Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Semi-Complete Data Augmentation for Efficient State Space Model Fitting

Agnieszka Borowska & Ruth King

To cite this article: Agnieszka Borowska & Ruth King (2022): Semi-Complete Data Augmentation for Efficient State Space Model Fitting, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2022.2077350

To link to this article: https://doi.org/10.1080/10618600.2022.2077350

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2022, VOL. 00, NO. 0, 1–17
https://doi.org/10.1080/01621459.2022.2077350

Semi-Complete Data Augmentation for Efficient State Space Model Fitting

Agnieszka Borowska[a] and Ruth King[b]

[a]School of Mathematics and Statistics, University of Glasgow, Glasgow, UK; [b]School of Mathematics and Maxwell Institute for Mathematical Sciences, University of Edinburgh, Edinburgh, UK

ABSTRACT

We propose a novel efficient model-fitting algorithm for state space models. State space models are an intuitive and flexible class of models, frequently used due to the combination of their natural separation of the different mechanisms acting on the system of interest: the latent underlying system process; and the observation process. This flexibility, however, often comes at the price of more complicated model-fitting algorithms due to the associated analytically intractable likelihood. For the general case a Bayesian data augmentation approach is often employed, where the true unknown states are treated as auxiliary variables and imputed within the MCMC algorithm. However, standard "vanilla" MCMC algorithms may perform very poorly due to high correlation between the imputed states and/or parameters, often leading to model-specific bespoke algorithms being developed that are nontransferable to alternative models. The proposed method addresses the inefficiencies of traditional approaches by combining data augmentation with numerical integration in a Bayesian hybrid approach. This approach permits the use of standard "vanilla" updating algorithms that perform considerably better than the traditional approach in terms of improved mixing and lower autocorrelation, and has the potential to be incorporated into bespoke model-specific algorithms. To demonstrate the ideas, we apply our semi-complete data augmentation algorithm to different application areas and models, leading to distinct implementation schemes and improved mixing and demonstrating improved mixing of the model parameters. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received January 2020
Accepted May 2022

KEYWORDS
Bayesian inference; Data augmentation; Effective sample size; Markov chain Monte Carlo; Numerical integration

1. Introduction

Inference about a latent state governing the dynamics of the system under study given only the observed noisy data is of interest in many contexts, for example, in applied statistics, ecology, engineering, or economics. A very intuitive way of describing such problems is provided by latent process models, also known as state space models (SSM), see Durbin and Koopman (2012) and West and Harrison (1997). Such models are frequently used due to the combination of their natural separation of the different mechanisms acting on the system of interest: the (unobserved) underlying system process; and the observation process. Considering each distinct process separately simplifies the model specification process and provides a very flexible modeling framework. This flexibility, however, typically comes at the price of substantially more complicated fitting of such models to data as for the general nonlinear non-Gaussian SSM the associated likelihood is analytically intractable. Only in certain circumstances the associated likelihood can be calculated explicitly: for linear Gaussian systems the likelihood can be obtained by applying the Kalman filter; for hidden Markov models (HMMs) with a discrete state space the likelihood may admit a closed-form but may become infeasible for a large number of states. In this article we focus on models for which the likelihood is intractable or for which it may be unfeasible to compute explicitly.

Dominant approaches to intractable likelihood problems include: (i) numerical or Monte Carlo integration to estimate the observed (or marginal) data likelihood; and (ii) data augmentation (DA), based on the *complete* (or joint) *data likelihood* of the observed and the imputed unobserved states, see Tanner and Wong (1987). Approach (i) includes sequential Monte Carlo (SMC) methods, see Doucet, de Freitas, and Gordon (2001), which can be used for parameter estimation within a standard Markov chain Monte Carlo (MCMC) algorithm (i.e., particle MCMC, Andrieu, Doucet, and Holenstein 2010). In general, numerical integration is efficient for low dimensional systems. Alternatively (ii) DA have become standard for inference for SSMs within a Bayesian framework, see Frühwirth-Schnatter (1994b, 2004). DA treats the true unknown states as auxiliary variables and imputes them within an MCMC algorithm. However, "vanilla" MCMC methods may perform very poorly due to high correlation between the imputed states and/or parameters, see Hobert, Royand, and Robert (2011) and the references therein.

A range of algorithms have been developed to improve the sampling efficiency for SSMs. One class of techniques includes reparameterizations for more efficient sampling, relating to centered and noncentered parameterizations described in Papaspiliopoulos, Roberts, and Sköld (2007). For instance, Strickland, Martin, and Forbes (2008) examine the effect of particular types of reparameterization in two specific

CONTACT Agnieszka Borowska *Agnieszka.Borowska@glasgow.ac.uk* School of Mathematics Statistics, University of Glasgow, Glasgow, UK.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JCGS*.

non-Gaussian state space models (the stochastic volatility, SV, model and the stochastic conditional duration model); while Kastner and Frühwirth-Schnatter (2014) and Bitto and Frühwirth-Schnatter (2019) propose to combine two DA schemes based the ancillarity-sufficiency interweaving strategy (ASIS) developed by Yu and Meng (2011). Alternatively, a group of subsampling MCMC methods aims to speed up the computations by taking a subsample of the data in each MCMC iteration, see Maclaurin and Adams (2014), Bardenet, Doucet, and Holmes (2017), Quiroz et al. (2019) and references therein.

In this article we propose a novel model-fitting algorithm to circumvent the inefficiencies of DA by using a combination of imputation and numerical integration within a Bayesian hybrid approach. In particular, we consider the case where the states are ordered, but permit the states to be discrete or continuous valued. Related ideas have been proposed before, for example, in the context of Bayesian forecasting, Harrison and Stevens (1976) approximate the posterior distribution of a SSM using discretization which they refer to as "multi-process filtering"; Frühwirth-Schnatter (1994a) uses Gauss-Hermite integration to develop an approximate Kalman filter for SSMs. More recently, Strickland et al. (2009) use a univariate representation of multivariate SMMs for their efficient MCMC estimation. The idea underlying our hybrid approach is to combine the "good" aspects of both methods by minimizing the problems that arise for each, that is, highly correlated latent states for DA and the curse of dimensionality for numerical integration. To this end, we use the structure of the unknown states and split them into two types: auxiliary variables, imputed within the MCMC algorithm; and "integrable" states, numerically integrated out within the likelihood function. We refer to this likelihood as the *semi-complete data likelihood*.

Since the resulting semi-complete data likelihood may still be analytically intractable, we propose to numerically approximate it by embedding it into an HMM form. To this end, we first notice that quadratures (with fixed or adaptive nodes) can be seen as "binning" of similar values of the integrated state vector within specified ranges ("bins"), which can then be interpreted as states of a (finite-dimensional) first-order HMM. Second, we use the specified states of the underlying Markov chain to express the approximation to the semi-complete data likelihood in a general form as the likelihood of the associated HMM. We note that this approximation approach assumes that the SSM of interests belongs to one of the following three important classes: discrete multivariate (e.g., HMMs and factor HMMs), integer count multivariate (illustrated in Section 5.2), and continuous multivariate (illustrated in Section 5.1).

There are several novel aspects of the proposed methodology. First, we specify the general semi-complete data likelihood methodology for the large class of SSMs. We note that a similar semi-complete data likelihood approach was applied to the particular case of capture-recapture data (King et al. 2016). Second, we propose a general framework for approximating the integrals that arise within the SSMs based on fixed and adaptive grids, generalizing the idea of Langrock, MacDonald, and Zucchini (2012b) who developed fixed-grid approximation for the special case of frequentist estimation for the stochastic volatility model. Third, we express the approximation to the semi-complete data likelihood as a likelihood of the associated HMM, where the value of the integrated state vector within specified range are interpreted as states of an HMM. Finally, we consider efficient numerical approximations for both continuous and discrete variables.

We note that the novel approach we propose does not rely on any particular MCMC updating scheme but on the specification of the auxiliary variables. It is the choice of the analyst what specific sampler is used for updating these auxiliary variables. In our experiments in Section 5 we apply the random walk Metropolis–Hastings (RW-MH) algorithm as it is often used in practice and it acts as a "vanilla" MCMC algorithm (see Marin and Robert 2007, chap. 4). We note, however, that one may use more advanced samplers such as Hamiltonian Monte Carlo (HMC, Neal 2011) or the No-U-Turn Sampler (NUTS, Hoffman and Gelman 2014) for continuous states or even probabilistic programming languages such as Stan (Stan Development Team 2019) or JAGS (Plummer 2017). Finally, we emphasize that the proposed method serves as a part of the analysts' toolbox of techniques, providing a new building block for the construction of complex algorithms, in which it is combined with other algorithms to achieve further efficiency gains. For example, in practice, reparameterization strategies such as centered/non-centered parameterizations or ASIS can be immediately combined with our proposed approach as these techniques do not affect the dependence structure of the model.

The structure of the article is as follows. Section 2 presents the general SSM specification and the standard approaches to their fitting. Section 3 introduces the proposed semi-complete data augmentation (SCDA) approach, while Section 4 develops a general HMM-based likelihood approximation. We consider the performance of our SCDA method in Section 5, based on empirical applications related to the unobserved component stochastic volatility model and abundance estimation. Section 6 concludes with a discussion.

2. State Space Models

Consider a state space model given by ($t = 1, \ldots, T$)

$$\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\theta} \sim p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\theta}), \tag{1}$$

$$\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{\theta} \sim p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{\theta}), \tag{2}$$

$$\boldsymbol{x}_0|\boldsymbol{\theta} \sim p(\boldsymbol{x}_0|\boldsymbol{\theta}). \tag{3}$$

Let $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$ (with $\boldsymbol{y}_t = (y_{1,t}, \ldots, y_{M,t})^T$ potentially multivariate, of dimension $M < \infty$, with $y_{m,t} \in \mathcal{Y}_m$) denote a time series of observations of length T, $\boldsymbol{x} = (\boldsymbol{x}_0, \ldots, \boldsymbol{x}_T)$ a series of latent states (with $\boldsymbol{x}_t = (x_{1,t}, \ldots, x_{D,t})^T$ potentially multivariate, of dimension $D < \infty$, with $x_{d,t} \in \mathcal{X}_d$) and $\boldsymbol{\theta}$ the model parameters for which we put a prior $p(\boldsymbol{\theta})$. To simplify notation, we use p as a general symbol for a probability mass function (pmf) or a probability density function (pdf), possibly conditional.

