the classical Kalman filter.

We also provide closed-form expressions for the predictive density of the multivariate binary response data $\mathbf{y}_t$. Indeed, the prediction of $\mathbf{y}_t \in \{0;1\}^m$ given the data $\mathbf{y}_{1:t-1}$, is a primary goal in applications of dynamic probit models. In our setting, this task requires the derivation of the predictive density $p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$ which coincides, under (1)–(2), with $\int \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})\mathrm{d}\boldsymbol{\theta}_t$, where $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ is the state predictive density from (9). Corollary 1 shows that $p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$ has an explicit form.

Corollary 1 *Under model (1)–(2), the observation predictive density* $p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$ *is*

$$p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}) = \frac{\Phi_{mt}(\boldsymbol{\gamma}_{t|t}; \boldsymbol{\Gamma}_{t|t})}{\Phi_{m(t-1)}(\boldsymbol{\gamma}_{t|t-1}; \boldsymbol{\Gamma}_{t|t-1})}, \tag{11}$$

for every time t, *with parameters* $\boldsymbol{\gamma}_{t|t}$, $\boldsymbol{\Gamma}_{t|t}$, $\boldsymbol{\gamma}_{t|t-1}$ *and* $\boldsymbol{\Gamma}_{t|t-1}$, *defined as in Theorem 1.*

Hence, the evaluation of probabilities of future events is possible via explicit calculations after marginalizing out analytically the states with respect to their predictive density. As is clear from (11), this requires the calculation of Gaussian cumulative distribution functions whose dimension increases with t and m. Efficient evaluation of such integrals is possible for small-to-moderate t and m via recent methods (Botev, 2017), but this solution is impractical for large t and m, as seen in Table 1. In Sect. 4, we develop novel Monte Carlo strategies to address this issue and enhance scalability. This is done by exploiting Theorem 1 to improve current solutions.

3.2 Smoothing distribution

We now consider smoothing distributions. In this case, the focus is on the distribution of the entire states' sequence $\boldsymbol{\theta}_{1:n}$, or a subset of it, given all data $\mathbf{y}_{1:n}$. Theorem 2 shows that also the smoothing density $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$ belongs to the SUN family. Direct consequences of this result, involving marginal smoothing and marginal likelihoods are reported in Corollaries 2 and 3.

Before stating the result, let us first introduce the two block-diagonal matrices, $\mathbf{D}$ and $\boldsymbol{\Lambda}$, with dimensions $(mn)\times(pn)$ and $(mn)\times(mn)$ respectively, and diagonal blocks $\mathbf{D}_{[ss]} = \mathbf{B}_s\mathbf{F}_s \in \mathbb{R}^{m\times p}$ and $\boldsymbol{\Lambda}_{[ss]} = \mathbf{B}_s\mathbf{V}_s\mathbf{B}_s \in \mathbb{R}^{m\times m}$, for every time point $s = 1,\dots,n$. Moreover, let $\boldsymbol{\xi}$ and $\boldsymbol{\Omega}$ denote the mean and covariance matrix of the multivariate Gaussian distribution for $\boldsymbol{\theta}_{1:n}$ induced by the state equations. Under (2), $\boldsymbol{\xi}$ is a $pn\times 1$ column vector obtained by stacking the p-dimensional blocks $\boldsymbol{\xi}_{[s]} = \mathbb{E}(\boldsymbol{\theta}_s) = \mathbf{G}_1^s\mathbf{a}_0 \in \mathbb{R}^p$ for every $s = 1,\dots,n$, with $\mathbf{G}_1^s = \mathbf{G}_s\cdots\mathbf{G}_1$. Similarly, letting $\mathbf{G}_l^s = \mathbf{G}_s\cdots\mathbf{G}_l$, also the $(pn)\times(pn)$ covariance matrix $\boldsymbol{\Omega}$ has a block structure with $(p\times p)$-dimensional blocks $\boldsymbol{\Omega}_{[ss]} = \mathrm{var}(\boldsymbol{\theta}_s) = \mathbf{G}_1^s\mathbf{P}_0\mathbf{G}_1^{s\intercal} + \sum_{l=2}^{s}\mathbf{G}_l^s\mathbf{W}_{l-1}\mathbf{G}_l^{s\intercal} + \mathbf{W}_s$, for $s = 1,\dots,n$, and $\boldsymbol{\Omega}_{[sl]} = \boldsymbol{\Omega}_{[ls]}^{\intercal} = \mathrm{cov}(\boldsymbol{\theta}_s,\boldsymbol{\theta}_l) = \mathbf{G}_{l+1}^s\boldsymbol{\Omega}_{[ll]}$, for $s > l$.

Theorem 2 *Under model (1)–(2), the joint smoothing distribution is*

$$(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n}) \sim \mathrm{SUN}_{pn,mn}(\boldsymbol{\xi}_{1:n|n}, \boldsymbol{\Omega}_{1:n|n}, \boldsymbol{\Delta}_{1:n|n}, \boldsymbol{\gamma}_{1:n|n}, \boldsymbol{\Gamma}_{1:n|n}), \quad (12)$$

with parameters defined as

$$\boldsymbol{\xi}_{1:n|n} = \boldsymbol{\xi}, \quad \boldsymbol{\Omega}_{1:n|n} = \boldsymbol{\Omega}, \quad \boldsymbol{\Delta}_{1:n|n} = \bar{\boldsymbol{\Omega}}\boldsymbol{\omega}\mathbf{D}^{\intercal}\mathbf{s}^{-1},$$
$$\boldsymbol{\gamma}_{1:n|n} = \mathbf{s}^{-1}\mathbf{D}\boldsymbol{\xi}, \quad \boldsymbol{\Gamma}_{1:n|n} = \mathbf{s}^{-1}(\mathbf{D}\boldsymbol{\Omega}\mathbf{D}^{\intercal} + \boldsymbol{\Lambda})\mathbf{s}^{-1},$$

where $\mathbf{s} = [(\mathbf{D}\boldsymbol{\Omega}\mathbf{D}^{\intercal} + \boldsymbol{\Lambda}) \odot \mathbf{I}_{mn}]^{1/2}$.

Since the SUN is closed under marginalization and linear combinations, it follows from Theorem 2 that the smoothing distribution for any combination of states is still a SUN. In particular, direct application of the results in Azzalini and Capitanio (2014, Sect. 7.1.2) yields the marginal smoothing distribution for any state $\boldsymbol{\theta}_t$ reported in Corollary 2.

Corollary 2 *Under the model in (1)–(2), the marginal smoothing distribution at any time $t \le n$ is*

$$(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:n}) \sim \mathrm{SUN}_{p,mn}(\boldsymbol{\xi}_{t|n}, \boldsymbol{\Omega}_{t|n}, \boldsymbol{\Delta}_{t|n}, \boldsymbol{\gamma}_{t|n}, \boldsymbol{\Gamma}_{t|n}), \quad (13)$$

with parameters defined as

$$\boldsymbol{\xi}_{t|n} = \boldsymbol{\xi}_{[t]}, \quad \boldsymbol{\Omega}_{t|n} = \boldsymbol{\Omega}_{[tt]}, \quad \boldsymbol{\Delta}_{t|n} = \boldsymbol{\Delta}_{1:n|n[t]},$$
$$\boldsymbol{\gamma}_{t|n} = \boldsymbol{\gamma}_{1:n|n}, \quad \boldsymbol{\Gamma}_{t|n} = \boldsymbol{\Gamma}_{1:n|n},$$

where $\boldsymbol{\Delta}_{1:n|n[t]}$ defines the t-th block of p rows in $\boldsymbol{\Delta}_{1:n|n}$. When $t = n$, (13) gives the filtering distribution at n.

Another important consequence of Theorem 2 is the availability of a closed-form expression for the marginal likelihood $p(\mathbf{y}_{1:n})$, which is provided in Corollary 3.

Corollary 3 *Under model (1)–(2), the marginal likelihood is $p(\mathbf{y}_{1:n}) = \Phi_{mn}(\boldsymbol{\gamma}_{1:n|n}; \boldsymbol{\Gamma}_{1:n|n})$, with $\boldsymbol{\gamma}_{1:n|n}$ and $\boldsymbol{\Gamma}_{1:n|n}$ defined as in Theorem 2.*

This closed-form result is useful in several contexts, including estimation of unknown system parameters via marginal likelihood maximization, and full Bayesian inference through MCMC or variational inference methods.

4 Inference via Monte Carlo methods

As discussed in Sects. 2 and 3, inference without sampling from (9), (10) or (12) is, theoretically, possible. Indeed, since the SUN densities of the filtering, predictive and smoothing distributions can be obtained from Theorems 1–2, the main functionals of interest can be computed via closed-form expressions (Arellano-Valle and Azzalini, 2006; Azzalini and Bacchieri, 2010; Gupta et al., 2013; Azzalini and Capitanio, 2014) or by relying on numerical integration. However, these strategies require evaluations of multivariate Gaussian cumulative distribution functions, which tend to be impractical as t grows or when the focus is on complex functionals.

In such situations, Monte Carlo integration provides an accurate solution to evaluate the generic functionals $\mathbb{E}[g(\boldsymbol{\theta}_t) \mid \mathbf{y}_{1:t}]$, $\mathbb{E}[g(\boldsymbol{\theta}_t) \mid \mathbf{y}_{1:t-1}]$ and $\mathbb{E}[g(\boldsymbol{\theta}_{1:n}) \mid \mathbf{y}_{1:n}]$ for the filtering, predictive and smoothing distribution via

$$\frac{1}{R}\sum_{r=1}^{R} g(\boldsymbol{\theta}_{t|t}^{(r)}), \quad \frac{1}{R}\sum_{r=1}^{R} g(\boldsymbol{\theta}_{t|t-1}^{(r)}), \quad \frac{1}{R}\sum_{r=1}^{R} g(\boldsymbol{\theta}_{1:n|n}^{(r)}),$$

with $\boldsymbol{\theta}_{t|t}^{(r)}$, $\boldsymbol{\theta}_{t|t-1}^{(r)}$ and $\boldsymbol{\theta}_{1:n|n}^{(r)}$ sampled from $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ and $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$, respectively. For example, if the evaluation of (11) is demanding, the observations predictive density can be easily computed as $\sum_{r=1}^{R}\Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_{t|t-1}^{(r)}; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)/R$.

Algorithm 1: Independent and identically distributed sampling from $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$

[1] Sample $\mathbf{U}_{0\ 1:n|n}^{(1)}, \ldots, \mathbf{U}_{0\ 1:n|n}^{(R)}$ independently from a $\mathrm{N}_{pn}(\mathbf{0}, \bar{\boldsymbol{\Omega}}_{1:n|n} - \boldsymbol{\Delta}_{1:n|n}\boldsymbol{\Gamma}_{1:n|n}^{-1}\boldsymbol{\Delta}_{1:n|n}^{\intercal})$.
[2] Sample $\mathbf{U}_{1\ 1:n|n}^{(1)}, \ldots, \mathbf{U}_{1\ 1:n|n}^{(R)}$ independently from a $\mathrm{TN}_{mn}(\mathbf{0}, \boldsymbol{\Gamma}_{1:n|n}; \mathbb{A}_{\boldsymbol{\gamma}_{1:n|n}})$.
[3] Compute $\boldsymbol{\theta}_{1:n|n}^{(1)}, \ldots, \boldsymbol{\theta}_{1:n|n}^{(R)}$ via $\boldsymbol{\theta}_{1:n|n}^{(r)} = \boldsymbol{\xi}_{1:n|n} + \boldsymbol{\omega}_{1:n|n}(\mathbf{U}_{0\ 1:n|n}^{(r)} + \boldsymbol{\Delta}_{1:n|n}\boldsymbol{\Gamma}_{1:n|n}^{-1}\mathbf{U}_{1\ 1:n|n}^{(r)})$, for $r = 1, \ldots, R$.

To be implemented, the above approach requires an efficient strategy to sample from (9), (10) and (12). Exploiting the SUN properties and recent results in Botev (2017), an algorithm to draw independent and identically distributed samples from the exact SUN distributions in (9), (10) and (12) is developed within Sect. 4.1. As illustrated in Sect. 5, such a technique is more accurate than state-of-the-art methods and can be efficiently implemented in small-to-moderate dimensional time series. In Sect. 4.2 we develop, instead, novel sequential Monte Carlo schemes that allow scalable online learning in high dimensional settings and have optimality properties (Doucet et al., 2000) which shed new light also on existing strategies (e.g, Andrieu and Doucet, 2002).

4.1 Independent identically distributed sampling

As discussed in Sect. 1, MCMC and sequential Monte Carlo methods to sample from $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ and $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$ are available. However, the commonly recommended practice, if feasible, is to rely on independent and identically distributed (i.i.d.) samples. Here, we derive a Monte Carlo algorithm to address this goal with a main focus on the smoothing distribution, and discuss direct modifications to allow sampling also in the filtering and predictive case. Indeed, Monte Carlo inference is particularly suitable for batch settings, although, as discussed later, the proposed routine is practically useful also when the focus is on filtering and predictive distributions, since i.i.d. samples are simulated rapidly, for each t, in small-to-moderate dimensions.

Exploiting the closed-form expression of the smoothing distribution in Theorem 2, and the additive representation (7) of the SUN, i.i.d. samples for $\boldsymbol{\theta}_{1:n|n}$ from the smoothing distribution (12) can be obtained via a linear combination between independent samples from (pn)-variate Gaussians and (mn)-variate truncated normals. Algorithm 1 provides the detailed pseudo-code for this novel strategy, whose outputs are i.i.d. samples from the joint smoothing density $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$. Here, the most computationally intensive step is the sampling from $\mathrm{TN}_{mn}(\mathbf{0}, \boldsymbol{\Gamma}_{1:n|n}; \mathbb{A}_{\boldsymbol{\gamma}_{1:n|n}})$, which denotes the multivariate Gaussian distribution $\mathrm{N}_{mn}(\mathbf{0}, \boldsymbol{\Gamma}_{1:n|n})$ truncated to the region $\mathbb{A}_{\boldsymbol{\gamma}_{1:n|n}} = \{\mathbf{u}_1 \in \mathbb{R}^{mn} : \mathbf{u}_1 + \boldsymbol{\gamma}_{1:n|n} > 0\}$. In fact, although efficient Hamiltonian Monte Carlo solutions are available (Pakman and Paninski, 2014), these strategies do not provide independent samples. More recently, an accept-reject method based on minimax tilting has been proposed by Botev (2017) to improve the acceptance rate of classical rejection sampling, while avoiding mixing issues of MCMC. This routine is available in the `R` library `TruncatedNormal` and allows efficient sampling from multivariate truncated normals with a dimension of few hundreds, thereby providing effective Monte Carlo inference via Algorithm 1 in small-to-moderate dimensional time series where mn is of the order of few hundreds.