The system process describing the evolution of $\boldsymbol{x}_t$, the true (unobserved) state of the system over time is defined by distribution (2). The observation process which generates $\boldsymbol{y}_t$, the observed data given the true underlying states, is specified by distribution (1). This separation of the different mechanisms acting on the system of interest makes SSM a very intuitive and flexible description of time series data. Figure 1 graphically

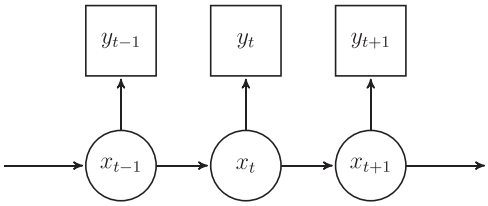


Figure 1. A graphical representation of the general first-order SSM. Squares represent the observed data, circles—the unknown variables.

presents the dependencies between states and observations in the SSM. An extensive discussion of SSMs is provided by Durbin and Koopman (2012) and Cappé, Moulines, and Ryden (2006), where the latter authors refer to this class of models as hidden Markov models (HMM). Throughout this article we refer to the general class of models as SSMs, and restrict the term HMMs to models with discrete states only (i.e., $\dim(\mathcal{X}_d) < \infty$); this is the convention used by for example, Zucchini, MacDonald, and Langrock (2016).

Modelling flexibility of SSMs is, however, often offset with the issue of estimating $\boldsymbol{\theta}$, the associated model parameters. The *observed data likelihood* for the system (1)–(3)

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{\theta}) &= \int p(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})d\boldsymbol{x} \\ &= \int p(\boldsymbol{x}_0|\boldsymbol{\theta}) \prod_{t=1}^{T} p(y_t|\boldsymbol{x}_t,\boldsymbol{\theta}) p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{\theta}) d\boldsymbol{x}_0 d\boldsymbol{x}_1 \ldots d\boldsymbol{x}_T, \end{aligned} \tag{4}$$

is typically not available in closed form due to the necessary integration over the latent variables. This is despite the tractability of $p(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})$, the joint distribution of the data and the auxiliary variables, often referred to as the *complete data likelihood*.

For models with discrete states the observed data likelihood is the likelihood of an HMM, where the states of the chain correspond to distinct values of the latent process, and the transition matrix can be derived from the transition equation (2). This likelihood can be efficiently calculated using the forward algorithm (see Zucchini, MacDonald, and Langrock 2016). However, for systems with multiple processes or processes with a large set of possible states this can lead to this approach being unfeasible.

To overcome the problem of the intractable likelihood, a DA technique is commonly adopted, see Tanner and Wong (1987), Frühwirth-Schnatter (1994b, 2004), and Hobert (2011). The unknown states $\boldsymbol{x}$ are treated as auxiliary variables and imputed leading to a closed-form complete data likelihood (5) which in a Bayesian framework is used to construct the joint posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{x}$ in (6):

$$p(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}) = p(\boldsymbol{x}_0|\boldsymbol{\theta}) \prod_{t=1}^{T} p(y_t|\boldsymbol{x}_t,\boldsymbol{\theta}) p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{\theta}), \tag{5}$$

$$p(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{6}$$

An MCMC algorithm (or other) can be used to obtain a sample from (6), from which we obtain $p(\boldsymbol{\theta}|\boldsymbol{y})$, the marginal posterior of $\boldsymbol{\theta}$. In practice the random walk Metropolis–Hastings (RW-MH) algorithm is often used and it acts as a "vanilla" MCMC algorithm (see Marin and Robert 2007, chap. 4).

However, this approach often results in posterior draws being highly correlated, indicating poor mixing and hence low efficiency of MCMC algorithms. This is particularly the case for SSMs which impose a strong dependence structure on the latent variables and parameters. Single-update algorithms can perform especially poorly and block updates can lead to improved mixing. However, the latter often require defining an appropriate partition of the states and parameters into blocks and specifying an efficient proposal distributions for each block. Thus, bespoke codes often need to be written dependent on model and data.

3. Semi-Complete Data Augmentation

We propose to combine DA with numerical integration within a Bayesian hybrid framework, which we call *semi-complete data augmentation* (SCDA). A key idea is to separate the latent state $\boldsymbol{x}$ into two components $\boldsymbol{x} = (\boldsymbol{x}_{\text{aug}}^T, \boldsymbol{x}_{\text{int}}^T)^T$. We will refer to $\boldsymbol{x}_{\text{int}}$ and $\boldsymbol{x}_{\text{aug}}$ as the "integrated" states and the "augmented" states, respectively. We specify the *semi-complete data likelihood* $p(\boldsymbol{y},\boldsymbol{x}_{\text{aug}}|\boldsymbol{\theta})$ as follows

$$p(\boldsymbol{y},\boldsymbol{x}_{\text{aug}}|\boldsymbol{\theta}) = \int p(\boldsymbol{y}|\boldsymbol{x}_{\text{aug}},\boldsymbol{x}_{\text{int}},\boldsymbol{\theta}) p(\boldsymbol{x}_{\text{aug}},\boldsymbol{x}_{\text{int}}|\boldsymbol{\theta}) d\boldsymbol{x}_{\text{int}}. \tag{7}$$

The joint posterior distribution of the parameters and augmented states is given by

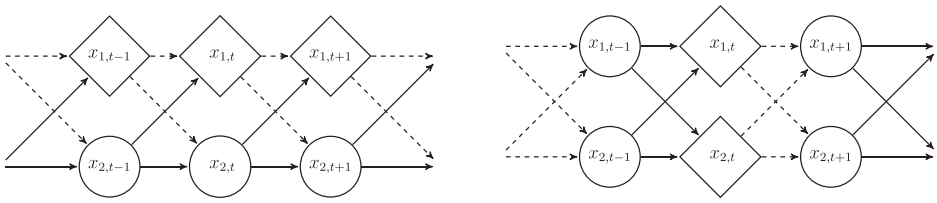
$$p(\boldsymbol{x}_{\text{aug}},\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y},\boldsymbol{x}_{\text{aug}}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{x}_{\text{aug}},\boldsymbol{\theta})p(\boldsymbol{x}_{\text{aug}}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{8}$$

We note that the approach of King et al. (2016), who propose a Bayesian hybrid approach for the particular case of capture-recapture data, is a special case of our general approach proposed here.

3.1. Specification of x_{aug}, x_{int}

Consider a series of latent states $\boldsymbol{x} = \{\boldsymbol{x}_t\}_{t=0}^T$ of length $T+1$, where the state at time t is D dimensional: $\boldsymbol{x}_t = (x_{1,t},\ldots,x_{D,t})^T$, for $t = 0,1,\ldots,T$. We want to integrate out D_{int} dimensions of the state at time points T_{int}, where $D_{\text{int}} \subset \{1,\ldots,D\}$ and $T_{\text{int}} \subset \{0,1,\ldots,T\}$ are "suitably" chosen subsets of dimension and time indices, respectively. Such a "suitable" specification of D_{int} and T_{int} depends on the dependence structure of the model so that the associated integral can be efficiently calculated. For instance, it can be low dimensional or it can be reduced to a product of low-dimensional integrals. We denote the compliments of both subsets D_{aug} and T_{aug}, respectively. We also let T_{int}^+ and T_{aug}^+ denote the corresponding sets without the initial observations, that is, excluding time $t = 0$, and we set $T^* = |T_{\text{int}}^+|$. The "integrated" and "augmented" states are then defined as the partition of $\boldsymbol{x}$ into $\boldsymbol{x}_{\text{int}} = \{x_{d,t}\}_{d\in D_{\text{int}},t\in T_{\text{int}}}$ and $\boldsymbol{x}_{\text{aug}} = \{x_{d,t}\}_{d\in D_{\text{aug}},t\in T_{\text{aug}}}$, where their corresponding elements at time t are denoted by $\boldsymbol{x}_{\text{int},t} = \{x_{d,t}\}_{d\in D_{\text{int}}}$ and $\boldsymbol{x}_{\text{aug},t} = \{x_{d,t}\}_{d\in D_{\text{aug}}}$, respectively. As example, consider the two following schemes.

(i) "*Horizontal*" *integration*: for example, for a $D = 2$ dimensional state we integrate out the second state at all time periods, so that $D_{\text{int}} = \{2\}$ (and hence, $D_{\text{aug}} = \{1\}$),



(a) **Horizontal integration.** (b) **Vertical integration**

Figure 2. Two examples of an integration/augmentation scheme. Diamonds represent the imputed states, circles—the integrated states. Dashed lines used for the relations *from* the imputed (known) states.

and $T_{\text{int}} = \{0, 1, \ldots, T\}$ (and hence $T_{\text{aug}} = T_{\text{int}}$), see Figure 2(a). We use this scheme in the lapwings data application in Section 5.2.

(ii) "*Vertical*" *integration*: for example, all D states are integrated out at odd time periods, so that $D_{\text{int}} = \{1, \ldots, D\}$ and $T_{\text{int}} = \{2t + 1\}_{t=0}^{\lfloor T/2 \rfloor}$ (and hence, $T_{\text{aug}} = \{2t\}_{t=0}^{\lfloor T/2 \rfloor}$), see Figure 2(b). We use this scheme in the unobserved component stochastic volatility (UCSV) model application in Section 5.1, for $D = 1$ dimensional state.

As we can see, in general T_{int} and T_{aug} do not need to be equal and their elements may not be consecutive numbers. However, we would like to iterate over both sets using the same index. Therefore, we introduce two functions $\tau(t)$ and $a(t)$ such that the image of τ is T_{int}^{+} and the image of a covers T_{aug}^{+}, both defined on $1, 2, \ldots, T^*$. We require τ to be bijective and allow a to take values in the power set of T_{aug}^{+}. The latter characteristic means that $a(t)$ can take two or more values in T_{aug}^{+} but also no value (i.e., $a(t) = \emptyset$). In the two examples above we have $\tau(t) = t$ and $a(t) = t$ for the horizontal integration (a) and $\tau(t) = 2t + 1$ and $a(t) = 2t$ for the vertical integration (b). Additionally, we specify a function for observations $o(t)$ with a similar role to τ and a, that is, allowing us to iterate over the set of observation indices $\{1, \ldots, T\}$ using the same index as to iterate over T_{int} and T_{aug}. Thus, we want the image of $o(t)$ to be $\{1, \ldots, T\}$, which may consists of elements from both T_{int} and T_{aug}. This means that we need to be able to assign multiple indices from $\{1, \ldots, T\}$ to t. Hence, we allow $o(t)$ to take values in the power set of $T_{\text{int}} \cup T_{\text{aug}}$. For illustration, consider vertical integration (b) together with conditionally independent observations $y_t|\boldsymbol{x}_t \sim p(y_t|\boldsymbol{x}_t)$. For $t = 1, 2, \ldots, T^*$ consider states in two different time periods: at $\tau(t) = 2t + 1$ for $\boldsymbol{x}_{\text{int}}$ and at $a(t) = 2t$ for $\boldsymbol{x}_{\text{aug}}$, so for each t we need to account for two different observations, $y_{\tau(t)}$ and $y_{a(t)}$. This means that $o(t) = \{2t, 2t + 1\}$ for $t \geq 1$. In this case we also need to account for y_1 so we additionally specify $o(t) = \{2t + 1\}$ for $t = 0$. For horizontal integration given in (a) $T_{\text{int}} = T_{\text{aug}}$, hence, $o(t) = t$.