Clearly, the availability of an i.i.d. sampling scheme from the smoothing distribution overcomes the need of MCMC methods and particle smoothers. The first set of strategies usually faces mixing or time-inefficiency issues, especially in imbalanced binary settings (Johndrow et al., 2019), whereas the second class of routines tends to be computationally intensive and subject to particles degeneracy (Doucet and Johansen, 2009).

When the focus is on Monte Carlo inference for the marginal smoothing density $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:n})$ at a specific time t, Algorithm 1 requires minor adaptations relying again on the additive representation of the SUN in (13), under similar arguments considered for the joint smoothing setting. This latter routine can be also used to sample from the filtering distribution in (10) by applying such a scheme with $n = t$ to obtain i.i.d. samples for $\boldsymbol{\theta}_{t|t}$ from $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$. Leveraging realizations from the filtering distribution at time $t-1$, i.i.d. samples for $\boldsymbol{\theta}_{t|t-1}$ from the predictive density $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$, can be simply obtained via the direct application of (2) which provides samples for $\boldsymbol{\theta}_{t|t-1}$ from $\mathrm{N}_p(\mathbf{G}_t\boldsymbol{\theta}_{t-1|t-1}, \mathbf{W}_t)$. As a result, efficient Monte Carlo inference in small-to-moderate dimensional dynamic probit models is possible also for filtering and predictive distributions.

4.2 Sequential Monte Carlo sampling

When the dimension of the dynamic probit model (1)–(2) grows, sampling from multivariate truncated Gaussians in Algorithm 1 might yield computational bottlenecks (Botev, 2017). This is particularly likely to occur in series monitored on a fine time grid. Indeed, in several applications, the number of time series m is typi-

cally small, whereas the length of the time window can be large. To address this issue and allow scalable online filtering and prediction also in large t settings, we first derive in Sect. 4.2.1 a particle filter which exploits the SUN results to obtain optimality properties, in the sense of Doucet et al. (2000). Despite covering a gap in the literature on dynamic probit models, as clarified in Sects. 4.2.1 and 4.2.2, such a strategy is amenable to further improvements since it induces unnecessary autocorrelation in the Gaussian part of the SUN generative representation. Motivated by this consideration and by the additive structure of the SUN filtering distribution, we further develop in Sect. 4.2.2 a partially collapsed sequential Monte Carlo procedure which recursively samples via lookahead methods (Lin et al., 2013) only the multivariate truncated normal term in the SUN additive representation, while keeping the Gaussian component exact. As outlined in Sect. 4.2.2, such a broad class of partially collapsed lookahead particle filters comprises, as a special case, the Rao–Blackwellized particle filter developed by Andrieu and Doucet (2002). This provides novel theoretical support to the notable performance of such a strategy, which was originally motivated, in the context of dynamic probit models, also by the lack of a closed-form optimal particle filter for the states.

4.2.1 "Optimal" particle filter

The first proposed strategy belongs to the class of sequential importance sampling-resampling (SISR) algorithms which provide default strategies in particle filtering (e.g., Doucet et al., 2000, 2001; Durbin and Koopman, 2012). For each time t, these routines sample R trajectories for $\boldsymbol{\theta}_{1:t|t}$ from $p(\boldsymbol{\theta}_{1:t} \mid \mathbf{y}_{1:t})$, known as *particles*, conditioned on those produced at $t-1$, by iterating, in time, between the two steps summarized below.

1. Sampling. Let $\boldsymbol{\theta}^{(1)}_{1:t-1|t-1}, \ldots, \boldsymbol{\theta}^{(R)}_{1:t-1|t-1}$ be the trajectories of the particles at time $t-1$, and denote with $\pi(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{1:t-1}, \mathbf{y}_{1:t})$ the proposal. Then, for $r = 1, \ldots, R$

[1.a] Sample $\bar{\boldsymbol{\theta}}^{(r)}_{t|t}$ from $\pi(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}^{(r)}_{1:t-1|t-1}, \mathbf{y}_{1:t})$ and set

$$\bar{\boldsymbol{\theta}}^{(r)}_{1:t|t} = (\boldsymbol{\theta}^{(r)\intercal}_{1:t-1|t-1}, \bar{\boldsymbol{\theta}}^{(r)\intercal}_{t|t})^\intercal.$$

[1.b] Set $w^{(r)}_t = w_t(\bar{\boldsymbol{\theta}}^{(r)}_{1:t|t})$, with

$$w_t(\bar{\boldsymbol{\theta}}^{(r)}_{1:t|t}) \propto \frac{p(\mathbf{y}_t \mid \bar{\boldsymbol{\theta}}^{(r)}_{t|t}) p(\bar{\boldsymbol{\theta}}^{(r)}_{t|t} \mid \boldsymbol{\theta}^{(r)}_{t-1|t-1})}{\pi(\bar{\boldsymbol{\theta}}^{(r)}_{t|t} \mid \boldsymbol{\theta}^{(r)}_{1:t-1|t-1}, \mathbf{y}_{1:t})},$$

and normalize the weights, so that their sum is 1.

2. Resampling. For $r = 1, \ldots, R$, sample updated particles' trajectories $\boldsymbol{\theta}^{(1)}_{1:t|t}, \ldots, \boldsymbol{\theta}^{(R)}_{1:t|t}$ from $\sum_{r=1}^R w^{(r)}_t \delta_{\bar{\boldsymbol{\theta}}^{(r)}_{1:t|t}}$.

From these particles, functionals of the filtering density $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$ can be computed using the terminal values $\boldsymbol{\theta}_{t|t}$ of each particles' trajectory for $\boldsymbol{\theta}_{1:t|t}$. Note that in point [1.a] we have presented the general formulation of SISR, where the importance density $\pi(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{1:t-1}, \mathbf{y}_{1:t})$ can, in principle, depend on the whole trajectory $\boldsymbol{\theta}_{1:t-1}$ (Durbin and Koopman, 2012, Sect. 12.3).

As is clear from the above steps, the performance of SISR relies on the choice of $\pi(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{1:t-1}, \mathbf{y}_{1:t})$. Such a density should allow tractable sampling along with efficient evaluation of the importance weights, and should be also carefully specified to propose effective candidate samples. Recalling Doucet et al. (2000), the optimal proposal is $\pi(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{1:t-1}, \mathbf{y}_{1:t}) = p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{y}_t)$, with importance weights $w_t \propto p(\mathbf{y}_t \mid \boldsymbol{\theta}_{t-1})$. Indeed, conditioned on $\boldsymbol{\theta}_{1:t-1|t-1}$ and $\mathbf{y}_{1:t}$, this choice minimizes the variance of the weights, thus limiting degeneracy issues and improving mixing. Unfortunately, in several dynamic models, tractable sampling from $p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{y}_t)$ and the direct evaluation of $p(\mathbf{y}_t \mid \boldsymbol{\theta}_{t-1})$ is not possible (Doucet et al., 2000). As outlined in Corollary 4, this is not the case for dynamic probit models. In particular, by leveraging the proof of Theorem 1 and the closure properties of the SUN, sampling from $p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{y}_t)$ is straightforward and $p(\mathbf{y}_t \mid \boldsymbol{\theta}_{t-1})$ has a simple form.

Corollary 4 *For every time $t = 1, \ldots, n$, the optimal importance distribution under model (1)–(2) is*

$$(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{y}_t) \sim \text{SUN}_{p,m}(\boldsymbol{\xi}_{t|t,t-1}, \boldsymbol{\Omega}_{t|t,t-1}, \boldsymbol{\Delta}_{t|t,t-1}, \boldsymbol{\gamma}_{t|t,t-1}, \boldsymbol{\Gamma}_{t|t,t-1}), \tag{14}$$

whereas the importance weights are

$$p(\mathbf{y}_t \mid \boldsymbol{\theta}_{t-1}) = \Phi_m(\boldsymbol{\gamma}_{t|t,t-1}; \boldsymbol{\Gamma}_{t|t,t-1}), \tag{15}$$

with parameters defined by the recursive equations

$$\begin{aligned}
&\boldsymbol{\xi}_{t|t,t-1} = \mathbf{G}_t \boldsymbol{\theta}_{t-1}, \quad \boldsymbol{\Omega}_{t|t,t-1} = \mathbf{W}_t, \\
&\boldsymbol{\Delta}_{t|t,t-1} = \bar{\boldsymbol{\Omega}}_{t|t,t-1} \boldsymbol{\omega}_{t|t,t-1} \mathbf{F}_t^\intercal \mathbf{B}_t \mathbf{c}_t^{-1}, \\
&\boldsymbol{\gamma}_{t|t,t-1} = \mathbf{c}_t^{-1} \mathbf{B}_t \mathbf{F}_t \boldsymbol{\xi}_{t|t,t-1}, \\
&\boldsymbol{\Gamma}_{t|t,t-1} = \mathbf{c}_t^{-1} \mathbf{B}_t \left(\mathbf{F}_t \boldsymbol{\Omega}_{t|t,t-1} \mathbf{F}_t^\intercal + \mathbf{V}_t \right) \mathbf{B}_t \mathbf{c}_t^{-1},
\end{aligned}$$

where $\mathbf{c}_t = \left[(\mathbf{F}_t \boldsymbol{\Omega}_{t|t,t-1} \mathbf{F}_t^\intercal + \mathbf{V}_t) \odot \mathbf{I}_m \right]^{1/2}$.

As clarified in Corollary 4, the weights $p(\mathbf{y}_t \mid \boldsymbol{\theta}_{t-1})$ for the generated trajectories are available analytically in (15) and do not depend on the sampled values of the particle at time t. This allows the implementation of the more efficient auxiliary particle filter (AUF) (Pitt and Shephard, 1999) by simply reversing the order of the sampling and resampling steps, thereby obtaining a performance gain (Andrieu and Doucet, 2002). Algorithm 2 illustrates the pseudo-code of the proposed "optimal" auxiliary filter, which exploits the additive representation of the SUN and Corollary 4. Note that, unlike

Algorithm 2: "Optimal" particle filter to sample from $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, for $t = 1, \dots, n$ [AUF version]

for t *from* 1 *to* n **do**

[1] Compute the weights $w_t^{(r)} = p(\mathbf{y}_t \mid \boldsymbol{\theta}_{t-1} = \boldsymbol{\theta}_{t-1|t-1}^{(r)})$ for $r = 1, \dots, R$, by applying equation (15).

[2] Resample updated particles $\bar{\boldsymbol{\theta}}_{t-1|t-1}^{(1)}, \dots, \bar{\boldsymbol{\theta}}_{t-1|t-1}^{(R)}$ from $\sum_{r=1}^{R} w_t^{(r)} \delta_{\boldsymbol{\theta}_{t-1|t-1}^{(r)}}$.

for r *from* 1 *to* R **do**

[3] Set $\boldsymbol{\xi}_{t|t,t-1}^{(r)} = \mathbf{G}_t \bar{\boldsymbol{\theta}}_{t-1|t-1}^{(r)}$ and $\boldsymbol{\gamma}_{t|t,t-1}^{(r)} = \mathbf{c}_t^{-1} \mathbf{B}_t \mathbf{F}_t \boldsymbol{\xi}_{t|t,t-1}^{(r)}$. Then, simulate $\boldsymbol{\theta}_{t|t}^{(r)}$ from (14), as follows:

[3.1] Sample $\mathbf{U}_{0\ t|t}^{(r)}$ from a $\mathrm{N}_p(\mathbf{0}, \bar{\boldsymbol{\Omega}}_{t|t,t-1} - \boldsymbol{\Delta}_{t|t,t-1} \boldsymbol{\Gamma}_{t|t,t-1}^{-1} \boldsymbol{\Delta}_{t|t,t-1}^{\intercal})$.

[3.2] Sample $\mathbf{U}_{1\ t|t}^{(r)}$ from a $\mathrm{TN}_m(\mathbf{0}, \boldsymbol{\Gamma}_{t|t,t-1}; \mathbb{A}_{\boldsymbol{\gamma}_{t|t,t-1}^{(r)}})$.

[3.3] Compute $\boldsymbol{\theta}_{t|t}^{(r)} = \boldsymbol{\xi}_{t|t,t-1}^{(r)} + \boldsymbol{\omega}_{t|t,t-1} (\mathbf{U}_{0\ t|t}^{(r)} + \boldsymbol{\Delta}_{t|t,t-1} \boldsymbol{\Gamma}_{t|t,t-1}^{-1} \mathbf{U}_{1\ t|t}^{(r)})$.

for Algorithm 1, such a sequential sampling strategy requires to sample at each step from a truncated normal whose dimension does not depend on t, thus facilitating scalable sequential inference in large t studies. Samples from the predictive distribution can be obtained from those of the filtering as discussed in Sect. 4.1.

Despite having optimality properties, a close inspection of Algorithm 2 shows that the states' particles at $t-1$ affect both the Gaussian component, via $\boldsymbol{\xi}_{t|t,t-1}$, and the truncated normal term, via $\boldsymbol{\gamma}_{t|t,t-1}$, in the SUN additive representation of $(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$. Although the autocorrelation in the multivariate truncated normal samples is justified by the computational intractability of this variable in high dimensions, inducing serial dependence also in the Gaussian terms seems unnecessary, as these quantities are tractable and their dimension does not depend on t; see Theorem 1. This suggests that a strategy which sequentially updates only the truncated normal term, while maintaining the Gaussian part exact, could further improve the performance of Algorithm 2. This new particle filter is derived in Sect. 4.2.2, inheriting also lookahead ideas (Lin et al., 2013).