In order to identify conditionally independent latent states to "integrate out," one can use the graphical structure of the problem: Figure 1 can be seen as an directed acyclic graph (DAG), for which the literature on dynamic Bayesian networks (see Murphy 2002) provides insights regarding the impact of conditioning on a certain node (*d-separation*). In the context of particle filters Doucet et al. (2000) note that the "tractable structure" of some state space models might by analytically marginalized out given imputed other nodes.

3.2. Approximate Marginal Likelihood

The semi-complete data likelihood $p(\boldsymbol{y}, \boldsymbol{x}_{\text{aug}}|\boldsymbol{\theta})$ in the joint posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{x}_{\text{aug}}$ in (8) may still be analytically intractable. In this case we can estimate it using simulation-based techniques. Consider a sample of length N of unknown variables of interest (i.e., $\boldsymbol{\theta}$ and $\boldsymbol{x}_{\text{aug}}$). Here, N is the number of points used for integration: for a deterministic integration it is the number of evaluation points, for a stochastic, that is, Monte Carlo (MC), integration it is the number of draws. We use such a sample to compute $\hat{p}_N(\boldsymbol{y}, \boldsymbol{x}_{\text{aug}}|\boldsymbol{\theta})$, the N-sample estimator of the semi-complete data likelihood, and consequently to approximate the posterior distribution as $\hat{p}_N(\boldsymbol{x}_{\text{aug}}, \boldsymbol{\theta}|\boldsymbol{y}) \propto \hat{p}_N(\boldsymbol{y}, \boldsymbol{x}_{\text{aug}}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. We set $\hat{p}_N(\boldsymbol{y}, \boldsymbol{x}_{\text{aug}}|\boldsymbol{\theta})$ such that $\hat{p}_N(\boldsymbol{y}, \boldsymbol{x}_{\text{aug}}|\boldsymbol{\theta}) \overset{N\to\infty}{\to} p(\boldsymbol{y}, \boldsymbol{x}_{\text{aug}}|\boldsymbol{\theta})$, so that $\hat{p}_N(\boldsymbol{x}_{\text{aug}}, \boldsymbol{\theta}|\boldsymbol{y}) \overset{N\to\infty}{\to} p(\boldsymbol{x}_{\text{aug}}, \boldsymbol{\theta}|\boldsymbol{y})$. Further properties of the resulting estimator depend on the approximation scheme. If it is unbiased and nonnegative, standard MCMC algorithms converge to the exact posterior distribution $p(\boldsymbol{x}_{\text{aug}}, \boldsymbol{\theta}|\boldsymbol{y})$, which follows from the pseudo-marginal argument, see Beaumont (2003), Andrieu and Roberts (2009), and Andrieu, Doucet, and Holenstein (2010). Pseudo-marginal algorithms are called "exact approximate" and we note that they are the extreme case of our approach with $\boldsymbol{x}_{\text{int}} = \boldsymbol{x}$. Whether our approximate MCMC algorithm is "exact approximate" or "just approximate" depends on whether or not $\hat{p}_N(\boldsymbol{y}, \boldsymbol{x}_{\text{aug}}|\boldsymbol{\theta})$ is an unbiased and nonnegative estimator of the marginal likelihood.

The "just approximate" algorithms, such as a quadrature, can be made arbitrarily close to the true integral by considering sufficiently many points (i.e., as $N \to \infty$). Alternatively, unbiased estimators using an MC approach might be characterized by large MC errors, particularly for a small number of samples, see, for example, Korattikara, Chen, and Welling (2014), Jacob and Thiery (2015). The choice between different likelihood approximation methods fits into the traditional discussion on the bias-variance tradeoff.

3.3. Dimensionality of Observations and States and Scalability

In the general SSM specification (1)–(3) we allow for multivariate observations $\boldsymbol{y}$, however, as observed data, the dimensionality of $\boldsymbol{y}$ does not play any particular role in the SCDA approach beyond the specification of the observation process, given the latent states, $\boldsymbol{x}$. The dimension of the latent states, $\boldsymbol{x}$, directly influences the basic DA algorithm, in terms of the number of auxiliary variables imputed within the algorithm, and hence,

associated computational implications. For the SCDA approach the dimension of $\boldsymbol{x}$ will impact both the specification of the set of imputed and integrated states (a larger dimension increases the number of possible divisions), as well as the number of auxiliary variables. In general, there remains some form of the curse of dimensionality due to the increased computational demands for increased numbers of auxiliary variables, however, for the SCDA approach there is some additional flexibility in terms of the specification of the set of integrated and imputed states. In general, feasible dimensions of $\boldsymbol{x}$ will be dependent on the specific model, associated correlation between the parameters and specific SCDA approach implemented in terms of the definition of the imputed and integrated out variables.

As an illustration, consider a time varying parameter (TVP) model, which is a linear regression model with coefficients varying over time (Durbin and Koopman 2012, sec. 3.6.1). This class of models has gained particular interest in macroeconometrics and finance as it allows for modeling the varying importance of groups of predictors over time (see, e.g., Primiceri 2005; Belmonte, Koop, and Korobilis 2014; Bitto and Frühwirth-Schnatter 2019). For $t = 1, \dots, T$, consider the following TVP model,

$$y_t = \boldsymbol{z}_t \boldsymbol{\beta}_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_t^2), \tag{9}$$

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{u}_t, \qquad \boldsymbol{u} \sim \mathcal{N}_d(0, \boldsymbol{\Omega}), \tag{10}$$

where y_t is a univariate observation and $\boldsymbol{z}_t = (z_{1,t}, \dots, z_{d,t})$ is a regressors vector, which can contain the lagged values of y_t (i.e., $y_{t-1}, y_{t-2}, \dots$). We assume that $\boldsymbol{\Omega}$ is diagonal, so that, $\boldsymbol{\Omega} = \text{diag}(\omega_{1,t}, \dots, \omega_{d,t})$, and the system is initialized with $\boldsymbol{\beta}_0 \sim \mathcal{N}_d(\boldsymbol{\beta}, \boldsymbol{\Omega})$, where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)$. For simplicity we assume homoscedastic observation errors, $\sigma_t^2 \equiv \sigma^2$, $t = 1, \dots, T$.

The time varying coefficients represent the d-dimensional state vector of the TVP model (9)–(10), in which individual states $\beta_{i,t}$, $i = 1, \dots, d$, are mutually independent. Thus, when updating $\beta_{i,t}$ with for example, RW-MH, the "update-relevant" term in the complete data likelihood is $p(\beta_{i,t}|\beta_{i,t-1})p(y_t|\beta_{1,t}, \dots, \beta_{d,t})p(\beta_{i,t+1}|\beta_{i,t})$, so that the other states $\beta_{j,t}$, $j \neq i$, appear only in the observation distribution. To obtain a SCDA scheme with one-dimensional integrals in which all the d states are involved in the integration[1] we may set

$$\boldsymbol{x}_{\text{int}} := \{\beta_{i,t}, t = nd + i, n \in \mathbf{N}, i = 1, \dots, d\},$$

that is, we integrate $\beta_{i,t}$ every dth time point starting from time i. For the time t' update of $\beta_{i,t'}$, $t' \neq nd + i$, only up to three disjoint one-dimensional integrals are required: potentially one per each of $p(\beta_{i,t}|\beta_{i,t-1})$ and $p(\beta_{i,t+1}|\beta_{i,t})$ terms (where the ith state itself is integrated), and one which involves the state that is integrated at t' (i.e. the mod(t', d)th state).

Extending the basic model (9)–(10) to allow for stochastic volatility of the observation errors ε_t can be treated in a similar manner. If we model $\sigma_t^2 = \exp(h_t)$, with h_t being either a random walk or AR(1) process, independent from $\boldsymbol{\beta}_t$, then this extra new state can be included in $\boldsymbol{x}_{\text{int}}$ and we would integrate $\beta_{i,t}$ every $(d + 1)$th time point starting from time i, and h_t at every $d + 1$ time point starting from time $d + 1$. We provide an empirical illustration of this scheme for a simple TVP model with $d = 1$ and $z_t \equiv 1$, $t = 1, \dots, T$ in Section 5.1.3.

[1] A naive scheme with one-dimensional integrals would impute all but one state, say the ith, and apply the vertical integration from Figure 2(b) to the ith state.

4. Approximations for MCMC Sampling

In this section we assume that the states of the SSM are ordered (we discuss the unordered case in Section 6). In particular, we restrict our attentions to the following three important classes of SSM: discrete multivariate (e.g., HMMs and factor HMMs), integer count multivariate (as in Section 5.2), and continuous multivariate (as in Section 5.1).

We focus on the case when $\hat{p}_N(\boldsymbol{y}, \boldsymbol{x}_{\text{aug}}|\boldsymbol{\theta})$ can be obtained as a product of one dimensional integrals. This assumption is less restrictive than it may appear at first: the choice of the auxiliary variables can often be made such that this condition is satisfied. There exist several methods to numerically estimate a single one dimensional integral including: (i) quadrature with fixed nodes; (ii) quadrature with adaptive nodes; (iii) stochastic (MC) integration. Approaches (i) and (ii) can be seen as "binning" of similar values of the integrated state vector within specified ranges ("bins"), which can then be interpreted as states of a (finite-dimensional) first-order HMM. In the context of bins of equal widths such an approach has been successfully applied for example, by Langrock et al. (2012a), Langrock, MacDonald, and Zucchini (2012b), and Langrock and King (2013). For approach (iii) the resulting estimator of the complete data likelihood is unbiased and an "exact approximate" algorithm is obtained. We note that in low dimensions all these methods are feasible, however, we focus on methods based on the two former approaches as they provide an intuitive interpretation in terms of state transition probabilities and conditional (augmented) observation distributions. There are two cases when such an approximation might be particularly useful. First, when the state vector is discrete but of a large size grouping of its elements into "bins" helps to reduce the size of the problem. Second, for continuous states any form of numerical integration basically reduces to splitting of the state space into "bins," which can then be further combined into larger groups to increase the efficiency of the algorithm.