4.2.2 Partially collapsed lookahead particle filter

As anticipated within Sect. 4.2, the most computationally intensive step to draw i.i.d. samples from the filtering distribution is sampling from the multivariate truncated normal $\mathbf{U}_{1\ 1:t|t} \sim \mathrm{TN}_{mt}(\mathbf{0}, \boldsymbol{\Gamma}_{1:t|t}; \mathbb{A}_{\boldsymbol{\gamma}_{1:t|t}})$ in Algorithm 1. Here, we present a class of procedures to sequentially generate these samples, which are then combined with realizations from the exact Gaussian component in the SUN additive representation, thus producing samples from the filtering distribution. With this goal in mind, define the region $\mathbb{A}_{\mathbf{y}_{s:t}} = \{\mathbf{z} \in \mathbb{R}^{m(t-s+1)} : (2\mathbf{y}_{s:t} - \mathbf{1}) \odot \mathbf{z} > \mathbf{0}\}$ for every $s = 1, \dots, t$, and let $\mathbf{V}_{1:t}$ be the $(mt) \times (mt)$ block-diagonal matrix having blocks $\mathbf{V}_{[ss]} = \mathbf{V}_s$, for $s = 1, \dots, t$. Moreover, denote with $\mathbf{B}_{s:t}$ and $\mathbf{F}_{s:t}$ two block-diagonal matrices of dimension $[m(t-s+1)] \times [m(t-s+1)]$ and $[m(t-s+1)] \times [p(t-s+1)]$, respectively, and diagonal blocks $\mathbf{B}_{s:t[ll]} = \mathbf{B}_{s+l-1}$ and $\mathbf{F}_{s:t[ll]} = \mathbf{F}_{s+l-1}$ for $l = 1, \dots, t-s+1$. Exploiting this notation and adapting results in Sect. 3.2 to the case $n = t$, it follows from standard properties of multivariate truncated normals (Horrace, 2005) that

$$\mathbf{U}_{1\ 1:t|t} \overset{\mathrm{d}}{=} -\boldsymbol{\gamma}_{1:t|t} + \mathbf{s}_{1:t|t}^{-1} \mathbf{B}_{1:t} \mathbf{z}_{1:t|t}, \tag{16}$$

with $\mathbf{z}_{1:t|t} \sim \mathrm{TN}_{mt}(\mathbf{F}_{1:t} \boldsymbol{\xi}_{1:t|t}, \mathbf{F}_{1:t} \boldsymbol{\Omega}_{1:t|t} \mathbf{F}_{1:t}^{\intercal} + \mathbf{V}_{1:t}; \mathbb{A}_{\mathbf{y}_{1:t}})$ and $\mathbf{s}_{1:t|t} = [(\mathbf{D} \boldsymbol{\Omega}_{1:t|t} \mathbf{D}^{\intercal} + \boldsymbol{\Lambda}) \odot \mathbf{I}_{mt}]^{1/2}$, where $\mathbf{D}$ and $\boldsymbol{\Lambda}$ are defined as in Sect. 3.2, setting $n = t$. Note that the multivariate truncated normal distribution for $\mathbf{z}_{1:t|t}$ actually coincides with the conditional distribution of $\mathbf{z}_{1:t}$ given $\mathbf{y}_{1:t}$ under model (3)–(5). Indeed, by marginalizing out $\boldsymbol{\theta}_{1:t}$ in $p(\mathbf{z}_{1:t} \mid \boldsymbol{\theta}_{1:t}) = \prod_{s=1}^{t} \phi_m(\mathbf{z}_s - \mathbf{F}_s \boldsymbol{\theta}_s; \mathbf{V}_s) = \phi_{mt}(\mathbf{z}_{1:t} - \mathbf{F}_{1:t} \boldsymbol{\theta}_{1:t}; \mathbf{V}_{1:t})$ with respect to its multivariate normal distribution derived in the proof of Theorem 2, we have $p(\mathbf{z}_{1:t}) = \phi_{mt}(\mathbf{z}_{1:t} - \mathbf{F}_{1:t} \boldsymbol{\xi}_{1:t|t}; \mathbf{F}_{1:t} \boldsymbol{\Omega}_{1:t|t} \mathbf{F}_{1:t}^{\intercal} + \mathbf{V}_{1:t})$ and, as a direct consequence, we obtain

$$\begin{aligned} p(\mathbf{z}_{1:t} \mid \mathbf{y}_{1:t}) &\propto p(\mathbf{z}_{1:t}) p(\mathbf{y}_{1:t} \mid \mathbf{z}_{1:t}), \\ &\propto p(\mathbf{z}_{1:t}) \mathbb{1}[(2\mathbf{y}_{1:t} - \mathbf{1}) \odot \mathbf{z}_{1:t} > \mathbf{0}], \end{aligned}$$

which is the kernel of a $\mathrm{TN}_{mt}(\mathbf{F}_{1:t} \boldsymbol{\xi}_{1:t|t}, \mathbf{F}_{1:t} \boldsymbol{\Omega}_{1:t|t} \mathbf{F}_{1:t}^{\intercal} + \mathbf{V}_{1:t}; \mathbb{A}_{\mathbf{y}_{1:t}})$ density.

The above analytical derivations clarify that in order to sample recursively from $\mathbf{U}_{1\ 1:t|t}$ it is sufficient to apply equation (16) to sequential realizations of $\mathbf{z}_{1:t|t}$ from the joint conditional density $p(\mathbf{z}_{1:t} \mid \mathbf{y}_{1:t})$, induced by model (3)–(5), after collapsing out $\boldsymbol{\theta}_{1:t}$. While basic SISR algorithms for $p(\mathbf{z}_{1:t} \mid \mathbf{y}_{1:t})$, combined with the exact sampling from the Gaussian component $\mathbf{U}_{0\ t|t}$, are expected to yield an improved performance relative to the particle filter developed in Sect. 4.2.1, here we adapt an even broader class of lookahead particle filters (Lin et al., 2013) — which includes the basic SISR as a special case. To introduce the general lookahead idea note that $p(\mathbf{z}_{1:t} \mid \mathbf{y}_{1:t}) = p(\mathbf{z}_{t-k+1:t} \mid \mathbf{z}_{1:t-k}, \mathbf{y}_{1:t}) p(\mathbf{z}_{1:t-k} \mid \mathbf{y}_{1:t})$, where k is a pre-specified delay offset. Moreover, as a direct consequence of the dependence structure displayed in Fig. 2, we also have that $p(\mathbf{z}_{t-k+1:t} \mid \mathbf{z}_{1:t-k}, \mathbf{y}_{1:t}) =$

$p(\mathbf{z}_{t-k+1:t} \mid \mathbf{z}_{1:t-k}, \mathbf{y}_{t-k+1:t})$ for any generic k. Hence, to sequentially generate realizations of $\mathbf{z}_{1:t|t}$ from $p(\mathbf{z}_{1:t} \mid \mathbf{y}_{1:t})$, we can first sample $\mathbf{z}_{1:t-k|t}$ from $p(\mathbf{z}_{1:t-k} \mid \mathbf{y}_{1:t})$ by extending, via SISR, the trajectory $\mathbf{z}_{1:t-k-1|t-1}$ with optimal proposal $p(\mathbf{z}_{t-k} \mid \mathbf{z}_{1:t-k-1} = \mathbf{z}_{1:t-k-1|t-1}, \mathbf{y}_{t-k:t})$, and then draw the last k terms in $\mathbf{z}_{1:t|t}$ from $p(\mathbf{z}_{t-k+1:t} \mid \mathbf{z}_{1:t-k} = \mathbf{z}_{1:t-k|t}, \mathbf{y}_{t-k+1:t})$. Note that when $k = 0$ this final operation is not necessary, and the particles' updating in the first step reduces to basic SISR. Values of k in $\{1; \ldots; n-1\}$ induce, instead, a lookahead structure in which at the current time t the optimal proposal for the delayed particles leverages information of response data $\mathbf{y}_{t-k:t}$ that are not only contemporaneous to $\mathbf{z}_{t-k}$, i.e., $\mathbf{y}_{t-k}$, but also *future*, namely $\mathbf{y}_{t-k+1}, \ldots, \mathbf{y}_t$. In this way, the samples from the sub-trajectory $\mathbf{z}_{1:t-k|t}$ of $\mathbf{z}_{1:t|t}$ at time t are more compatible with the sampling density $p(\mathbf{z}_{1:t} \mid \mathbf{y}_{1:t})$ of interest and hence, when completed with the last k terms drawn from $p(\mathbf{z}_{t-k+1:t} \mid \mathbf{z}_{1:t-k} = \mathbf{z}_{1:t-k|t}, \mathbf{y}_{t-k+1:t})$, produce a sequential sampling scheme from $p(\mathbf{z}_{1:t} \mid \mathbf{y}_{1:t})$ with improved mixing and reduced degeneracy issues relative to basic SISR. Although the magnitude of such gains clearly grows with k, as illustrated in Sect. 5, setting $k = 1$ already provides empirical evidence of improved performance relative to basic SISR, without major computational costs.

To implement the aforementioned strategy it is first necessary to ensure that the lookahead proposal belongs to a class of random variables which allow efficient sampling, while having a tractable closed-form expression for the associated importance weights. Proposition 1 shows that this is the case under model (3)–(5).

Proposition 1 *Under the augmented model in (3)–(5), the lookahead proposal mentioned above has the form*

$$p(\mathbf{z}_{t-k} \mid \mathbf{z}_{1:t-k-1}, \mathbf{y}_{t-k:t}) = \int p(\mathbf{z}_{t-k:t} \mid \mathbf{z}_{1:t-k-1}, \mathbf{y}_{t-k:t}) d\mathbf{z}_{t-k+1:t}, \tag{17}$$

where $p(\mathbf{z}_{t-k:t} \mid \mathbf{z}_{1:t-k-1}, \mathbf{y}_{t-k:t})$ *is the density of a truncated normal* $\text{TN}_{m(k+1)}(\mathbf{r}_{t-k:t|t-k-1}, \mathbf{S}_{t-k:t|t-k-1}; \mathbb{A}_{\mathbf{y}_{t-k:t}})$ *with parameters* $\mathbf{r}_{t-k:t|t-k-1} = \mathbb{E}(\mathbf{z}_{t-k:t} \mid \mathbf{z}_{1:t-k-1})$ *and* $\mathbf{S}_{t-k:t|t-k-1} = \text{var}(\mathbf{z}_{t-k:t} \mid \mathbf{z}_{1:t-k-1})$. *The importance weights* $w_t = w(\mathbf{z}_{1:t-k})$ *are, instead, proportional to*

$$\frac{p(\mathbf{y}_{t-k:t} \mid \mathbf{z}_{1:t-k-1})}{p(\mathbf{y}_{t-k:t-1} \mid \mathbf{z}_{1:t-k-1})} = \frac{\Phi_{m(k+1)}(\boldsymbol{\mu}_t; \boldsymbol{\Sigma}_t)}{\Phi_{mk}(\bar{\boldsymbol{\mu}}_t; \bar{\boldsymbol{\Sigma}}_t)}, \tag{18}$$

where the mean vectors are $\boldsymbol{\mu}_t = \mathbf{B}_{t-k:t}\mathbf{r}_{t-k:t|t-k-1}$ *and* $\bar{\boldsymbol{\mu}}_t = \mathbf{B}_{t-k:t-1}\mathbf{r}_{t-k:t-1|t-k-1}$, *whereas the covariance matrices are defined as* $\boldsymbol{\Sigma}_t = \mathbf{B}_{t-k:t}\mathbf{S}_{t-k:t|t-k-1}\mathbf{B}_{t-k:t}$ *and* $\bar{\boldsymbol{\Sigma}}_t = \mathbf{B}_{t-k:t-1}\mathbf{S}_{t-k:t-1|t-k-1}\mathbf{B}_{t-k:t-1}$.

To complete the procedure for sampling from $p(\mathbf{z}_{1:t} \mid \mathbf{y}_{1:t})$ we further require $p(\mathbf{z}_{t-k+1:t} \mid \mathbf{z}_{1:t-k}, \mathbf{y}_{t-k+1:t})$. As clarified in Proposition 2, also such a quantity is the density of a multivariate truncated normal.

Proposition 2 *Under model (3)–(5), it holds*

$$(\mathbf{z}_{t-k+1:t} \mid \mathbf{z}_{1:t-k}, \mathbf{y}_{t-k+1:t}) \sim \text{TN}_{mk}(\mathbf{r}_{t-k+1:t|t-k}, \mathbf{S}_{t-k+1:t|t-k}; \mathbb{A}_{\mathbf{y}_{t-k+1:t}}), \tag{19}$$

with parameters $\mathbf{r}_{t-k+1:t|t-k} = \mathbb{E}(\mathbf{z}_{t-k+1:t} \mid \mathbf{z}_{1:t-k})$ *and* $\mathbf{S}_{t-k+1:t|t-k} = \text{var}(\mathbf{z}_{t-k+1:t} \mid \mathbf{z}_{1:t-k})$.