4.1. Approximation Bins as Hidden Markov Model States

We consider two ways to specify the bins, or quadrature points: a deterministic one, with bins of a fixed size (but varying probability of occurring), and a stochastic one, with bins of a fixed probability (but varying size). To simplify the exposition, we assume that $\boldsymbol{x}_{\text{int},\tau(t)}$ is univariate and we write $x_{\text{int},\tau(t)}$. For multivariate $\boldsymbol{x}_{\text{int},\tau(t)}$ we may consider separate bins for each integrated state dimension $d \in D_{\text{int}}$ at time $\tau(t)$. We interpret the bins as states of a latent (first-order) Markov process, which allows us to give the resulting integration/augmentation scheme an HMM embedding.

Fixed bins. A straightforward approach to binning is via bins of a fixed size as it relates to a deterministic approximation of the likelihood with a quadrature and allows for a natural HMM interpretation. Discretizing the state space to perform numerical integration dates back to Kitagawa (1987) and is discussed in

Zucchini, MacDonald, and Langrock (2016). The state space $\mathcal{X}_{\text{int}}$ of the state to be integrated out is split into B bins of length k (for integer-valued variables we assume $k \in \mathbb{N}$) and for example, the midpoints of the bins are considered for integration. Then the values that fall in a given bin are approximated by the value of the midpoint of that bin. Such an approach is used by Langrock, MacDonald, and Zucchini (2012b) to efficiently approximate the likelihood for stochastic volatility models (with continuous bins) in a classical framework.

For infinitely dimensional states, either discrete or continuous, an "allowed integration range" needs to be specified. For instance, for a normal variable this means setting a lower and an upper bound for the integration b_0 and b_B, while for a Poisson variable only of an upper bound b_B since $b_0 = 0$ in this case. We divide the resulting domain into intervals:

$$\underbrace{[b_0, \dots, b_1)}_{\mathcal{B}_1,\text{ bin } 1}, \underbrace{[b_1, \dots, b_2)}_{\mathcal{B}_2,\text{ bin } 2}, \dots, \underbrace{[b_{j-1}, \dots, b_j)}_{\mathcal{B}_j,\text{ bin } j}, \dots,$$
$$\underbrace{[b_{B-1}, \dots, b_B)}_{\mathcal{B}_B,\text{ bin } B}, \quad b_i - b_{i-1} = k, \quad i = 1, \dots, B.$$

For continuous variables $\mathcal{B}_i$ is simply a continuous interval of length k, while for discrete variables it consists of k subsequent integers, for example, for a Poisson variable we have $\mathcal{B}_i = \{ik, \dots, (i+1)k\}$. We specify the midpoints of the bins as $b_i^* = \frac{b_{i-1}+b_i}{2}$ (for integer-valued variables rounding is required for even k).

We define $\{z_t\}$, $t \in 1, \dots, T^*$, as a B-state, discrete-time (not necessarily homogeneous) Markov chain with transition probabilities $\gamma_{jk,t} = \mathbb{P}(z_t = k | z_{t-1} = j)$ defined as

$$\gamma_{jk,t} := \mathbb{P}(x_{\text{int},\tau(t)} \in \mathcal{B}_k | x_{\text{int},\tau(t-1)} \in \mathcal{B}_j, \boldsymbol{x}_{\text{aug},a(t-1)}).$$

The transition of $z_{t-1} = j$ to $z_t = k$ is equivalent to $x_{\text{int},\tau(t)}$ belonging to bin k given $x_{\text{int},\tau(t-1)}$ was in bin j (and $\boldsymbol{x}_{\text{aug},a(t-1)}$). For computationally intensive probabilities we can further approximate these as $\tilde{\gamma}^*_{jk,t} := p(b_k^* | b_j^*, \boldsymbol{x}_{\text{aug},a(t-1)})$, which for discrete variables means $\mathbb{P}(x_{\text{int},\tau(t)} = b_k^* | x_{\text{int},\tau(t-1)} = b_j^*, \boldsymbol{x}_{\text{aug},a(t-1)})$. To get the valid probability values we normalize the transition probabilities as $\gamma^*_{jk,t} := \tilde{\gamma}^*_{jk,t} / \sum_{c=1}^{B} \tilde{\gamma}^*_{jc,t}$. Notice that this corresponds to treating the values in a bin uniformly. Alternatively, we can compute the transition probabilities between bins directly as follows

$$\mathbb{P}(x_{\text{int},\tau(t)} \in \mathcal{B}_k | x_{\text{int},\tau(t-1)} \in \mathcal{B}_j, \boldsymbol{x}_{\text{aug},a(t-1)})$$
$$\propto \int_{\mathcal{B}_k \times \mathcal{B}_j} p(x_{\text{int},\tau(t)} | x_{\text{int},\tau(t-1)}, \boldsymbol{x}_{\text{aug},a(t-1)}) dx_{\text{int},\tau(t-1)} dx_{\text{int},\tau(t)}.$$

However, such an analytical integration will typically be possible only in simple cases.

Adaptive bins. An alternative approach is to use adaptive intervals which do not require a bounded integration range. This can be done by transforming the variable of interest to the $[0, 1]$ range by applying a cdf. The bins are specified on the $[0, 1]$ interval and their limits or midpoints are transformed back to obtain the values of the original variable. In particular, quantiles of the distribution associated with the variable of interest can be used.

Suppose $x_{\text{int},\tau(t)} \sim p(\vartheta_{\tau(t)})$, $\tau(t) \in T_{\text{int}}$, where $\vartheta_{\tau(t)}$ is a vector of possibly time varying parameters, with the corresponding cdf $F(\vartheta_{\tau(t)})$. Consider a vector of $B+1$ quantiles $\boldsymbol{q} = [q_0, q_2, \dots, q_B]$. The corresponding B mid-quantiles $\boldsymbol{q}^* = [q_1^*, q_2^*, \dots, q_B^*]$ are given by $q_i^* = \frac{q_{i-1}+q_i}{2}$ (e.g., $\boldsymbol{q} = [0.0, 0.1, 0.2, \dots, 1.0]$ and $\boldsymbol{q}^* = [0.05, 0.15, \dots, 0.95]$). For $F(\vartheta_t)$ continuous and strictly monotonically increasing the bin midpoints at time t are determined by the mid-quantiles as $b_i^* = F^{-1}\left(q_i^* | \vartheta_{\tau(t)}\right)$. For discrete variables one can either use the generalized inverse distribution function, or use a continuous approximation to the associated discrete distribution.

4.2. Hidden Markov Model Likelihood

Having specified the states of the underlying Markov chain in the previous section, we aim to use them to approximate the joint semi-complete data likelihood (7) by embedding it into an HMM form (below, to ease the notation, we skip $\boldsymbol{\theta}$ in conditioning). We relate each state of the hidden Markov process with the relevant augmented states and observations. This imposes a time structure on the semi-complete data likelihood integral with respect to the "integration time" and thus allows us to cast it into a likelihood of an HMM.

Motivating example. Consider the state specification from Figure 2(a) to which we add conditionally independent observations to result in an SSM (see Appendix C.1, supplementary materials for a graphical illustration and more details). We specify $\boldsymbol{x}_{\text{aug}} = \{x_{1,t}\}_{t=0}^{T} =: \boldsymbol{x}_1$ and $\boldsymbol{x}_{\text{int}} = \{x_{2,t}\}_{t=0}^{T} =: \boldsymbol{x}_2$, which corresponds to the "horizontal" integration. Hence, we put $T_{\text{int}} = T_{\text{aug}} = \{0, 1, \dots, T\}$, $\tau(t) = t$, $a(t) = t$ and $o(t) = t$. Using the temporal dependence in this system, the semi-complete data likelihood $p(\boldsymbol{y}, \boldsymbol{x}_{\text{aug}})$ can be expressed as

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{x}_{\text{aug}}) &= p(x_{1,0}) \prod_{t=1}^{T} p(y_t | x_{1,t}) p(x_{1,t} | x_{1,t-1}) \\ &= p(x_{1,0}) \prod_{t=1}^{T^*} p(y_{o(t)} | x_{1,a(t)}) p(x_{1,a(t)} | x_{1,a(t-1)}), \end{aligned}$$

which is not tractable without integrating out $\boldsymbol{x}_2$. Hence, we marginalize over $\boldsymbol{x}_2$ and approximate the resulting integral using a quadrature with B bins $\mathcal{B}_k$, $k = 1, \dots, B$, as

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{x}_{\text{aug}}) = \int \cdots \int & p(x_{1,0}) p(x_{2,0}) \prod_{t=1}^{T^*} p(y_{o(t)} | x_{1,a(t)}, x_{2,\tau(t)}) \\ & \times p(x_{1,a(t)} | x_{1,a(t-1)}, x_{2,\tau(t-1)}) \\ & \quad p(x_{2,\tau(t)} | x_{1,a(t-1)}, x_{2,\tau(t-1)}) \\ & \quad dx_{2,\tau(T^*)} \dots dx_{2,\tau(1)} \qquad (11) \\ \approx \sum_{k_0=1}^{B} \cdots \sum_{k_{T^*}=1}^{B} & p(x_{1,0}) p(x_{2,0} \in \mathcal{B}_{k_0}) \\ & \prod_{t=1}^{T^*} p(y_{o(t)} | x_{1,a(t)}, x_{2,\tau(t)} \in \mathcal{B}_{k_t}) \\ & \times p(x_{1,a(t)} | x_{1,a(t-1)}, x_{2,\tau(t-1)} \in \mathcal{B}_{k_{t-1}}) \\ & \quad p(x_{2,\tau(t)} \in \mathcal{B}_{k_t} | x_{1,a(t-1)}, x_{2,\tau(t-1)} \in \mathcal{B}_{k_{t-1}}). \end{aligned}$$

The above approximation has a natural interpretation in terms of HMM by associating the events $x_{2,\tau(t)} \in \mathcal{B}_k$ with states of a hidden Markov process on B states. The transition matrix of this process is

$$\Gamma_t = [\mathbb{P}(x_{2,\tau(t)} \in \mathcal{B}_k | x_{1,a(t-1)}, x_{2,\tau(t-1)} \in \mathcal{B}_l)]_{k,l=1,\dots,B}, \quad (12)$$

for $t \in 1, 2, \dots, T^*$. We specify two further matrices for the "augmented data": P_t for the augmented states $\boldsymbol{x}_{\text{aug}}$ and Q_t for the real observations $\boldsymbol{y}$, as follows

$$P_t = \text{diag}\left(p(x_{1,a(t)} | x_{1,a(t-1)}, x_{2,\tau(t-1)} \in \mathcal{B}_l)\right)_{l=1,\dots,B}, \quad (13)$$

$$Q_t = \text{diag}\left(p(y_{o(t)} | x_{1,a(t)}, x_{2,\tau(t)} \in \mathcal{B}_k)\right)_{k=1,\dots,B}. \quad (14)$$

This is different compared to standard HMMs in which only the matrix for $\boldsymbol{y}$ is used. Notice that the conditioning in (12) and (13) is with respect to the previous realizations of the states, whilst for the observations in (14) it is with respect to the current states. Finally, the quadrature based approximation to the semi-complete data likelihood (11) can be expressed as

$$\hat{p}_B(\boldsymbol{y}, \boldsymbol{x}_{\text{aug}}) = p(x_{1,0}) \boldsymbol{u}_0 \left(\prod_{t=1}^{T^*} P_t \Gamma_t Q_t \right) \mathbf{1}, \quad (15)$$

where $\boldsymbol{u}_0 = (\mathbb{P}(x_{2,0} \in \mathcal{B}_1), \dots, \mathbb{P}(x_{2,0} \in \mathcal{B}_B))$ is the initial distribution of the Markov chain.

General formulation. The generic matrices of the HMM-based approximation are

$$\Gamma_t = [\mathbb{P}(\boldsymbol{x}_{\text{int},\tau(t)} \in \mathcal{B}_k | \boldsymbol{x}_{\text{int},\tau(t-1)} \in \mathcal{B}_l, \boldsymbol{x}_{\text{aug},a(t-1)})]_{k,l=1,\dots,B},$$

$$P_t = \text{diag}\left(p(\boldsymbol{x}_{\text{aug},a(t)} | \boldsymbol{x}_{\text{int},\tau(t-1)} \in \mathcal{B}_l), \boldsymbol{x}_{\text{aug},a(t-1)}\right)_{l=1,\dots,B},$$

$$Q_t = \text{diag}\left(p(\boldsymbol{y}_{o(t)} | \boldsymbol{x}_{\text{int},\tau(t)} \in \mathcal{B}_k, \boldsymbol{x}_{\text{aug},a(t)})\right)_{k=1,\dots,B},$$

for $t \in 1, 2, \dots, T^*$ and lead to the following form of the HMM approximation

$$\hat{p}_B(\boldsymbol{y}, \boldsymbol{x}_{\text{aug}}) = p(x_{1,0}) \boldsymbol{u}_0 Q_0 \left(\prod_{t=1}^{T^*} P_t \Gamma_t Q_t \right) \mathbf{1}, \quad (16)$$

which differs from (15) by including $Q_0 := \text{diag}\left(p(y_{o(0)} | x_{\text{int},0} \in \mathcal{B}_k^*)\right)_{k=1,\dots,B}$, which allows for a dependence of some observations on the initial state of the Markov process. The SV model example in Appendix A.4.3, supplementary materials demonstrates the role of Q_0.

5. Applications

We assess the performance of the proposed SCDA method in two case studies with distinctively different features resulting in different integration schemes. The first application relates to the unobserved component stochastic volatility model (UCSV)[2] proposed by Stock and Watson (2007), which has become a popular tool for inflation forecasting, see Chan (2013), Chan (2018), Li and Koopman (2021). We also consider a modification of the UCSV model—a time-varying parameter model introduced in Section 3.3—with univariate state vector and SV for the observation errors, which can be fitted using off-the-shelf R package `shrinkTVP` (Knaus et al. 2021). The second application involves the dataset on the Northern lapwing (*Vanellus vanellus*), which has been extensively analyzed in statistical ecology, see Besbeas et al. (2002), Brooks, King, and Morgan (2004), or King et al. (2008). Importantly, in the lapwing model the two latent states follow discrete distributions. We conclude this section with some guidance relating to the practical choices required for implementing the SCDA approach to state space models in general.

We are interested in comparing the performance of the standard DA approach with that of SCDA. For comparability, for each method we use a "vanilla" RW-MH (single-update) algorithm for estimation. We tune each sampler so that the acceptance rates for each element of $\boldsymbol{\theta}$ and the average acceptance rates for each of the imputed states are "reasonable," that is, 20%–40% (Gelman, Roberts, and Gilks 1996; Roberts and Rosenthal 2001). We note, however, that the idea of SCDA does not rely on any particular MCMC updating scheme but on the specification of the auxiliary variables. It is the choice of the analyst what specific sampler is used for updating these auxiliary variables. We are consistent in using the same RW-MH algorithm to consider the relative performance of the DA and SCDA algorithms, but note that one may use, for example, Hamiltonian Monte Carlo (HMC, Neal 2011) or the No-U-Turn Sampler (NUTS, Hoffman and Gelman 2014) for continuous states. The proposed methodology can be also implemented using probabilistic programming languages such as Stan (Stan Development Team 2019) or JAGS (Plummer 2017).

5.1. Example: Econometric Model

We consider the following specification of the UCSV model, for $t = 1, \dots, T$,

$$\begin{aligned} y_t &= \tau_t + \exp(h_t/2)\varepsilon_t^y, & h_t &= h_{t-1} + \sqrt{\omega_h^2}\varepsilon_t^h, \\ \tau_t &= \tau_{t-1} + \exp(g_t/2)\varepsilon_t^\tau, & g_t &= g_{t-1} + \sqrt{\omega_g^2}\varepsilon_t^g, \\ \tau_1 &\sim \mathcal{N}\left(m_\tau, V_\tau \exp(g_0)\right), & \varepsilon_t^x &\overset{\text{iid}}{\sim} \mathcal{N}(0,1), \quad x \in \{y, \tau, h, g\}, \end{aligned} \quad (17)$$

where the inflation rate, y_t, is decomposed into a random walk trend component, τ_t, and a transitory noise component ε_t^y, each of which is subject to stochastic volatility. Log-volatilites h_t and g_t have a random walk specification.[3] The dynamics in (17) follows the specification of Stock and Watson (2007), with the only exception that we allow for distinct volatility-of-volatility coefficients for h_t and g_t, that is, ω_h^2 and ω_g^2. This parametrisation choice, together with putting priors on the initial log-volatility values, h_0 and g_0, is similar to the one considered by Chan (2018). However, Chan (2018) puts a normal prior on the standard deviations ω_h and ω_g, while we adopt a more standard inverse-gamma priors on variances (see, e.g., Chan 2013; Kroese and Chan 2013). In sum, we adopt the following

[2] To provide more intuition behind the UCSV example, we present in Appendix A, supplementary materials an additional conceptual exercise based on the basic SV model, which is a building block of the UCSV model.

[3] As pointed out by Chan (2018), a random walk specification is popular in macroeconometrics, while an alternative modeling approach of adopting a stationary autoregressive process is typical for the financial literature.

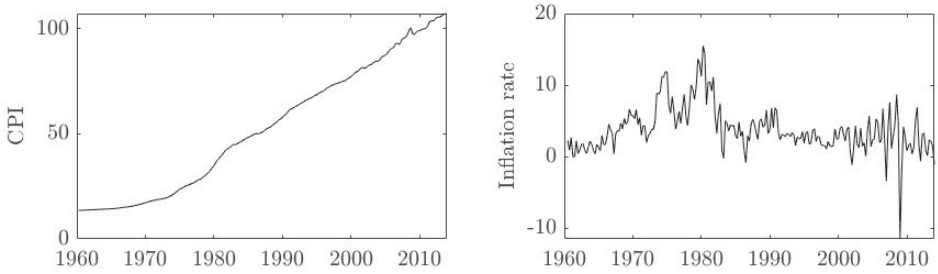


Figure 3. UCSV model: U.S. quarterly CPI (left) and CPI inflation (right) from 1960Q1 to 2013Q4.

prior specification

$$h_0 \sim \mathcal{N}(m_h, V_{h_0}), \qquad g_0 \sim \mathcal{N}(m_g, V_{g_0}), \tag{18}$$

$$\omega_h^2 \sim \mathcal{IG}(\alpha_h, \lambda_h), \qquad \omega_g^2 \sim \mathcal{IG}(\alpha_g, \lambda_g), \tag{19}$$

with $m_h = m_g = 0$, $V_{h_0} = V_{g_0} = 10$, $\alpha_h = \alpha_g = 10$, $\lambda_h = \lambda_g = 0.2^2(\alpha_h - 1)$ and we set $V_\tau = V_h = V_g = 10$, $m_\tau = 0$. We collect the model parameters into $\boldsymbol{\theta} = \{m_h, m_g, \omega_h, \omega_g\}^T$.

The estimation of the UCSV model is challenging due to the intractable likelihood

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \int p(\boldsymbol{y}, \boldsymbol{\tau}, \boldsymbol{h}, \boldsymbol{g}|\boldsymbol{\theta}) d\boldsymbol{\tau} d\boldsymbol{h} d\boldsymbol{g},$$

where $\boldsymbol{\tau} = \{\tau_t\}_{t=1}^T$, $\boldsymbol{h} = \{h_t\}_{t=1}^T$, $\boldsymbol{g} = \{g_t\}_{t=1}^T$. In the frequentist literature, Li and Koopman (2021) propose to estimate parameters of the UCSV model using a simulated maximum likelihood method based on the Kalman filter. However, Bayesian MCMC methods seem to be a more popular tool to this end, see Stock and Watson (2007), Chan (2013), and Shephard (2015). In particular, imputing the hidden states in Bayesian DA leads to the complete data likelihood with a closed form:

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{\tau}, \boldsymbol{h}, \boldsymbol{g}) &= p(\boldsymbol{y}|\boldsymbol{\tau}, \boldsymbol{h}) p(\boldsymbol{\tau}|\boldsymbol{g}) p(\boldsymbol{g}) p(\boldsymbol{h}) \\ &= p(\tau_1) p(g_1) p(h_1) \prod_{t=2}^{T} p(y_t|\tau_t, h_t) \\ &\quad p(\tau_t|\tau_{t-1}, g_t) p(g_t|g_{t-1}) p(h_t|h_{t-1}). \end{aligned}$$

However, standard DA often leads to poorly mixing chains.