Note that the expression of the importance weights in equation (18) does not depend on $\mathbf{z}_{t-k}$, and, hence, also in this case the resampling step can be performed before sampling from (17), thus leading to an AUF routine. Besides improving efficiency, such a strategy allows to combine the particle generation in (17) and the completion of the last k terms of $\mathbf{z}_{1:t|t}$ in (19) within a single sampling step from the joint $[m(k+1)]$-variate truncated normal distribution for $(\mathbf{z}_{t-k:t} \mid \mathbf{z}_{1:t-k-1}, \mathbf{y}_{t-k:t})$ reported in Proposition 1. The first m-dimensional component of this vector yields the new delayed particle for $\mathbf{z}_{t-k|t}$ from (17), whereas the whole sub-trajectory provides the desired sample from $p(\mathbf{z}_{t-k:t} \mid \mathbf{z}_{1:t-k-1}, \mathbf{y}_{t-k:t})$ which is joined to the previously resampled particles for $\mathbf{z}_{1:t-k-1|t}$ to form a realization of $\mathbf{z}_{1:t|t}$ from $p(\mathbf{z}_{1:t} \mid \mathbf{y}_{1:t})$. Once this sample is available, one can obtain a draw of $\boldsymbol{\theta}_{t|t}$ from the filtering density $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$ of interest by exploiting the additive representation of the SUN and the analogy between $\mathbf{U}_{1\ 1:t|t}$ and $\mathbf{z}_{1:t|t}$ in (16). In practice, as clarified in Algorithm 3, the updating of $\mathbf{U}_{1\ 1:t|t}$ via lookahead recursion on $\mathbf{z}_{1:t|t}$ and the exact sampling from the Gaussian component of the SUN filtering distribution for $\boldsymbol{\theta}_t$ can be effectively combined in a single online routine based on Kalman filter steps.

To clarify Algorithm 3, note that $p(\boldsymbol{\theta}_t \mid \mathbf{z}_{1:t})$ is the filtering density of the Gaussian dynamic linear model defined in (4)–(5), for which the Kalman filter can be directly implemented, once the trajectory $\mathbf{z}_{1:t|t}$ has been generated from $p(\mathbf{z}_{1:t} \mid \mathbf{y}_{1:t})$ via the lookahead filter. Let $\mathbf{a}_{t-k-1|t-k-1} = \mathbb{E}(\boldsymbol{\theta}_{t-k-1} \mid \mathbf{z}_{1:t-k-1})$, $\mathbf{P}_{t-k-1|t-k-1} = \text{var}(\boldsymbol{\theta}_{t-k-1}|\mathbf{z}_{1:t-k-1})$ and $\mathbf{a}_{t-k|t-k-1} = \mathbb{E}(\boldsymbol{\theta}_{t-k}|\mathbf{z}_{1:t-k-1})$, $\mathbf{P}_{t-k|t-k-1} = \text{var}(\boldsymbol{\theta}_{t-k} \mid \mathbf{z}_{1:t-k-1})$ be the mean vector and covariance matrices for the Gaussian filtering and predictive distributions produced by the standard Kalman filter recursions at time $t-k-1$ under model (4)–(5). Besides being necessary to draw values from the states' filtering and predictive distributions, conditioned on the trajectories of $\mathbf{z}_{1:t|t}$ sampled from $p(\mathbf{z}_{1:t} \mid \mathbf{y}_{1:t})$, such quantities are also sufficient to update online the lookahead parameters $\mathbf{r}_{t-k:t|t-k-1}$ and $\mathbf{S}_{t-k:t|t-k-1}$ that are required to compute the importance weights in Proposition 1, and to sample from the $[m(k+1)]$-variate truncated normal density $p(\mathbf{z}_{t-k:t} \mid \mathbf{z}_{1:t-k-1}, \mathbf{y}_{t-k:t})$ under the auxiliary filter. In particular, the formulation of the dynamic model in (4)–(5) implies that $\mathbf{r}_{t-k:t|t-k-1} = \mathbb{E}(\mathbf{z}_{t-k:t} \mid \mathbf{z}_{1:t-k-1}) = \mathbb{E}(\mathbf{F}_{t-k:t}\boldsymbol{\theta}_{t-k:t} \mid \mathbf{z}_{1:t-k-1})$, and,

Algorithm 3: Lookahead particle filter to draw from $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, for $t = 1, \dots, n$ [AUF version with KF steps]

Set k, and initialize $\mathbf{a}_{0|0}^{(r)} = \mathbf{a}_0$ for $r = 1, \dots, R$ and $\mathbf{P}_{0|0} = \mathbf{P}_0$.

for t *from* 1 *to* k **do**

[1] Sample $\boldsymbol{\theta}_{t|t}^{(1)}, \dots, \boldsymbol{\theta}_{t|t}^{(R)}$ from Algorithm 1 [this can be done efficiently in an exact manner since k is usually small].

for t *from* $k+1$ *to* n **do**

[2] Define the vectors and matrices that are required to perform steps [3] and [4].

[2.1] Set $\mathbf{P}_{t-k|t-k-1} = \mathbf{G}_{t-k}\mathbf{P}_{t-k-1|t-k-1}\mathbf{G}_{t-k}^{\intercal} + \mathbf{W}_{t-k}$ [KF] and compute $\mathbf{S}_{t-k:t|t-k-1}$ as in Sect. 4.2.2.

[2.2] Set $\mathbf{P}_{t-k|t-k} = \mathbf{P}_{t-k|t-k-1} - \mathbf{P}_{t-k|t-k-1}\mathbf{F}_{t-k}^{\intercal}\mathbf{S}_{t-k|t-k-1}^{-1}\mathbf{F}_{t-k}\mathbf{P}_{t-k|t-k-1}$ [KF] .

[2.3] For $r = 1, \dots, R$, set $\mathbf{a}_{t-k|t-k-1}^{(r)} = \mathbf{G}_{t-k}\mathbf{a}_{t-k-1|t-k-1}^{(r)}$ [KF] and compute $\mathbf{r}_{t-k:t|t-k-1}^{(r)}$ as in Sect. 4.2.2.

[3] Implement the resampling step under the AUF version.

[3.1] For $r = 1, \dots, R$, calculate the importance weight $w_t^{(r)}$ via (18).

[3.2] Sample $(\bar{\mathbf{a}}_{t-k|t-k-1}^{(1)}, \bar{\mathbf{r}}_{t-k:t|t-k-1}^{(1)}), \dots, (\bar{\mathbf{a}}_{t-k|t-k-1}^{(R)}, \bar{\mathbf{r}}_{t-k:t|t-k-1}^{(R)})$ from $\sum_{r=1}^{R} w_t^{(r)} \delta_{(\mathbf{a}_{t-k|t-k-1}^{(r)}, \mathbf{r}_{t-k:t|t-k-1}^{(r)})}$.

for r *from* 1 *to* R **do**

[4] Update the delayed particle $\mathbf{z}_{t-k|t}^{(r)}$ and sample $\boldsymbol{\theta}_{t|t}^{(r)}$.

[4.1] Sample $(\mathbf{z}_{t-k|t}^{(r)\intercal}, \bar{\mathbf{z}}_{t-k+1:t|t}^{(r)\intercal})^{\intercal}$ from a $\mathrm{TN}_{m(k+1)}(\bar{\mathbf{r}}_{t-k:t|t-k-1}, \mathbf{S}_{t-k:t|t-k-1}; \mathbb{A}_{\mathbf{y}_{t-k:t}})$.

[4.2] Set $\mathbf{a}_{t-k|t-k}^{(r)} = \bar{\mathbf{a}}_{t-k|t-k-1}^{(r)} + \mathbf{P}_{t-k|t-k-1}\mathbf{F}_{t-k}^{\intercal}\mathbf{S}_{t-k|t-k-1}^{-1}(\mathbf{z}_{t-k|t}^{(r)} - \bar{\mathbf{r}}_{t-k|t-k-1}^{(r)})$ [KF].

[4.3] Compute $\mathbf{a}_{t|t}^{*(r)}$ and $\mathbf{P}_{t|t}^{*(r)}$ by performing k recursions of the KF updates applied to (4)–(5) from $t-k+1$ to t with observations $\mathbf{z}_{t-k+1:t} = \bar{\mathbf{z}}_{t-k+1:t|t}^{(r)}$ and starting moments $\mathbf{a}_{t-k|t-k}^{(r)}$ and $\mathbf{P}_{t-k|t-k}$.

[4.4] Sample $\boldsymbol{\theta}_{t|t}^{(r)}$ from the $\mathrm{N}_p(\mathbf{a}_{t|t}^{*(r)}, \mathbf{P}_{t|t}^{*(r)})$.

therefore, $\mathbf{r}_{t-k:t|t-k-1}$ can be expressed as a function of $\mathbf{a}_{t-k|t-k-1}$ via the direct application of the law of the iterated expectations by stacking the m-dimensional vectors $\mathbf{F}_{t-k}\mathbf{a}_{t-k|t-k-1}$, $\mathbf{F}_{t-k+1}\mathbf{G}_{t-k+1}\mathbf{a}_{t-k|t-k-1}, \dots,$ $\mathbf{F}_t\mathbf{G}_{t-k+1}^{t}\mathbf{a}_{t-k|t-k-1}$, with $\mathbf{G}_l^s$ defined as in Sect. 3.2.

A similar reasoning can be applied to write the covariance matrix $\mathbf{S}_{t-k:t|t-k-1} = \mathrm{var}(\mathbf{z}_{t-k:t} \mid \mathbf{z}_{1:t-k-1})$ as a function of $\mathbf{P}_{t-k|t-k-1}$. In particular letting $l_- = l-1$, the $m \times m$ diagonal blocks of $\mathbf{S}_{t-k:t|t-k-1}$ can obtained sequentially after noticing that

$$\begin{aligned}\mathbf{S}_{t-k:t|t-k-1[ll]} &= \mathrm{var}(\mathbf{z}_{t-k+l_-} \mid \mathbf{z}_{1:t-k-1}) \\ &= \mathbf{F}_{t-k+l_-}\mathbf{P}_{t-k+l_-|t-k-1}\mathbf{F}_{t-k+l_-}^{\intercal} + \mathbf{V}_{t-k+l_-},\end{aligned}$$

for every $l = 1, \dots, k+1$, where the states' covariance matrix $\mathbf{P}_{t-k+l_-|t-k-1}$ at time $t-k+l_-$ can be expressed as a function of $\mathbf{P}_{t-k|t-k-1}$ via the recursive equations $\mathbf{P}_{t-k+l_-|t-k-1} = \mathbf{G}_{t-k+l_-}\mathbf{P}_{t-k+l_--1|t-k-1}\mathbf{G}_{t-k+l_-}^{\intercal} + \mathbf{W}_{t-k+l_-}$, for every $l = 2, \dots, k+1$. Moreover, letting $l_- = l-1$ and $s_- = s-1$, also the off-diagonal blocks can be obtained in a related manner, after noticing that the generic block of $\mathbf{S}_{t-k:t|t-k-1}$ is defined as

$$\begin{aligned}\mathbf{S}_{t-k:t|t-k-1[sl]} &= \mathbf{S}_{t-k:t|t-k-1[ls]}^{\intercal} \\ &= \mathrm{cov}(\mathbf{F}_{t-k+s_-}\boldsymbol{\theta}_{t-k+s_-}, \mathbf{F}_{t-k+l_-}\boldsymbol{\theta}_{t-k+l_-} \mid \mathbf{z}_{1:t-k-1}) \\ &= \mathbf{F}_{t-k+s_-}\mathbf{G}_{t-k+l}^{t-k+s_-}\mathbf{P}_{t-k+l_-|t-k-1}\mathbf{F}_{t-k+l_-}^{\intercal},\end{aligned}$$

for every $s = 2, \dots, k+1$ and $l = 1, \dots, s-1$, where the matrix $\mathbf{P}_{t-k+l_-|t-k-1}$ can be expressed as a function of $\mathbf{P}_{t-k|t-k-1}$ via the recursive equations reported above.

According to these results, the partially collapsed lookahead particle filter for sampling recursively from $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$ simply requires to store and update, for each particle trajectory, the sufficient statistics $\mathbf{a}_{t-k|t-k-1}$ and $\mathbf{P}_{t-k|t-k-1}$ via Kalman filter recursions applied to the model (4)–(5), with every $\mathbf{z}_t$ replaced by the particles generated under the lookahead routine. As previously discussed, also this updating requires only the moments $\mathbf{a}_{t-k|t-k-1}$ and $\mathbf{P}_{t-k|t-k-1}$ computed recursively as a function of the delayed particles' trajectories. This yields to a computational complexity per iteration that is constant with time, as it does not require to compute quantities whose dimension grows with t. In addition, as discussed in Remark 1, such a dual interpretation combined with our SUN closed-form results, provides novel theoretical support to the Rao–Blackwellized particle filter introduced by Andrieu and Doucet (2002).

Remark 1 *The Rao–Blackwellized particle filter by Andrieu and Doucet (2002) for $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$ can be directly obtained as a special case of Algorithm 3, setting $k = 0$.*

Consistent with Remark 1, the Rao–Blackwellized idea (Andrieu and Doucet, 2002) actually coincides with a partially collapsed filter which only updates, without lookahead strategies, the truncated normal component

in the SUN additive representation of the states' filtering distribution, while maintaining the Gaussian term exact. Hence, although this method was originally motivated, in the context of dynamic probit models, also by the apparent lack of an "optimal" closed-form SISR for the states' filtering distribution, our results actually show that such a strategy is expected to yield improved performance relative to the "optimal" particle filter for sampling directly from $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$. In fact, unlike this filter, which is actually available according to Sect. 4.2.1, the Rao–Blackwellized idea avoids the unnecessary autocorrelation in the Gaussian component of the SUN representation, and relies on an optimal particle filter for the multivariate truncated normal part. In addition, Remark 1 and the derivation of the whole class of partially collapsed lookahead filters suggest that setting $k > 0$ is expected to yield further gains relative to the Rao–Blackwellized particle filter; see Sect. 5 for quantitative evidence supporting these results.

5 Illustration on financial time series

Recalling Sects. 1–4, our core contribution in this article is not on developing innovative dynamic models for binary data with improved ability in recovering some ground-truth generative process, but on providing novel closed-form expressions for the filtering, predictive and smoothing distributions under a broad class of routine-use dynamic probit models, along with new Monte Carlo and sequential Monte Carlo strategies for accurate learning of such distributions and the associated functionals in practical applications.