5.1.1. Dependence Structure and Semi-Complete Data Likelihood

The UCSV model concerns three states on the real line, all of which follow random walks. The sampling inefficiency in this context originates from a high persistence of the latent processes.[4] In order to break this dependence, we propose two schemes based on the model dependence structure (illustrated in Online Appendix B.1, supplementary materials). We start with integrating out $\boldsymbol{g}_{2T+1}$, the permanent log-volatility in odd time periods, and imputing $\boldsymbol{g}_{2T}$, the permanent log-volatility in even time periods, together with $\boldsymbol{h}$ and $\boldsymbol{\tau}$ (*Scheme 1*). This corresponds to *naive vertical* integration from Section 3.3, with $\boldsymbol{x}_{\text{int}} = \boldsymbol{g}_{2T+1}$ and $\boldsymbol{x}_{\text{aug}} = \boldsymbol{g}_{2T}$, as integration is applied to one state only. This simplicity is useful for illustrative purposes as well as demonstrates a general heuristic that in practice it may be advisable to start with a simple scheme and gradually extend it toward more complex schemes (see Section 5.3 for a discussion of practical heuristics for SCDA). Note that g_t feeds into another latent state, that is, the trend component τ_t, via which it affects the mean of y_t. Thus, next we include the transitory log-volatility h_t, which drives the variance not the mean of y_t, in the integration scheme. This results in integrating out $\boldsymbol{h}_{2T}$ as well as $\boldsymbol{g}_{2T+1}$ (*Scheme 2*), which is another instance of *vertical* integration, with $\boldsymbol{x}_{\text{int}} = \{\boldsymbol{g}_{2T+1}, \boldsymbol{h}_{2T}\}$ and $\boldsymbol{x}_{\text{aug}} = \{\boldsymbol{g}_{2T}, \boldsymbol{h}_{2T+1}\}$ (see Section 3.3). Note that in both schemes the semi-complete data likelihood (7) can be represented as a product of one dimensional integrals,[5] which would not be the case in Scheme 2 if we integrated same period g_t and h_t (that would result in two-dimensional integrals).

5.1.2. Application

We consider $\tilde{y}_t$, the quarterly consumer price indices (CPI) for the United States, from 1960Q1 to 2013Q4, also used by Chan (2018), which we transform to annualized inflation rate $y_t = 400 \log(\tilde{y}_t / \tilde{y}_{t-1})$. Figure 3 illustrates the data. Given the independence of both log-volatilities, $\boldsymbol{h}$ and $\boldsymbol{g}$, we expect Scheme 2 to be more efficient than Scheme 1 and indeed, our pilot study confirmed this conjecture. Thus, below we focus on results for Scheme 2.

We consider 10 independent runs of Scheme 2 and full DA (with random initialization), which we run with 20,000 draws after a burn-in of 10,000. For comparison, we also include results for a DA run with much more draws (2,000,000) and a longer burn-in period (500,000), denoted "DA long." For SCDA we investigate $10, 15, \ldots, 30$ adaptive bins and $20, 25, \ldots, 40$ fixed bins. Figure 4 illustrates parameter posterior means. We observe that all methods deliver means close to those from DA long, however, most variants of SDCA lead to much more precise estimates compared with DA with the same number of draws. Interestingly, the number of bins does not seem to change the reliability of SCDA too much, especially for fixed bins: 20 fixed bins give the same range of posterior means across 10 runs as 40 fixed bins. This is an encouraging result as it demonstrates the stability of the proposed numerical approximations as well as shows that as few as 20 fixed bins (or even 10 adaptive bins) are sufficient to accurately estimate the parameters. In Appendix F.2, supplementary materials we quantitatively summarize Figure 4

[4] The random walks may be regarded as a limiting case of AR(1) process, with the persistence coefficient equal to one.

[5] We provide technical details of the semi-complete data likelihood formulations for both schemes in Appendix B.1, supplementary materials where we also discuss the HMM approximations based on fixed and adaptive bins (introduced in Section 4.1).

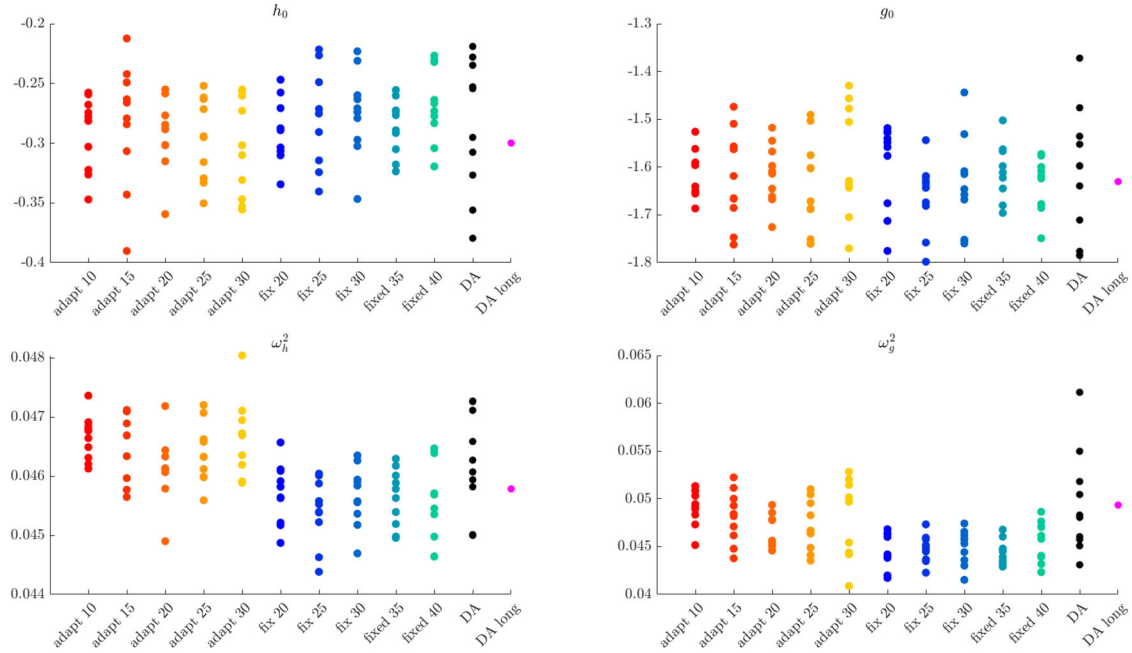


Figure 4. UCSV model: parameter posterior means from 10 replications for SCDA Scheme 2 and DA together with the benchmark from DA long.

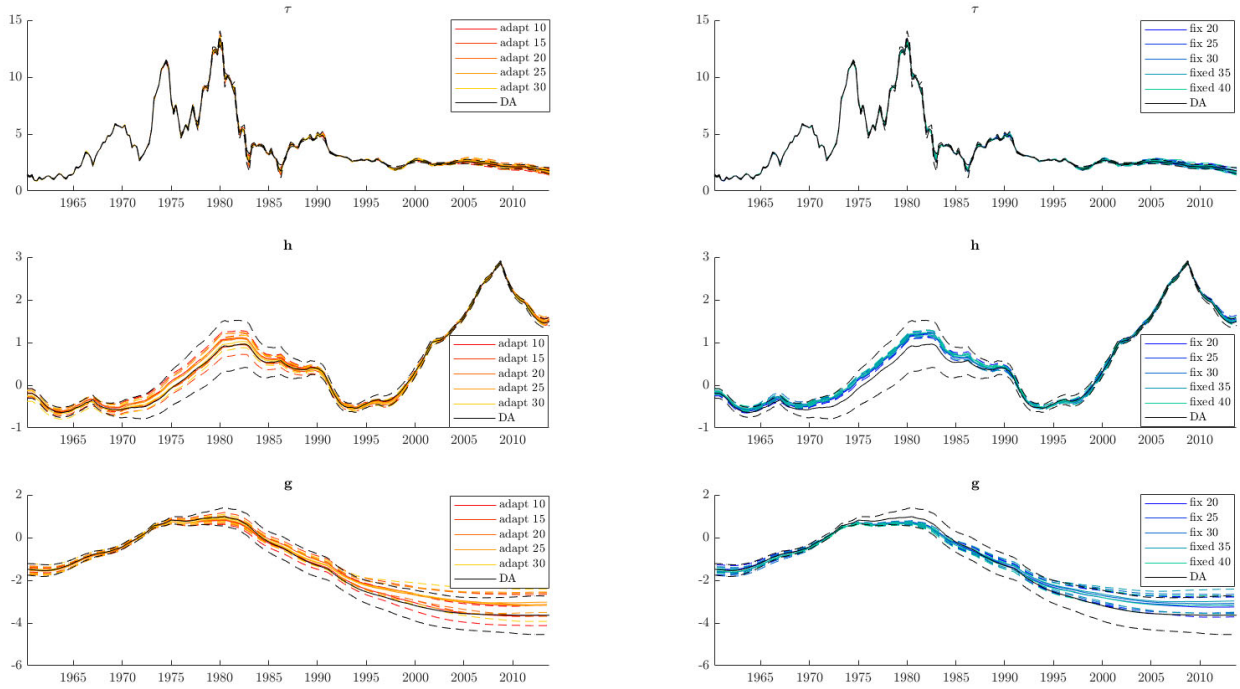


Figure 5. UCSV model: state posterior means (solid lines) averaged over 10 replications for SCDA Scheme 2 and DA, $\pm$ their standard deviations (dashed lines). Left column: adaptive bins, right column: fixed bins.

as well as provide its "quantile counterparts" (for 5%, 25%, 75%, and 95% quantiles) which lead to similar conclusions for the quantiles as for the mean. A similar pattern can be observed for the state posterior means, as shown in Figure 5. As expected, on average SCDA and DA lead to very similar posterior mean, but SCDA delivers more precise estimates (which is especially visible for the hard-to-estimate log-volatilities $\boldsymbol{h}$ and $\boldsymbol{g}$).

Figure 6 presents ESSs for parameters for 10 replications of SCDA Scheme 2 and DA, while Table 1 reports medians of those ESSs together with computing times of each algorithm. We observe that the SCDA approach leads to improved mixing (up to over three times higher ESSs) for all parameters. Moreover, the ESS values for DA often happen to be very low, while it is rare for SCDA to deliver such low ESSs.