Consistent with the above discussion, we illustrate the practical utility of the closed-form results for the filtering, predictive and smoothing distributions derived in Sect. 3 directly on a realistic real-world dataset, and assess the performance gains of the Monte Carlo strategies developed in Sect. 4. The focus will be on the accuracy in recovering the whole exact SUN distributions of interest, and not just pre-selected functionals. In fact, accurate learning of the entire exact distribution is more challenging and implies, as a direct consequence, accuracy in approximating the associated exact functionals. These assessments are illustrated with a focus on a realistic financial application considering a dynamic probit regression for the daily opening directions of the French CAC40 stock market index from January 4th, 2018 to March 29th, 2019. In this study, the variable y_t is defined on a binary scale, with $y_t = 1$ if the opening value of the CAC40 on day t is greater than the corresponding closing value in the previous day, and $y_t = 0$ otherwise. Financial applications of this type have been a source of particular interest in past and recent years (e.g., Kim and Han, 2000; Kara et al., 2011; Atkins et al., 2018), with common approaches combining a wide variety of technical indicators and news information to forecast stock markets directions via complex machine learning methods. Here, we show how a similar predictive performance can be obtained via a simple and interpretable dynamic probit regression for y_t, which combines past information on the opening directions of CAC40 with those of the NIKKEI225, regarded as binary covariates x_t with dynamic coefficients. Since the Japanese market opens before the French one, x_t is available prior to y_t and, hence, provides a valid predictor for each day t.

Recalling the above discussion and leveraging the default model specifications in these settings (e.g., Soyer and Sung, 2013), we rely on a dynamic probit regression for y_t with two independent random walk processes for the coefficients $\boldsymbol{\theta}_t = (\theta_{1t}, \theta_{2t})^\intercal$. Letting $\mathbf{F}_t = (1, x_t)$ and $\text{pr}(y_t = 1 \mid \boldsymbol{\theta}_t) = \Phi(\theta_{1t} + \theta_{2t} x_t; 1)$, such a model can be expressed as in equations (1)–(2) via

$$\begin{aligned} &p(y_t \mid \boldsymbol{\theta}_t) = \Phi[(2y_t - 1)\mathbf{F}_t\boldsymbol{\theta}_t; 1], \\ &\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \overset{\text{i.i.d.}}{\sim} \text{N}_2(\mathbf{0}, \mathbf{W}), \quad t = 1, \ldots n, \end{aligned} \tag{20}$$

where $\boldsymbol{\theta}_0 \sim \text{N}_2(\mathbf{a}_0, \mathbf{P}_0)$, whereas $\mathbf{W}$ is a time-invariant diagonal matrix. In (20), the element θ_{1t} of $\boldsymbol{\theta}_t$ measures the trend in the directions of the CAC40 when the NIKKEI225 has a negative opening on day t, whereas θ_{2t} characterizes the shift in such a trend if the opening of the NIKKEI225 index is positive, thereby providing an interpretable probit model with dynamic coefficients.

To evaluate performance in smoothing, filtering and prediction, we split the time window in two parts. Observations from January 4th, 2018 to May 31st, 2018 are used as batch data to study the smoothing distribution and to compare the particle filters developed in Sect. 4.2 with other relevant competitors. In the subsequent time window, spanning from June 1st, 2018 to March 29th, 2019, the focus is instead on illustrating performance in online filtering and prediction for streaming data via the lookahead routine derived in Sect. 4.2.2 — which yields the highest approximation accuracy among the online filters evaluated in the first time window.

Figure 3 shows the pointwise median and interquartile range of the smoothing distribution for θ_{1t} and θ_{2t}, $t = 1, \ldots, 97$, based on $R = 10^5$ samples from Algorithm 1. To implement this routine, we set $\mathbf{a}_0 = (0, 0)^\intercal$ and $\mathbf{P}_0 = \text{diag}(3, 3)$ following the guidelines in Gelman et al. (2008) and Chopin and Ridgway (2017) for probit regression. The errors' variances in the diagonal matrix $\mathbf{W}$ are instead set equal to 0.01 as suggested by a graphical search of the maximum for the marginal likelihood computed under different combinations of $(\text{W}_{11}, \text{W}_{22})$ via the analytical formula in Corollary 3.

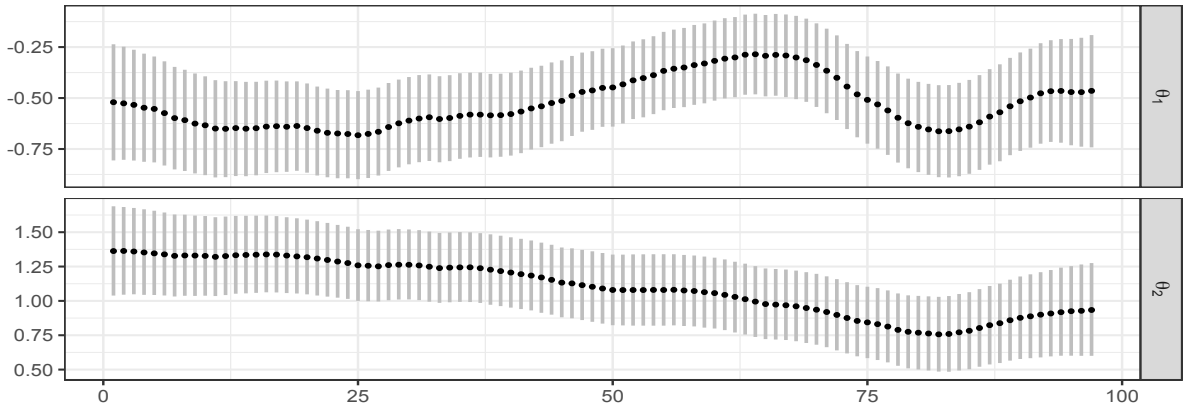

Fig. 3: Pointwise median and interquartile range for the smoothing distributions of θ_{1t} and θ_{2t} in model (20), for the time window from January 4th, 2018 to May 31st, 2018. The quartiles are computed from 10^5 samples produced by Algorithm 1.

As shown in Fig. 3, the dynamic states θ_{1t} and θ_{2t} tend to concentrate around negative and positive values, respectively, for the entire smoothing window, thus highlighting a general concordance between CAC40 and NIKKEI225 opening patterns. However, the strength of this association varies in time, supporting our proposed dynamic probit over static specifications. For example, it is possible to observe a decay in θ_{1t} and θ_{2t} on April–May, 2018 which reduces the association among CAC40 and NIKKEI225, while inducing a general negative trend for the opening directions of the French market. This could be due to the overall instability in the Eurozone on April–May, 2018 caused by the uncertainty after the Italian and British elections during those months.

To clarify the computational improvements of the methods developed in Sects. 4.1 and 4.2, we also compare, in Fig. 4 and in Table 1, their performance against the competing strategies mentioned in Sect. 1. Here, the focus is on the accuracy and computational cost in approximating the exact filtering distribution at time $t = 1, \ldots, 97$, thereby allowing the implementation of the filters discussed in Sect. 1. The competing methods include the extended Kalman filter (Uhlmann, 1992) (EKF), the bootstrap particle filter (Gordon et al., 1993) (BOOT), and the Rao–Blackwellized (RAO-B) sequential Monte Carlo strategy by Andrieu and Doucet (2002), which has been discussed in Sect. 4.2.2 and exploits the hierarchical representation (3)–(5) of model (1)–(2). Although being a popular solution in routine implementations, the extended Kalman filter relies on a quadratic approximation of the probit log-likelihood which leads to Gaussian filtering distributions, thereby affecting the quality of online learning when imbalances in the data induce skewness. The bootstrap particle filter (Gordon et al., 1993) provides, instead, a general SISR that relies on the importance density $p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1})$, thus failing to account effectively for information in $\mathbf{y}_t$, when proposing particles. Rao–Blackwellized sequential Monte Carlo (Andrieu and Doucet, 2002) aims at providing an alternative particle filter, which also addresses the apparent unavailability of an analytic form for the "optimal" particle filter (Doucet et al., 2000). The authors overcome this issue by proposing a sequential Monte Carlo strategy for the Rao–Blackwellized density $p(\mathbf{z}_{1:t} \mid \mathbf{y}_{1:t})$ of the partially observed Gaussian responses $\mathbf{z}_{1:t}$ in model (3)–(5) and compute, for each trajectory $\mathbf{z}_{1:t|t}$, relevant moments of $(\boldsymbol{\theta}_t \mid \mathbf{z}_{1:t|t})$ via classical Kalman filter updates — applied to model (4)–(5) — which are then averaged across the particles to obtain Monte Carlo estimates for the moments of $(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$. As specified in Remark 1, this solution, when adapted to draw samples from $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, is a special case of the sequential strategy in Sect. 4.2.2, with no lookahead, i.e., $k = 0$.

Although the above methods yield state-of-the-art solutions, the proposed strategies are motivated by the apparent absence of a closed-form filter for (1)–(2), that is, in fact, available according to our findings in Sect. 3. Consistent with this argument, we evaluate the accuracy of EFK, BOOT and RAO-B in approximating the exact filtering distribution obtained, for each $t = 1, \ldots, 97$, via direct evaluation of the density from (10). These performances are also compared with those of the new methods proposed in Sect. 4. These include the filtering version of the i.i.d. sampler (I.I.D.) in Sect. 4.1, along with the "optimal" particle filter (OPT) presented in Sect. 4.2.1, and the lookahead sequential Monte Carlo routine derived in Sect. 4.2.2, setting $k = 1$ (LA-1).

For the two dynamic state variables θ_{1t} and θ_{2t}, the accuracy of each sampling scheme is measured via the Wasserstein distance (e.g., Villani, 2008) between the empirical filtering distribution computed, for every time $t = 1, \ldots, 97$, from $R = 10^3$, $R = 10^4$ and $R = 10^5$ particles produced by that specific scheme and the one obtained via the direct evaluation of the associated exact

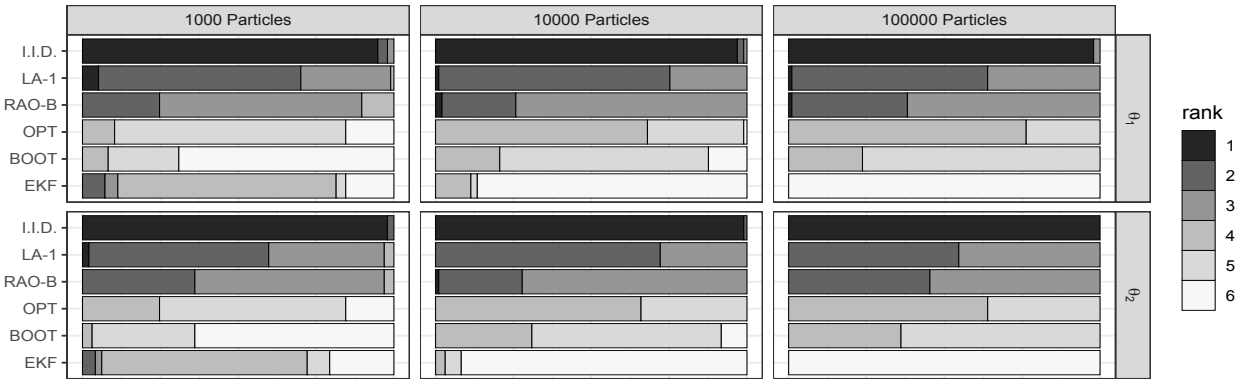

Fig. 4: For the states θ_{1t} and θ_{2t}, barplots representing the relative frequencies of global rankings for the six sampling schemes, in terms of accuracy in approximating the exact SUN filtering distributions over the time window analyzed. For each scheme and time $t = 1, \ldots, 97$, the accuracy is measured via the median Wasserstein distance (over 100 replicated experiments) between the empirical filtering distribution computed from $10^3, 10^4$ and 10^5 particles, respectively, and the one obtained by direct evaluation of the associated exact density from (10) on two grids of 2000 equally spaced values for θ_{1t} and θ_{2t}. This allows to compute, for every $t = 1, \ldots, 97$, the ranking of each sampling scheme in terms of accuracy in approximating the exact filtering density at time t, and to derive the associated barplot summarizing the distribution of the rankings over the whole window.

	ACCURACY					
	θ_{1t} $[R = 10^3]$	θ_{2t} $[R = 10^3]$	θ_{1t} $[R = 10^4]$	θ_{2t} $[R = 10^4]$	θ_{1t} $[R = 10^5]$	θ_{2t} $[R = 10^5]$
I.I.D.	0.01917 [**1**]	0.02362 [**1**]	0.00606 [**1**]	0.00748 [**1**]	0.00199 [**1**]	0.00245 [**1**]
LA–1	0.02558 [**2**]	0.03588 [**2**]	0.00838 [**2**]	0.01133 [**2**]	0.00273 [**2**]	0.00379 [**2**]
RAO–B	0.02700 [**3**]	0.03700 [**3**]	0.00885 [**3**]	0.01201 [**3**]	0.00278 [**3**]	0.00383 [**3**]
OPT	0.06642 [**5**]	0.09063 [**4**]	0.02196 [**4**]	0.03077 [**4**]	0.00687 [**4**]	0.00958 [**4**]
BOOT	0.07237 [**6**]	0.10021 [**5**]	0.02325 [**5**]	0.03225 [**5**]	0.00728 [**5**]	0.00992 [**5**]
EKF	0.06108 [**4**]	0.10036 [**6**]	0.05853 [**6**]	0.09824 [**6**]	0.05829 [**6**]	0.09802 [**6**]
	COMPUTATIONAL COST					
I.I.D.	$\mathcal{O}(tp^3 + t^3m^3 + R[p^2 + t^2m^2C(mt)])$					
LA–1	$\mathcal{O}(t(p^3 + m^3) + tR[p^2 + pm + m^2C(2m)] + tM[m^2 + Rm])$					
RAO–B	$\mathcal{O}(t(p^3 + m^3) + tR[p^2 + pm + m^2C(m)] + tM[m^2 + Rm])$					
OPT	$\mathcal{O}(t(p^3 + m^3) + tR[p^2 + pm + m^2C(m)] + tM[m^2 + Rm])$					
BOOT	$\mathcal{O}(t(p^3 + m^3) + tR(p^2 + pm) + tM[m^2 + Rm])$					
EKF	$\mathcal{O}(t[p^3 + m^3 + Mm^2])$					

Table 1: For the states θ_{1t} and θ_{2t}, averaged accuracy in approximating the exact SUN filtering distribution at $t = 1, \ldots, 97$, and computational cost for obtaining a sample of dimension R from such a filtering distribution at time t. For each scheme, the accuracy is measured via the Wasserstein distance between the empirical filtering distribution computed from $10^3, 10^4$ and 10^5 particles, respectively, and the one obtained via direct evaluation of the associated exact SUN density from (10) on two grids of 2000 equally spaced values for θ_{1t} and θ_{2t}. For each t, we first compute the median Wasserstein distance from 100 replicated experiments, and then average such quantities across time. Numbers in square brackets denote the ranking in each column. The costs are derived for the case in which the importance weights are evaluated via Monte Carlo based on M samples. For the EKF, we provide the cost of the KF recursions, when the probit likelihood is evaluated via M Monte Carlo samples.

density from (10) on two grids of 2000 equally spaced values for θ_{1t} and θ_{2t}. For the sake of clarity, with a little abuse of terminology, the term *particle* refers both to the samples of the sequential Monte Carlo methods and to those obtained under i.i.d. sampling from the SUN. The Wasserstein distance is computed via the `R` function `wasserstein1d`. Note also that, although EKF and RAO-B focus, mostly, on moments of $(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, such strategies can be adapted to sample from an approximation of the filtering distribution. Figure 4 displays, for the two states and for varying number of particles, the frequencies of the global rankings of the different schemes, out of the 97 time instants. Such rankings are computed according to the median Wasserstein distance obtained, for each $t = 1, \ldots, 97$, from 100 replicated experiments. The overall averages across time of these median Wasserstein distances are reported in Table 1, along with computational costs for obtaining R samples from the filtering at time t under each scheme; see Appendix B for detailed derivations of such costs.

Figure 4 and Table 1 confirm that the I.I.D. sampler in Sect. 4.1 over-performs the competitors in accuracy, since the averaged median Wasserstein distances from the exact filtering distribution are lower than those of

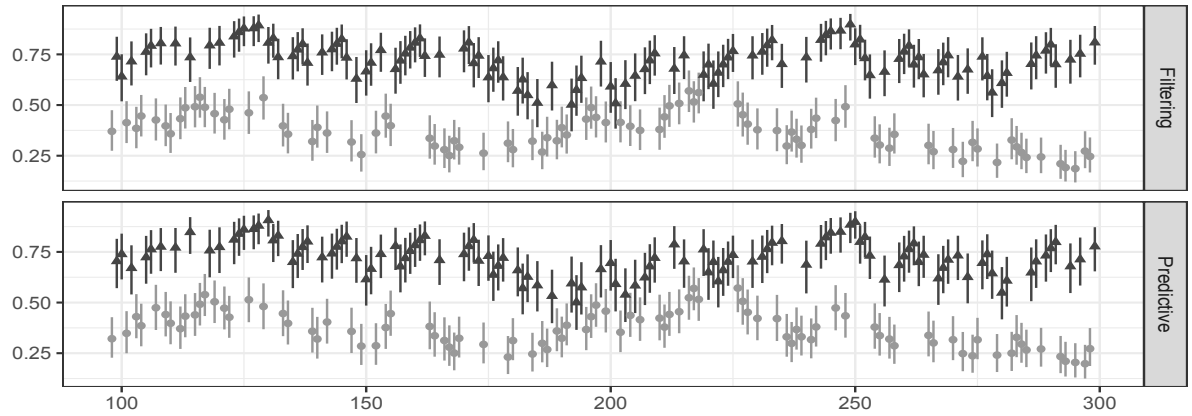

Fig. 5: Median and interquartile range of the filtering and predictive distributions for $\Phi(\theta_{1t} + x_t\theta_{2t}; 1)$ computed from 10^5 particles produced by the lookahead particle filter in Algorithm 3 for the second time window. Black and grey segments denote days in which $x_t = 1$ and $x_t = 0$, respectively.

the other schemes under all settings, and the ranking of the I.I.D. is 1 in almost all the 97 times. This improved performance comes, however, with a higher computational complexity, especially in the sampling from (mt)-variate truncated normals in the SUN additive representation, which yields a cost depending on $C(mt)$, i.e., the average number of proposed draws required to accept one sample. While the improved accuracy of I.I.D. justifies such a cost in small-to-moderate dimensions, as t increases the I.I.D. becomes progressively impractical, thus motivating scalable particle filters with linear cost in t, such as BOOT, RAO-B, OPT and LA-1. In our basic `R` implementation, we found that the proposed I.I.D. sampler has reasonable runtimes (of a couple of minutes) also for larger series with $mt \approx 300$. However, in much higher dimensions the particle filters become orders of magnitude faster and still practically effective.

As expected, the OPT filter in Sect. 4.2.1 tends to improve the performance of BOOT, since this strategy is optimal within the class where BOOT is defined. However, as discussed in Sects. 4.2.1 and 4.2.2, both methods induce unnecessary autocorrelation in the Gaussian part of the SUN filtering distribution, thus yielding suboptimal solutions relative to particle filters that perform sequential Monte Carlo only on the multivariate truncated normal component. The accuracy gains of RAO-B and LA-1 relative to BOOT and OPT in Fig. 4 and Table 1 provide empirical evidence in support of this argument, while displaying additional improvements of the lookahead strategy derived in Sect. 4.2.2 over RAO-B, even when k is set just to 1, i.e., LA-1. As shown in Table 1, the complexities of LA-1 and RAO-B are of the same order, except for sampling from bivariate truncated normals under LA-1 instead of univariate ones as in RAO-B. This holds for any fixed k, with the additional sampling cost being $C(m[k+1])$. However, consistent with the results in Fig. 4 and Table 1 it suffices to set k quite small to already obtain some accuracy gains, thus making such increments in computational cost affordable in practice. The EKF is, overall, the less accurate solution since, unlike the other methods, it relies on a Gaussian approximation of the SUN filtering distribution. This is only beneficial relative to BOOT and OPT when the number of particles is small, due to the reduced mixing of such strategies induced by the autocorrelation in the Gaussian component of the SUN additive representation. All these results remained consistent also when comparing other quantiles of the Wasserstein distance across experiments and when studying the accuracy in approximating pre-selected functionals of interest.

Motivated by the accurate performance of the novel lookahead strategy in Sect. 4.2.2, we apply LA-1 to provide scalable online filtering and prediction for model (20) from June 1st, 2018 to March 29th, 2019. Following the idea of sequential inference, the particles are initialized exploiting the marginal smoothing distribution of May 31, 2018 from the batch analysis. Figure 5 outlines median and interquartile range for the filtering and predictive distribution of the probability that CAC40 has a positive opening in each day of the window considered for online inference. These two distributions can be easily obtained by applying the function $\Phi(\theta_{1t} + x_t\theta_{2t}; 1)$ to the particles of the states filtering and predictive distribution. In line with Fig. 3, a positive opening of the NIKKEI225 provides, in general, a high estimate for the probability that $y_t = 1$, whereas a negative opening tends to favor the event $y_t = 0$. However, the strength of this result evolves over time with some periods showing less evident shifts in the probabilities process when x_t changes from 1 to 0. One-step-ahead prediction, leveraging the samples of the predictive distribution for the probability process, led to a correct classification rate

of 66.34% which is comparable to those obtained under more complex procedures combining a wide variety of inputs to predict stock markets directions via state-of-the-art machine learning methods (e.g., Kim and Han, 2000; Kara et al., 2011; Atkins et al., 2018).

6 Discussion

This article shows that filtering, predictive and smoothing densities in multivariate dynamic probit models have a SUN kernel and the associated parameters can be computed via tractable expressions. As discussed in Sects. 3–5, this result provides advances in online inference and facilitates the implementation of tractable methods to draw i.i.d. samples from the exact filtering, predictive and smoothing distributions, thereby allowing improved Monte Carlo inference in small-to-moderate settings. Filtering in higher dimensions can be, instead, implemented via scalable sequential Monte Carlo which exploits SUN properties to provide novel particle filters.

Such advances motivate future research. For example, a relevant direction is to extend the results in Sect. 3 to dynamic tobit, binomial and multinomial probit models, for which closed-form filters are unavailable. In the multinomial setting a viable solution is to exploit the results in Fasano and Durante (2021) for the static case. Joint filtering and prediction of continuous and binary time series is also of interest (Liu et al., 2009). A natural state-space model for these data can be obtained by allowing only the sub-vector of Gaussian variables associated with the binary data to be partially observed in (3)–(5). However, also in this case, closed-form filters are unavailable. By combining our results in Sect. 3 with classical Kalman filter, this gap may now be covered.

As mentioned in Sects. 1 and 3.2, estimation of possible unknown parameters characterizing the state-space model in (1)–(2) is another relevant problem, that can be addressed by maximizing the marginal likelihood derived in Sect. 3.2. This quantity can be explicitly evaluated as in Corollary 3 for any small-to-moderate n. A more scalable option in large n settings is to rely on equations (62) and (66) in Doucet et al. (2000) which allow to evaluate the marginal likelihood leveraging samples from particle filters. In this respect, the improved lookahead filter developed in Sect. 4.2.2 is expected to yield accuracy gains also in parameter estimation, when used as a scalable strategy to evaluate marginal likelihoods. This routine can be also adapted to sample from the joint smoothing distribution via a backward recursion. However, unlike the i.i.d. sampler in Algorithm 1, this approach yields an additional computational cost which is quadratic in the total number of particles R (e.g., Doucet et al., 2000). Since R is much higher than n in most applications, the i.i.d. sampler developed in Algorithm 1 is preferable over particle smoothers in routine studies having small-to-moderate dimension, since it also yields improved accuracy by avoiding sequential Monte Carlo. Finally, additional quantitative studies beyond those in Sect. 5 can be useful for obtaining further insights on the performance of our proposed algorithms relative to state-of-the-art strategies, including recent ensemble sampling (Deligiannidis et al., 2020).

Data and Codes. The dataset considered in Sect. 5 is available at `Yahoo Finance`. Pseudo-codes that can be easily implemented with any software are provided in Algorithms 1–3.

Acknowledgments. We are grateful to the Editor, the Associate Editor, and the two anonymous referees for their precious comments and constructive feedbacks, which helped us in improving the preliminary version of this paper.

Appendix A: Proofs of the main results

Proof of Lemma 1. To prove Lemma 1, note that, by applying the Bayes' rule, we obtain

$$p(\boldsymbol{\theta}_1 \mid \mathbf{y}_1) \propto p(\boldsymbol{\theta}_1)p(\mathbf{y}_1 \mid \boldsymbol{\theta}_1),$$

where $p(\boldsymbol{\theta}_1) = \phi_p(\boldsymbol{\theta}_1 - \mathbf{G}_1\mathbf{a}_0; \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1)$ and $p(\mathbf{y}_1 \mid \boldsymbol{\theta}_1) = \Phi_m(\mathbf{B}_1\mathbf{F}_1\boldsymbol{\theta}_1; \mathbf{B}_1\mathbf{V}_1\mathbf{B}_1)$. The expression for $p(\boldsymbol{\theta}_1)$ can be easily obtained by noting that $\boldsymbol{\theta}_1 = \mathbf{G}_1\boldsymbol{\theta}_0 + \boldsymbol{\varepsilon}_1$ in (2), with $\boldsymbol{\theta}_0 \sim \mathrm{N}_p(\mathbf{a}_0, \mathbf{P}_0)$ and $\boldsymbol{\varepsilon}_1 \sim \mathrm{N}_p(\mathbf{0}, \mathbf{W}_1)$. The form for the probability mass function of $(\mathbf{y}_1 \mid \boldsymbol{\theta}_1)$ is instead a direct consequence of equation (1). Hence, combining these expressions and recalling (6), it is clear that $p(\boldsymbol{\theta}_1 \mid \mathbf{y}_1)$ is proportional to the density of a SUN with suitably–specified parameters, such that the kernel of (6) coincides with $\phi_p(\boldsymbol{\theta}_1 - \mathbf{G}_1\mathbf{a}_0; \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1)\Phi_m(\mathbf{B}_1\mathbf{F}_1\boldsymbol{\theta}_1; \mathbf{B}_1\mathbf{V}_1\mathbf{B}_1)$. In particular, letting

$$\begin{aligned}
&\boldsymbol{\xi}_{1|1} = \mathbf{G}_1\mathbf{a}_0, \quad \boldsymbol{\Omega}_{1|1} = \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1,\\
&\boldsymbol{\Delta}_{1|1} = \bar{\boldsymbol{\Omega}}_{1|1}\boldsymbol{\omega}_{1|1}\mathbf{F}_1^\intercal\mathbf{B}_1\mathbf{s}_1^{-1}, \quad \boldsymbol{\gamma}_{1|1} = \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{F}_1\boldsymbol{\xi}_{1|1},\\
&\boldsymbol{\Gamma}_{1|1} = \mathbf{s}_1^{-1}\mathbf{B}_1(\mathbf{F}_1\boldsymbol{\Omega}_{1|1}\mathbf{F}_1^\intercal + \mathbf{V}_1)\mathbf{B}_1\mathbf{s}_1^{-1},
\end{aligned}$$

we have that

$$\begin{aligned}
&\boldsymbol{\gamma}_{1|1} + \boldsymbol{\Delta}_{1|1}^\intercal\bar{\boldsymbol{\Omega}}_{1|1}^{-1}\boldsymbol{\omega}_{1|1}^{-1}(\boldsymbol{\theta}_1 - \boldsymbol{\xi}_{1|1})\\
&= \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{F}_1\boldsymbol{\xi}_{1|1} + \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{F}_1(\boldsymbol{\theta}_1 - \boldsymbol{\xi}_{1|1}) = \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{F}_1\boldsymbol{\theta}_1,\\
&\boldsymbol{\Gamma}_{1|1} - \boldsymbol{\Delta}_{1|1}^\intercal\bar{\boldsymbol{\Omega}}_{1|1}^{-1}\boldsymbol{\Delta}_{1|1}\\
&= \mathbf{s}_1^{-1}[\mathbf{B}_1(\mathbf{F}_1\boldsymbol{\Omega}_{1|1}\mathbf{F}_1^\intercal + \mathbf{V}_1)\mathbf{B}_1 - \mathbf{B}_1(\mathbf{F}_1\boldsymbol{\Omega}_{1|1}\mathbf{F}_1^\intercal)\mathbf{B}_1]\mathbf{s}_1^{-1}\\
&= \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{V}_1\mathbf{B}_1\mathbf{s}_1^{-1}.
\end{aligned}$$

with $\mathbf{s}_1^{-1}$ as in Lemma 1. Note that this term is introduced to make $\boldsymbol{\Gamma}_{1|1}$ a correlation matrix, as required in the SUN parametrization (Arellano-Valle and Azzalini,