We conclude that the SCDA algorithm leads to improved ESS values compared to the full DA algorithm, however, at the cost of an increased computing time. However, SCDA brings an additional advantage in terms of typically more precise parameter and state estimates. We note that despite its relatively complex dependence structure, the UCSV model is computationally

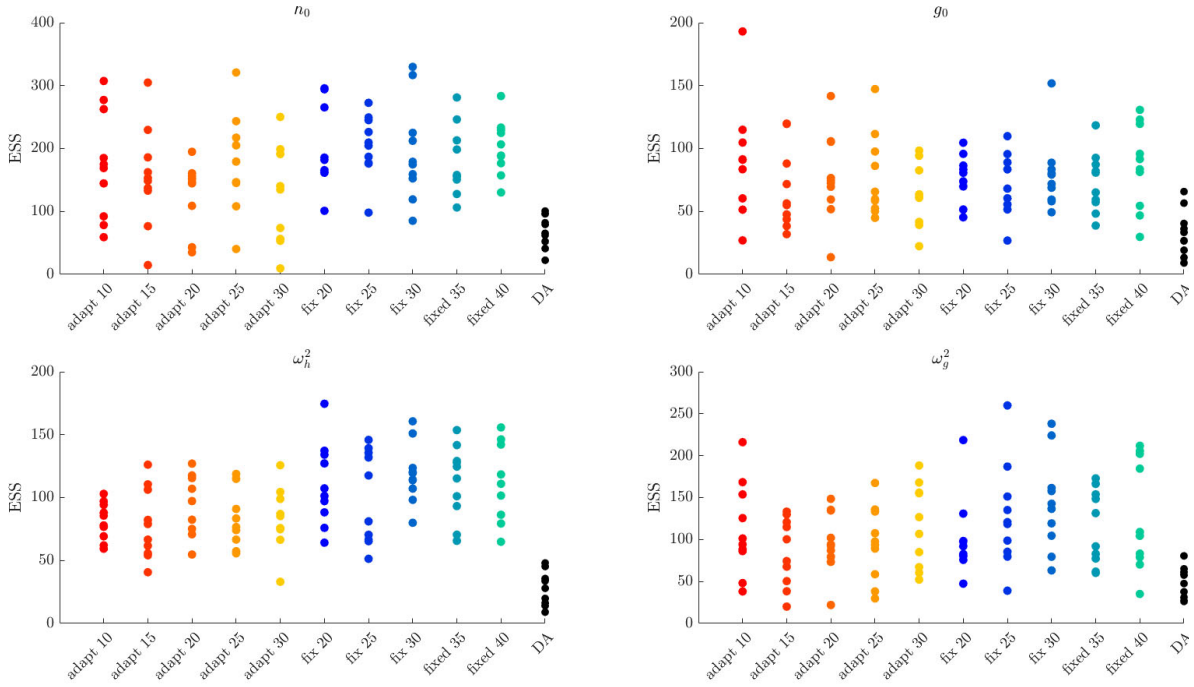


Figure 6. UCSV model: ESS for the parameters.

Table 1. UCSV model: ESS for parameters and computing times (in seconds)—medians for 10 replications.

ESS	Adapt10	Adapt15	Adapt20	Adapt25	Adapt30	Fix20	Fix25	Fix30	Fix35	Fix40	DA
h_0	344	302	299	359	209	348	414	354	314	395	145
g_0	87	56	74	63	62	77	76	76	73	88	30
ω_h^2	164	146	204	160	161	209	199	234	240	229	62
ω_g^2	98	87	94	96	117	87	120	140	112	107	53
Time	71	76	81	86	91	69	73	76	80	83	11

pretty quick due to the univariate normal distributions present in its specification. As we observe in the following study, non-normal distributions, and in particular discrete distributions that tend to be much more computationally expensive, can lead to SCDA being computationally advantageous over DA and hence to much more pronounces efficiency gains.

5.1.3. TVP Model and Comparison with Off-the-Shelf Packages

It would be of interest to compare the performance of the SCDA algorithm with that of (optimised) off-the-shelf packages. However, to the best of our knowledge, there are no available packages for Bayesian estimation of the previous UCSV model (17). However, if we simplify the model we can use the `shrinkTVP` R package (Knaus et al. 2021), which fits the class of *time-varying parameter* (TVP) models. In particular, we replace the stochastic volatility for the trend process τ_t with a homoscedastic noise, model the stochastic volatility of inflation, h_t, as an AR(1) process (instead of a random walk) and use different priors.

Time varying parameter model. The resulting model has the following specification, for $t = 1, \dots, T$,

$$y_t = \tau_t + \exp(h_t/2)\varepsilon_t^y, \quad \varepsilon_t^x \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \quad x \in \{y, \tau, h\},$$
$$\tau_t = \tau_{t-1} + \sqrt{\omega_\tau^2}\varepsilon_t^\tau, \quad h_t = \mu + \phi(h_{t-1} - \mu) + \sqrt{\omega_h^2}\varepsilon_t^h$$

$$\tau_0 \sim \mathcal{N}\left(\bar{\tau}, \omega_\tau^2\right), \quad h_0 \sim \mathcal{N}\left(\mu, \frac{\omega_h^2}{1-\phi^2}\right), \tag{20}$$

with the following prior specification for the mean of the initial trend value and the standard deviation of the innovations,

$$\bar{\tau}|\lambda_B^2 \sim \mathcal{N}\left(0, \frac{2}{\lambda_B^2}\right), \qquad \omega_\tau|\kappa_B^2 \sim \mathcal{N}\left(0, \frac{2}{\kappa_B^2}\right),$$

which are a special case of the normal-gamma-gamma prior (Cadonna, Frühwirth-Schnatter, and Knaus 2020) known as ridge regression. Here, κ_B^2 and λ_B^2 are shrinkage parameters, as the larger their values the stronger the "pull" toward zero of the corresponding parameters. In our experiments we set both κ_B^2 and λ_B^2 to 0.02, which corresponds to relatively vague priors. For the parameters of the SV process h_t we adopt priors from Kastner and Frühwirth-Schnatter (2014) and Knaus et al. (2021) given as

$$\mu \sim \mathcal{N}(b_\mu, B_\mu), \quad \frac{\phi+1}{2} \sim \mathcal{B}(\alpha_\phi, \beta_\phi),$$
$$\omega_h^2 \sim \mathcal{G}\left(\frac{1}{2}, \frac{1}{2B_{\omega_h}}\right) = B_{\omega_h} \cdot \chi_1^2,$$

with $b_\mu = 0, B_\mu = 1, \alpha_\phi = 5, \beta_\phi = 1.5, B_{\omega_h} = 1$.

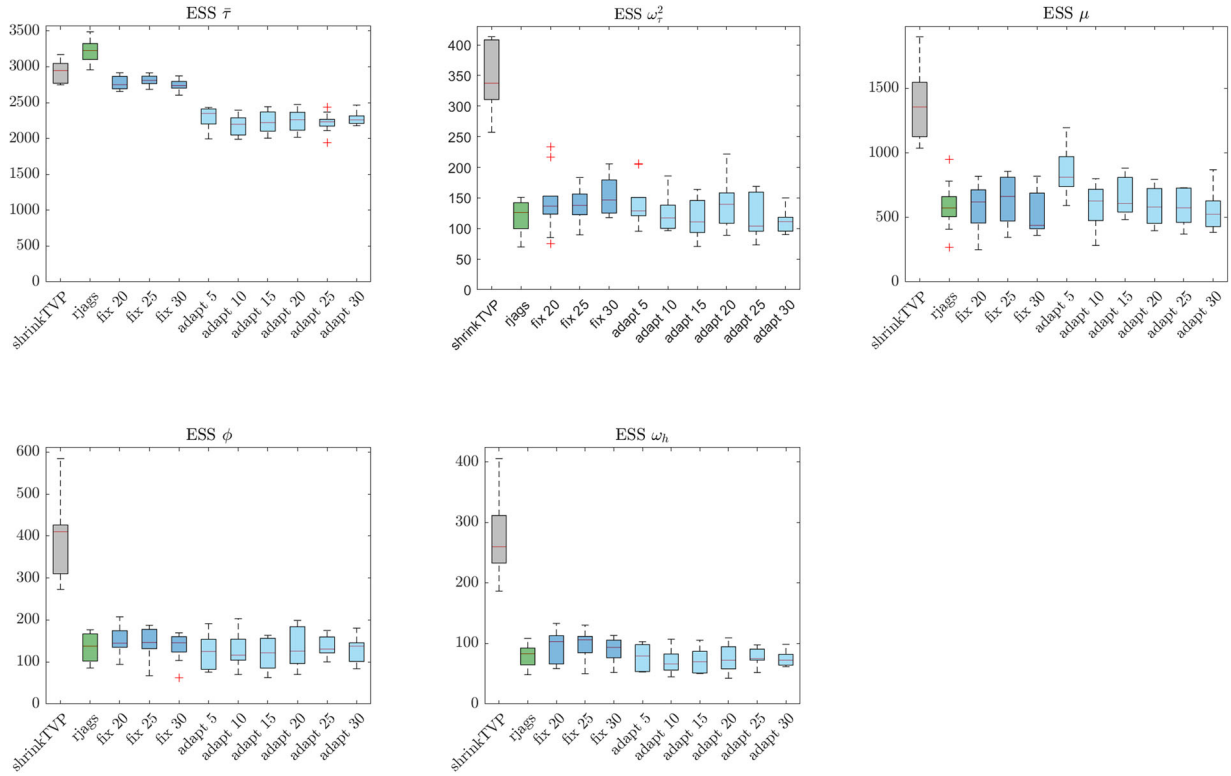


Figure 7. TVP model: ESS for the parameters.

Alternative packages. `shrinkTVP` implements a DA algorithm using Gibbs sampling with MH steps and applies a number of specialist algorithms to improve mixing. In particular, this includes the ancillarity-sufficiency interweaving strategy (ASIS) of Kastner and Frühwirth-Schnatter (2014) or the band and sparse matrix algorithm for linear Gaussian SSMs of McCausland, Miller, and Pelletier (2011) for jointly sampling τ_t, $t = 0, \ldots, T$. In addition, we implement a standard DA algorithm using the R package `rjags`, which is based on the JAGS (Plummer 2017) probabilistic programming language.

SCDA scheme. We specify SCDA by setting $\boldsymbol{x}_{\text{int}} = \{\boldsymbol{h}_{2T}, \boldsymbol{\tau}_{2T+1}\}$. Alternating the time periods for the integrated h_t and τ_t states allows us to avoid double integrals in the algorithm as the integrals associated with $\boldsymbol{h}_{2T}$ and $\boldsymbol{\tau}_{2T+1}$ disjoint. This can be observed since the integrals associated with h_t and τ_t are dependent only via the likelihood term $p(y_t|\tau_t, h_t)$ so that integrating same-t h_t and τ_t results in double integrals; while integrating out alternating t (i.e., $h(t2)$ and $\tau(2t+1)$) leads to single integrals only.