2006). Recalling Durante (2019), and substituting these quantities in the kernel of the SUN density (6), we have

$$\begin{aligned}&\phi_p(\boldsymbol{\theta}_1 - \mathbf{G}_1\mathbf{a}_0; \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1)\\&\quad\cdot \Phi_m(\mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{F}_1\boldsymbol{\theta}_1; \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{V}_1\mathbf{B}_1\mathbf{s}_1^{-1})\\&= \phi_p(\boldsymbol{\theta}_1 - \mathbf{G}_1\mathbf{a}_0; \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1)\Phi_m(\mathbf{B}_1\mathbf{F}_1\boldsymbol{\theta}_1; \mathbf{B}_1\mathbf{V}_1\mathbf{B}_1)\\&= p(\boldsymbol{\theta}_1)p(\mathbf{y}_1 \mid \boldsymbol{\theta}_1) \propto p(\boldsymbol{\theta}_1 \mid \mathbf{y}_1),\end{aligned}$$

thus proving Lemma 1. To prove that $\boldsymbol{\Omega}^*_{1|1}$ is a correlation matrix, replace the indentity $\mathbf{I}_m$ with $\mathbf{B}_1\mathbf{V}_1\mathbf{B}_1$ in the proof of Theorem 1 by Durante (2019). □

Proof of Theorem 1. Recalling equation (2), the proof for $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ in (9) requires studying the variable $\mathbf{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\varepsilon}_t$, given $\mathbf{y}_{1:t-1}$, where

$$\begin{aligned}(\boldsymbol{\theta}_{t-1} \mid \mathbf{y}_{1:t-1}) \sim \text{SUN}_{p,m(t-1)}(&\boldsymbol{\xi}_{t-1|t-1}, \boldsymbol{\Omega}_{t-1|t-1},\\&\boldsymbol{\Delta}_{t-1|t-1}, \boldsymbol{\gamma}_{t-1|t-1}, \boldsymbol{\Gamma}_{t-1|t-1}),\end{aligned}$$

and $\boldsymbol{\varepsilon}_t \sim \text{N}_p(\mathbf{0}, \mathbf{W}_t)$, with $\boldsymbol{\varepsilon}_t \perp \mathbf{y}_{1:t-1}$. To address this goal, first note that, by the closure properties of the SUN under linear transformations (Azzalini and Capitanio, 2014, Sect. 7.1.2), we have that $(\mathbf{G}_t\boldsymbol{\theta}_{t-1} \mid \mathbf{y}_{1:t-1})$ is still a SUN with parameters $\mathbf{G}_t\boldsymbol{\xi}_{t-1|t-1}$, $\mathbf{G}_t\boldsymbol{\Omega}_{t-1|t-1}\mathbf{G}_t^\intercal$, $[(\mathbf{G}_t\boldsymbol{\Omega}_{t-1|t-1}\mathbf{G}_t^\intercal)\odot\mathbf{I}_p]^{-\frac{1}{2}}\mathbf{G}_t\boldsymbol{\omega}_{t-1|t-1}\boldsymbol{\Delta}_{t-1|t-1}$, $\boldsymbol{\gamma}_{t-1|t-1}$ and $\boldsymbol{\Gamma}_{t-1|t-1}$. Hence, to conclude the proof of equation (9), we only need to obtain the distribution of the sum among this variable and the noise $\boldsymbol{\varepsilon}_t \sim \text{N}_p(\mathbf{0}, \mathbf{W}_t)$. This can be accomplished by considering the moment generating function of such a sum — as done by Azzalini and Capitanio (2014, Sect. 7.1.2) to prove closure under convolution. Indeed, it is straightforward to note that the product of the moment generating functions for $\boldsymbol{\varepsilon}_t$ and $(\mathbf{G}_t\boldsymbol{\theta}_{t-1} \mid \mathbf{y}_{1:t-1})$ leads to the moment generating function of a SUN having parameters $\boldsymbol{\xi}_{t|t-1} = \mathbf{G}_t\boldsymbol{\xi}_{t-1|t-1}$, $\boldsymbol{\Omega}_{t|t-1} = \mathbf{G}_t\boldsymbol{\Omega}_{t-1|t-1}\mathbf{G}_t^\intercal + \mathbf{W}_t$, $\boldsymbol{\Delta}_{t|t-1} = \boldsymbol{\omega}^{-1}_{t|t-1}\mathbf{G}_t\boldsymbol{\omega}_{t-1|t-1}\boldsymbol{\Delta}_{t-1|t-1}$, $\boldsymbol{\gamma}_{t|t-1} = \boldsymbol{\gamma}_{t-1|t-1}$ and $\boldsymbol{\Gamma}_{t|t-1} = \boldsymbol{\Gamma}_{t-1|t-1}$. To prove (10) note that

$$p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}) \propto \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$$

coincides with the posterior density in the probit model having likelihood $\Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)$, and SUN prior $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ from (9). Hence, (10) can be derived from Corollary 4 in Durante (2019), replacing matrix $\mathbf{I}_m$ in the classical probit likelihood with $\mathbf{B}_t\mathbf{V}_t\mathbf{B}_t$. □

Proof of Corollary 1. To prove Corollary 1, re-write $\int \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})\text{d}\boldsymbol{\theta}_t$ as

$$\frac{\int \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)K(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})\text{d}\boldsymbol{\theta}_t}{\Phi_{m(t-1)}(\boldsymbol{\gamma}_{t|t-1}; \boldsymbol{\Gamma}_{t|t-1})},$$

with $K(\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1}) = p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1})\Phi_{m(t-1)}(\boldsymbol{\gamma}_{t|t-1};\boldsymbol{\Gamma}_{t|t-1})$ denoting the kernel of the predictive density from (9). Consistent with this result, Corollary 1 follows by noting that $\Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)K(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ is the kernel of the filtering density from (10), whose normalizing constant $\int \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)K(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})\text{d}\boldsymbol{\theta}_t$ is equal to $\Phi_{mt}(\boldsymbol{\gamma}_{t|t}; \boldsymbol{\Gamma}_{t|t})$. □

Proof of Theorem 2. First notice that $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n}) \propto p(\boldsymbol{\theta}_{1:n})p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}_{1:n})$. Therefore, $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$ can be seen as the posterior density in the Bayesian model with likelihood $p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}_{1:n})$ and prior $p(\boldsymbol{\theta}_{1:n})$ for the vector $\boldsymbol{\theta}_{1:n} = (\boldsymbol{\theta}_1^\intercal, \ldots, \boldsymbol{\theta}_n^\intercal)^\intercal$. As pointed out in Sect. 3.2, it follows from (2) that $\boldsymbol{\theta}_{1:n} \sim \text{N}_{pn}(\boldsymbol{\xi}, \boldsymbol{\Omega})$, with $\boldsymbol{\xi}$ and $\boldsymbol{\Omega}$ defined in Sect. 3.2. The form of $p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}_{1:n})$ can be obtained from (1), by noticing that $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are conditionally independent given $\boldsymbol{\theta}_{1:n}$, thus providing the joint likelihood $p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}_{1:n}) = \prod_{s=1}^n \Phi_m(\mathbf{B}_s\mathbf{F}_s\boldsymbol{\theta}_s; \mathbf{B}_s\mathbf{V}_s\mathbf{B}_s)$. This quantity can be re-written as $\Phi_{mn}(\mathbf{D}\boldsymbol{\theta}_{1:n}; \boldsymbol{\Lambda})$ with $\mathbf{D}$ and $\boldsymbol{\Lambda}$ as in Sect. 3.2. Combining these results and recalling the proof of Lemma 1, if follows that $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n}) \propto \phi_{pn}(\boldsymbol{\theta}_{1:n} - \boldsymbol{\xi}; \boldsymbol{\Omega})\Phi_{mn}(\mathbf{D}\boldsymbol{\theta}_{1:n}; \boldsymbol{\Lambda})$, which coincides with the kernel of the SUN in Theorem 2. □

Proof of Corollary 3. The expression for the marginal likelihood follows by noting that $p(\mathbf{y}_{1:n})$ is the normalizing constant of the smoothing density. Indeed, $p(\mathbf{y}_{1:n}) = \int p(\mathbf{y}_{1:n}|\boldsymbol{\theta}_{1:n})p(\boldsymbol{\theta}_{1:n})d\boldsymbol{\theta}_{1:n}$. Hence, the integrand coincides with the kernel of the smoothing density, so that the whole integral is equal to $\Phi_{mn}(\boldsymbol{\gamma}_{1:n|n}; \boldsymbol{\Gamma}_{1:n|n})$. □

Proof of Corollary 4. The proof of Corollary 4 is similar to that of Lemma 1. Indeed, the proposal $p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{y}_t)$ is proportional to the product between the likelihood $p(\mathbf{y}_t \mid \boldsymbol{\theta}_t) = \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)$ and the prior $p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}) = \phi_p(\boldsymbol{\theta}_t - \mathbf{G}_t\boldsymbol{\theta}_{t-1}; \mathbf{W}_t)$. To derive the importance weights in (15), it suffices to notice that the marginal likelihood $p(\mathbf{y}_t \mid \boldsymbol{\theta}_{t-1})$ coincides with the normalizing constant of the SUN in (14). □

Proof of Proposition 1. To derive the form of the proposal, first notice that $p(\mathbf{z}_{t-k:t} \mid \mathbf{z}_{1:t-k-1}, \mathbf{y}_{t-k:t}) \propto p(\mathbf{z}_{t-k:t} \mid \mathbf{z}_{1:t-k-1})p(\mathbf{y}_{t-k:t} \mid \mathbf{z}_{1:t})$. Recalling model (3)–(5) and Sect. 4.2.2, we have that $(\mathbf{z}_{t-k:t} \mid \mathbf{z}_{1:t-k-1}) \sim \text{N}_{m(k+1)}(\mathbf{r}_{t-k:t|t-k-1}, \mathbf{S}_{t-k:t|t-k-1})$ and $p(\mathbf{y}_{t-k:t}|\mathbf{z}_{1:t}) = \mathbb{1}(\mathbf{z}_{t-k:t} \in \mathbb{A}_{\mathbf{y}_{t-k:t}})$. Hence, $p(\mathbf{z}_{t-k:t} \mid \mathbf{z}_{1:t-k-1})p(\mathbf{y}_{t-k:t} \mid \mathbf{z}_{1:t})$ is the kernel of the $[m(k+1)]$-variate truncated normal in Proposition 1. The form of the weights in (18) follows from their general expression (e.g., Andrieu and Doucet, 2002, Sect. 2.2.1), combined with the sequential formulation of the model. Note also that, when written as a function of $\mathbf{z}_s$ from the proposal, $p(\mathbf{y}_s \mid \mathbf{z}_s) = 1$, for any $s = 1, \ldots, t-k$. Therefore, with the convention

that $p(\mathbf{z}_1 \mid \mathbf{z}_0) = p(\mathbf{z}_1)$, the weights are proportional to

$$\begin{aligned}
&\frac{p(\mathbf{z}_{1:t-k} \mid \mathbf{y}_{1:t})}{p(\mathbf{z}_{1:t-k-1} \mid \mathbf{y}_{1:t-1})p(\mathbf{z}_{t-k} \mid \mathbf{z}_{1:t-k-1}, \mathbf{y}_{t-k:t})} \\
&\propto \frac{p(\mathbf{y}_{1:t} \mid \mathbf{z}_{1:t-k})p(\mathbf{z}_{1:t-k})/p(\mathbf{z}_{1:t-k-1})}{p(\mathbf{y}_{1:t-1} \mid \mathbf{z}_{1:t-k-1})p(\mathbf{z}_{t-k} \mid \mathbf{z}_{1:t-k-1}, \mathbf{y}_{t-k:t})} \\
&= \frac{p(\mathbf{y}_{1:t} \mid \mathbf{z}_{1:t-k})p(\mathbf{z}_{t-k} \mid \mathbf{z}_{1:t-k-1})}{p(\mathbf{y}_{1:t-1} \mid \mathbf{z}_{1:t-k-1})p(\mathbf{z}_{t-k} \mid \mathbf{z}_{1:t-k-1}, \mathbf{y}_{t-k:t})} \\
&= \frac{p(\mathbf{y}_{1:t} \mid \mathbf{z}_{1:t-k})p(\mathbf{y}_{t-k:t} \mid \mathbf{z}_{1:t-k-1})}{p(\mathbf{y}_{1:t-1} \mid \mathbf{z}_{1:t-k-1})p(\mathbf{y}_{t-k:t} \mid \mathbf{z}_{1:t-k})} \\
&= \frac{p(\mathbf{y}_{t-k:t} \mid \mathbf{z}_{1:t-k-1})}{p(\mathbf{y}_{1:t-1} \mid \mathbf{z}_{1:t-k-1})} = \frac{p(\mathbf{y}_{t-k:t} \mid \mathbf{z}_{1:t-k-1})}{p(\mathbf{y}_{t-k:t-1} \mid \mathbf{z}_{1:t-k-1})},
\end{aligned}$$

where the last equality follows from the fact that $p(\mathbf{y}_{1:t} \mid \mathbf{z}_{1:t-k}) = p(\mathbf{y}_{t-k:t} \mid \mathbf{z}_{1:t-k})$. To obtain the final form of equation (18) if suffices to notice that $p(\mathbf{y}_{t-k:t} \mid \mathbf{z}_{1:t-k-1}) = \text{pr}(\mathbf{B}_{t-k:t}\tilde{\mathbf{z}} > \mathbf{0}) = \Phi_{m(k+1)}(\boldsymbol{\mu}_t; \boldsymbol{\Sigma}_t)$, where $\tilde{\mathbf{z}} \sim \text{N}_{m(k+1)}(\mathbf{r}_{t-k:t|t-k-1}, \mathbf{S}_{t-k:t|t-k-1})$, with $\mathbf{r}_{t-k:t|t-k-1}$, $\mathbf{S}_{t-k:t|t-k-1}$, and $\mathbf{B}_{t-k:t}$ defined as in Sect. 4.2.2. A similar argument holds for the denominator of (18). □

Appendix B: Derivation of computational costs

In this section we derive the computational costs of the algorithms discussed in Sects. 4 and 5. Let us first consider Algorithm 1 with an initial focus on the smoothing distribution. For this routine, the matrix computations to obtain the parameters of interest require $\mathcal{O}(n^3[p^3 + m^3])$ operations. Regarding the sampling cost to obtain R draws, step [1] requires $\mathcal{O}(p^3n^3 + Rp^2n^2)$ operations since we have to first compute the Cholesky decomposition of $\bar{\boldsymbol{\Omega}}_{1:n|n} - \boldsymbol{\Delta}_{1:n|n}\boldsymbol{\Gamma}^{-1}_{1:n|n}\boldsymbol{\Delta}^{\intercal}_{1:n|n}$ in $\mathcal{O}(p^3n^3)$, and then multiply each independent sample for the resulting lower triangular matrix, at $\mathcal{O}(Rp^2n^2)$ total cost. Step [2] requires, instead, to obtain a minimax exponentially-tilted estimate at $\mathcal{O}(m^3n^3)$ cost (Botev, 2017) and then perform $\mathcal{O}(n^2m^2C(mn))$ operations for each independent sample, where $C(d)$ denotes the average number of proposed draws required per accepted sample in Botev (2017), when the dimension of the truncated normal is d. Hence, the overall cost of Algorithm 1 is $\mathcal{O}(n^3(p^3 + m^3) + Rn^2[p^2 + m^2C(mn)])$. If the interest is in the filtering distribution, which coincides with the marginal smoothing at $n = t$, it is sufficient to sample $\mathbf{U}_{0\ n|n}$ instead of $\mathbf{U}_{0\ 1:n|n}$. Hence, the overall cost for R samples reduces to $\mathcal{O}(tp^3 + t^3m^3 + R[p^2 + t^2m^2C(mt)])$.

We now consider the computational costs of the particle filters considered in Sect. 4 and 5. For each t, the cost is due to computation of parameters, sampling and evaluation of the importance weights. Starting with the "optimal" particle filter in Sect. 4.2.1, the matrix operations for computing the quantities in steps [3.1]–[3.3] of Algorithm 2 have an overall cost for the R samples of $\mathcal{O}(m^3 + pm^2 + p^2m + Rpm + Rp^2)$. The sampling costs are, instead, $\mathcal{O}(p^3+Rp^2)$ and $\mathcal{O}(m^3+Rm^2C(m))$ for the Gaussian and truncated normal terms, respectively. To conclude the derivation of the computational costs, it is necessary to derive those associated with the evaluation of the importance weights. For all the particle filters analyzed, such weights are obtained by evaluating in R different points the cumulative distribution function of a zero mean multivariate normal with fixed covariance matrix. To facilitate comparison, we assume that this evaluation relies on a Monte Carlo estimate based on M samples in all the particle filters. For the "optimal" particle filter, this step requires $\mathcal{O}(m^3 + Mm^2)$ operations to obtain the samples, plus $\mathcal{O}(MRm)$ for computing the Monte Carlo estimate. Combining these results, the overall cost for the "optimal" particle filter at time t is $\mathcal{O}(t(p^3+m^3)+tR[p^2+pm+m^2C(m)]+tM[m^2+Rm])$.

Let us now derive the cost of the Rao–Blackwellized algorithm by Andrieu and Doucet (2002). In this case, adapting the notation of the original paper to the one of Sect. 4.2.2, it can be noticed that one KF step requires $\mathcal{O}(p^3+Rp^2+Rpm+m^3)$ operations for the computation of $\mathbf{P}_{t|t-1}, \mathbf{a}_{t|t-1}, \mathbf{S}_{t|t-1}, \mathbf{r}_{t|t-1}, \mathbf{P}_{t|t}$ and $\mathbf{a}_{t|t}$, at any t. As for the sampling part, it first requires R draws from an m-variate truncated normal. Exploiting the same arguments considered for the previous algorithms, this step has an $\mathcal{O}(m^3+Rm^2C(m))$ cost. The sampling from the final Gaussian filtering distribution $p(\boldsymbol{\theta}_t \mid \mathbf{z}_{1:t} = \mathbf{z}_{1:t|t})$ of direct interest requires instead $\mathcal{O}(p^3 + Rp^2)$ operations. Leveraging again the derivations for the previous algorithms, the computation of the importance weights has cost $\mathcal{O}(m^3 + Mm^2 + RMm)$. Therefore, the overall cost of the sequential filtering procedure at time t is $\mathcal{O}(t(p^3+m^3)+tR[p^2+pm+m^2C(m)]+tM[m^2+Rm])$.

The above derivations for the Rao–Blackwellized algorithm directly extend to the partially collapsed lookahead particle filter shown in Algorithm 3. In fact, while at each t the Rao–Blackwellized solution requires one KF recursion combined with sampling from m-variate truncated normals and evaluation of cumulative distribution functions of m-variate Gaussians, the lookahead routine relies on samples from $[m(k+1)]$-variate truncated normals along with $k+1$ KF steps, and computation of cumulative distribution functions for $[m(k+1)]$-dimensional Gaussians. Hence, adapting the cost of the Rao–Blackwellized algorithm to this broader setting, we have that the overall cost of Algorithm 3 at time t is $\mathcal{O}(t(k_+p^3+k_+^3m^3)+tR[k_+p^2+k_+pm+k_+^2m^2C(k_+m)]+tM[k_+^2m^2 + Rk_+m])$, where $k_+ = k + 1$. Note that, in practice, k is set equal to a pre-specified small constant and, therefore, the actual implementation cost reduces to $\mathcal{O}(t(p^3+m^3)+tR[p^2+pm+m^2C(k_+m)]+tM[m^2+Rm])$, where k_+ only enters in $C(k_+m)$.

The bootstrap particle filter leverages the proposal $p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1})$, with importance weights given by the likelihood in equation (1). Hence, exploiting similar arguments considered for the previous routines yields a cost $\mathcal{O}(t(p^3 + m^3) + tR(p^2 + pm) + tM[m^2 + Rm])$.

Finally, note that the cost of the extended Kalman filter (Uhlmann, 1992) is lower than the one of the particle filters since no sampling is involved, except for the Monte Carlo evaluation of the multivariate probit likelihood. In particular, at each t, one has to invert a $p \times p$ and an $m \times m$ matrix, plus computing the likelihood, which yields a total cost at t of $\mathcal{O}(t[p^3 + m^3 + Mm^2])$.

References

Albert J.H., Chib S.: Bayesian analysis of binary and polychotomous response data. J. Am. Stat. Assoc. **88**(422) 669–679 (1993)

Andrieu C., Doucet A.: Particle filtering for partially observed Gaussian state space models. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **64**(4), 827–836 (2002)

Arellano-Valle R.B., Azzalini A.: On the unification of families of skew-normal distributions. Scand. J. Stat. **33**(3), 561–574 (2006)

Arnold B.C., Beaver R.: Hidden truncation models. Sankhyā Ser. A **62**(1), 23–35 (2000)

Arnold B.C., Beaver R.J., Azzalini A., Balakrishnan N., Bhaumik A., Dey D., Cuadras C., Sarabia J.M.: Skewed multivariate models related to hidden truncation and/or selective reporting. Test **11**(1), 7–54 (2002)

Atkins A., Niranjan M., Gerding E.: Financial news predicts stock market volatility better than close price. J. Fin. Data Sci. **4**(2), 120–137 (2018)

Azzalini A., Bacchieri A.: A prospective combination of phase II and phase III in drug development. Metron **68**(3), 347–369 (2010)

Azzalini A., Capitanio A.: Statistical applications of the multivariate skew normal distribution. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **61**(3), 579–602 (1999)

Azzalini A., Capitanio A.: The Skew-normal and Related Families. Cambridge University Press. (2014)

Azzalini A., Dalla Valle A.: The multivariate skew-normal distribution. Biometrika **83**(4), 715–726 (1996)

Botev Z.: The normal law under linear restrictions: simulation and estimation via minimax tilting. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **79**(1), 125–148 (2017)

Carlin B.P., Polson N.G., Stoffer D.S.: A Monte Carlo approach to non-normal and nonlinear state-space modeling. J. Am. Stat. Assoc. **87**(418), 493–500 (1992)

Chib S., Greenberg E.: Analysis of multivariate probit models. Biometrika **85**(2), 347–361 (1998)

Chopin N., Ridgway J.: Leave Pima indians alone: Binary regression as a benchmark for Bayesian computation. Stat. Sci. **32**(1), 64–87 (2017)

Deligiannidis G., Doucet A., Rubenthaler S.: Ensemble rejection sampling. arXiv preprint arXiv:200109188 (2020)

Doucet A., Johansen A.M.: A tutorial on particle filtering and smoothing: fifteen years later. In: Handbook of Nonlinear Filtering 12, 656–704 (2009)

Doucet A., Godsill S., Andrieu C.: On sequential Monte Carlo sampling methods for Bayesian filtering. Stat. Comput. **10**(3), 197–208 (2000)

Doucet A., De Freitas N., Gordon N.: Sequential Monte Carlo Methods in Practice. Springer (2001)

Durante D.: Conjugate Bayes for probit regression via unified skew-normal distributions. Biometrika **106**(4), 765–779 (2019)

Durbin J., Koopman S.J.: Time Series Analysis by State Space Methods. Oxford University Press (2012)

Fasano A., Durante D.: A class of conjugate priors for multinomial probit models which includes the multivariate normal one. arXiv preprint arXiv:200706944 (2021)

Gelman A., Jakulin A., Pittau M.G., Su Y.S.: A weakly informative default prior distribution for logistic and other regression models. Ann. Appl. Stat. **2**(4), 1360–1383 (2008)

González-Farías G., Domínguez-Molina A., Gupta A.K.: Additive properties of skew normal random vectors. J. Stat. Plan. Infer. **126**(2), 521–534 (2004)

Gordon N.J., Salmond D.J., Smith A.F.: Novel approach to nonlinear/non-Gaussian Bayesian state estimation. IEE Proc-F. **140**(2), 107–113 (1993)

Gupta A.K., González-Farías G., Domínguez-Molina A.: A multivariate skew normal distribution. J. Multivariate Anal. **89**(1), 181–190 (2004)

Gupta A.K., Aziz M.A., Ning W.: On some properties of the unified skew-normal distribution. J. Stat. Theor. Pract. **7**(3), 480–495 (2013)

Horrace W.C.: Some results on the multivariate truncated normal distribution. J. Multivariate Anal. **94**(1), 209–221 (2005)

Johndrow J.E., Smith A., Pillai N., Dunson D.B.: MCMC for imbalanced categorical data. J. Am. Stat. Assoc. **114**(527), 1394–1403 (2019)

Julier S.J., Uhlmann J.K.: New extension of the Kalman filter to nonlinear systems. In: Proceedings SPIE 3068, Signal Processing, Sensor Fusion, and Target Recognition, pp. 182–194 (1997)

Kalman R.E.: A new approach to linear filtering and prediction problems. J. Basic Eng. **82**(1), 35–45 (1960)

Kara Y., Boyacioglu M.A., Baykan ÖK.: Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul stock exchange. Expert Syst. Appl. **38**(5), 5311–5319 (2011)

Keane M.P., Wolpin K.I.: Empirical applications of discrete choice dynamic programming models. Rev. Econ. Dynam. **12**(1), 1–22 (2009)

Kim K., Han I. Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. Expert Syst. Appl. **19**(2), 125–132 (2000)

Kitagawa G.: Monte Carlo filter & smoother for non-Gaussian nonlinear state space models. J. Comput. Gr. Stat. **5**(1), 1–25 (1996)

Lin M., Chen R., Liu J.S.: Lookahead strategies for sequential Monte Carlo. Stat. Sci. **28**(1):69–94 (2013)

Liu J., Chen R.: Sequential Monte Carlo methods for dynamic systems. J. Am. Stat. Assoc. **93**(443), 1032–1044 (1998)

Liu X., Daniels M.J., Marcus B.: Joint models for the association of longitudinal binary and continuous processes with application to a smoking cessation trial. J. Am. Stat. Assoc. **104**(485), 429–438 (2009)

MacDonald I.L., Zucchini W.: Hidden Markov and Other Models for Discrete-Valued Time Series. CRC Press (1997)

Pakman A., Paninski L.: Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. J. Comput. Gr. Stat. **23**(2), 518–542 (2014)

Petris G., Petrone S., Campagnoli P.: Dynamic Linear Models with R. Springer (2009)
Pitt M.K., Shephard N.: Filtering via simulation: Auxiliary particle filters. J. Am. Stat. Assoc. **94**(446), 590–599 (1999)
Shephard N.: Partial non-Gaussian state space. Biometrika **81**(1), 115–131 (1994)
Soyer R., Sung M.: Bayesian dynamic probit models for the analysis of longitudinal data. Comput. Stat. Data Anal. **68**, 388–398 (2013)
Uhlmann J.K.: Algorithms for multiple-target tracking. Am. Sci. **80**(2), 128–141 (1992)
Villani C.: Optimal Transport: Old and New. Springer Science & Business Media (2008)
West M., Harrison J.: Bayesian Forecasting and Dynamic Models. Springer Science & Business Media (2006)