Results. The results below are based 10 algorithm replications, each with 20,000 draws after 10,000 draws for burn-in (or adaptation for `rjags`). For SCDA we consider 20, 25, and 30 fixed bins (10 and 15 fixed bins performed poorly, as for the previous example) and 5, 10, 15, 20, 25, and 30 adaptive bins. The comparisons of computational time are less informative due to the differences in computational packages used: `R` for the two alternative packages with the sampler in `shrinkTVP` coded in C++, and MATLAB, so that we focus on the absolute effective sample size for comparisons. However, to give some idea of computing times,[6] `shrinkTVP` took around 10 sec, JAGS 20 sec, and SCDA between 23 and 45 sec dependent on adaptive/fixed and number of bins.

Figures 7 and 8 present ESSs for the parameters and states, respectively. For the latter each boxplot is based on 10 time series of ESSs for the corresponding state (even τ_t or odd h_t); we provide ESSs for a selection of individual τ_ts and h_ts in Appendix F.3, supplementary materials. While we would expect the specialist `shrinkTVP` package with model-specific sampling techniques to perform much better *in this specific context* than the proposed general SCDA algorithm, we note that the sampling efficiency of SCDA even with *vanilla RW-MH updates* is relatively comparable with that of `shrinkTVP` in that the ESSs are of the same order; and very similar to `rjags` that uses a slice-sampling algorithm. Finally, we note that sampling efficiency of SCDA does not change substantially with the number of bins used, demonstrating the general stability of the algorithm.

5.2. Example: Ecological Model

We consider $\boldsymbol{y} = (y_1, \ldots, y_T)$, a time series of observations relating to census data (abundance index) of adult British lapwings (*Vanellus vanellus*, see Appendix E, supplementary materials for details). The lapwings dataset plays an important role in statistical ecology where it has frequently served as an illustration (see Besbeas et al. 2002; King 2011).

[6] On a standard personal laptop, with i7 Intel(R) Core(TM), 2.80 GHz, 16 GB RAM.

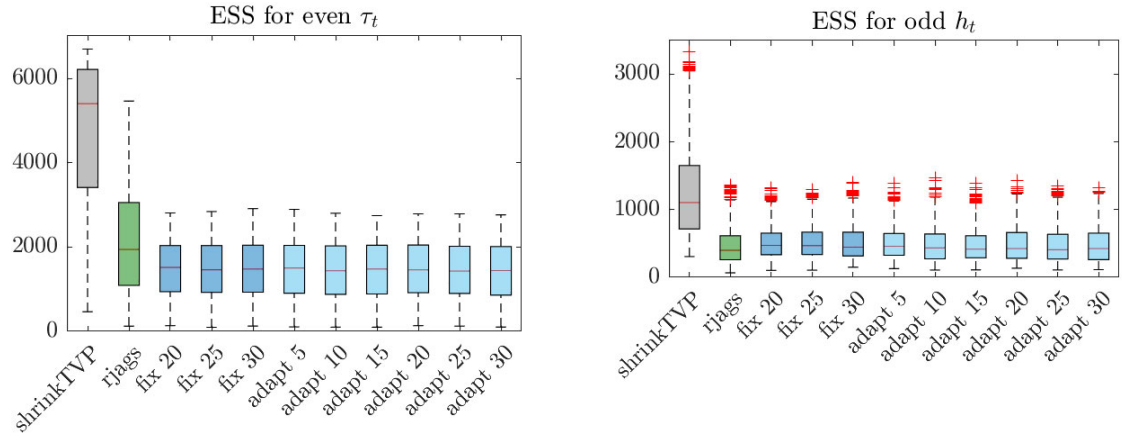


Figure 8. TVP model: ESS for the states. Left—even τ_t, right—odd h_t.

The counts are only estimates of the true unknown population size, which is assumed to change over time according to a first order Markov process. The latent population is related to two times series: for first-years and adults, denoted $\boldsymbol{N_1} = (N_{1,1}, \ldots, N_{1,T})^T$ and $\boldsymbol{N_a} = (N_{a,1}, \ldots, N_{a,T})^T$, respectively. Hence, the latent state is given by $\boldsymbol{x} = (\boldsymbol{N_1}^T, \boldsymbol{N_a}^T)^T$. Following Besbeas et al. (2002) we model the count data via the following SSM

$$y_t|N_{a,t}, \boldsymbol{\theta} \sim \mathcal{N}(N_{a,t}, \sigma_y^2), \qquad t = 1, \ldots, T, \quad (21)$$

$$N_{1,t}|N_{a,t-1}, \boldsymbol{\theta} \sim \text{Poisson}(N_{a,t-1}\rho_{t-1}\phi_{1,t-1}), \quad (22)$$

$$N_{a,t}|N_{1,t-1}, N_{a,t-1}, \boldsymbol{\theta} \sim \text{Bin}\big((N_{1,t-1} + N_{a,t-1}), \phi_{a,t-1}\big), \quad (23)$$

where $N_{1,0} \sim$ Neg–bin$(r_{1,0}, p_{1,0})$ and $N_{a,0} \sim$ Neg–bin $(r_{a,0}, p_{a,0})$. The model is parameterized by the time-varying productivity rate ρ_t, and time-varying survival rates $\phi_{1,t}$ and $\phi_{a,t}$, for first-years and adults, respectively, while $a_{i,0}$ and $p_{i,0}$ are hyperparameters of the prior distribution on the initial state value $N_{i,0}$, $i \in \{1, a\}$.

We let the SMM parameters follow regressions specified by Besbeas et al. (2002)

$$\text{logit}\phi_{1,t} = \alpha_1 + \beta_1 f_t, \quad \text{logit}\phi_{a,t} = \alpha_a + \beta_a f_t,$$
$$\log \rho_t = \alpha_\rho + \beta_\rho \tilde{t},$$

where $\tilde{t}$ the normalized time index and f_t denotes the normalized value of frost days *fdays* in year t, see Appendix E, supplementary materials for the explanation of this covariate.

To improve the estimation, Besbeas et al. (2002) consider an integrated population model combining the census data with ring-recovery data (see Appendix E, supplementary materials for the formula of the additional regression parameterized by α_λ and β_λ and further details). We refer to Besbeas et al. (2002) for a more detailed description of the integrated model. The set of model parameters is collected in a vector $\boldsymbol{\theta} = (\alpha_1, \alpha_a, \alpha_\rho, \alpha_\lambda, \beta_1, \beta_a, \beta_\rho, \beta_\lambda, \sigma_y^2)^T$.

Finally, we set independent vague $\mathcal{N}(0, 100)$ priors for the regression coefficients α_i and β_i, $i \in \{1, a, \rho, \lambda\}$ and $\Gamma^{-1}(a_y, b_y)$ on σ_y^2 with $a_y = 0.001 = b_y$. For the initial states, we set $r_{1,0} = 4$ and $p_{1,0} = 0.98$ so that the prior mean and variance of first-year birds is approximately 200 and 10,000, respectively; and $r_{a,0} = 111$ and $p_{a,0} = 0.9$, so that the prior mean and variance of adults is approximately 1000 and 10,000, respectively.

System (21)–(23) is non-Gaussian and nonlinear with the associated likelihood unavailable in a closed form. The standard vanilla DA approach leads to poorly mixing MCMC algorithms as demonstrated by King (2011). To this end, we first consider the dependence structure in the model to determine sensible $\boldsymbol{x}_{\text{int}}$ and $\boldsymbol{x}_{\text{aug}}$.

The two-dimensional state $(N_{1,t}, N_{a,t})^T$ follows the first-order Markov process with a nontrivial transition kernel. First-year birds in t only feed into adults in $t + 1$, however, adults in t contribute to both the number of first-years and adults in $t + 1$ as well as to the observed y_t. This suggests that reducing the strength of the dependence structure can be obtained by integrating out $\boldsymbol{N_1}$ while imputing $\boldsymbol{N_a}$. This corresponds to the *horizontal* integration scheme with $\boldsymbol{x}_{\text{int}} = \boldsymbol{N_1}$ and $\boldsymbol{x}_{\text{aug}} = \boldsymbol{N_a}$. The resulting modified dependence structure is presented in Figure 9. Marginalizing over $\boldsymbol{N_1}$ simplifies the analysis as only $\boldsymbol{N_a}$ need to be considered and they now follow a second-order Markov process. A similar second order structure in this context has also been noted by Besbeas and Morgan (2019).

5.2.1. Hidden Markov Model Approximation

The semi-complete data likelihood for the augmented data $(\boldsymbol{y}^T, \boldsymbol{N_a}^T)^T$ is

$$p(\boldsymbol{y}, \boldsymbol{N_a}|\boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{N_a}, \boldsymbol{\theta})p(\boldsymbol{N_a}|\boldsymbol{\theta}), \quad (24)$$

which is still intractable, thus, we approximate it using the HMM embedding (16). We provide the derivation details in Appendix B.2, supplementary materials, where we show that the HMM approximation to the SCDL (24) can be expressed as

$$p(\boldsymbol{y}, \boldsymbol{N_a}|\boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{N_a})p(\boldsymbol{N_a}) = \boldsymbol{u}_0 p(N_{a,0}) \left(\prod_{t=1}^{T} P_t \Gamma_t Q_t \right) \mathbf{1},$$

where $\boldsymbol{u}_0 = [p(N_{1,0} = 0), \ldots, p(N_{1,0} = N^*)]$ is the $(1 \times N^*)$ row vector of the initial state distribution; P_t is a diagonal matrix with $p_{k,t}$ elements on the diagonal, such that $p_{k,t} = p(N_{a,t}|N_{a,t-1}, N_{1,t-1} = k)$ denoting the conditional pmf of $N_{a,t}$ given $N_{1,t-1} = k$ and $N_{a,t-1}$; $\Gamma_t = [\gamma_{lk,t}]_{l,k=1}^{N^*}$ is a latent state transition matrix with transition probabilities $\gamma_{lk,t} = \mathbb{P}(N_{1,t} = k|N_{a,t-1})$; $Q_t = p(y_t|N_{a,t})\mathbb{I}$ is the observation matrix.=; and $\mathbf{1}$ denotes the $(N^* \times 1)$ column vector with each element equal to 1.

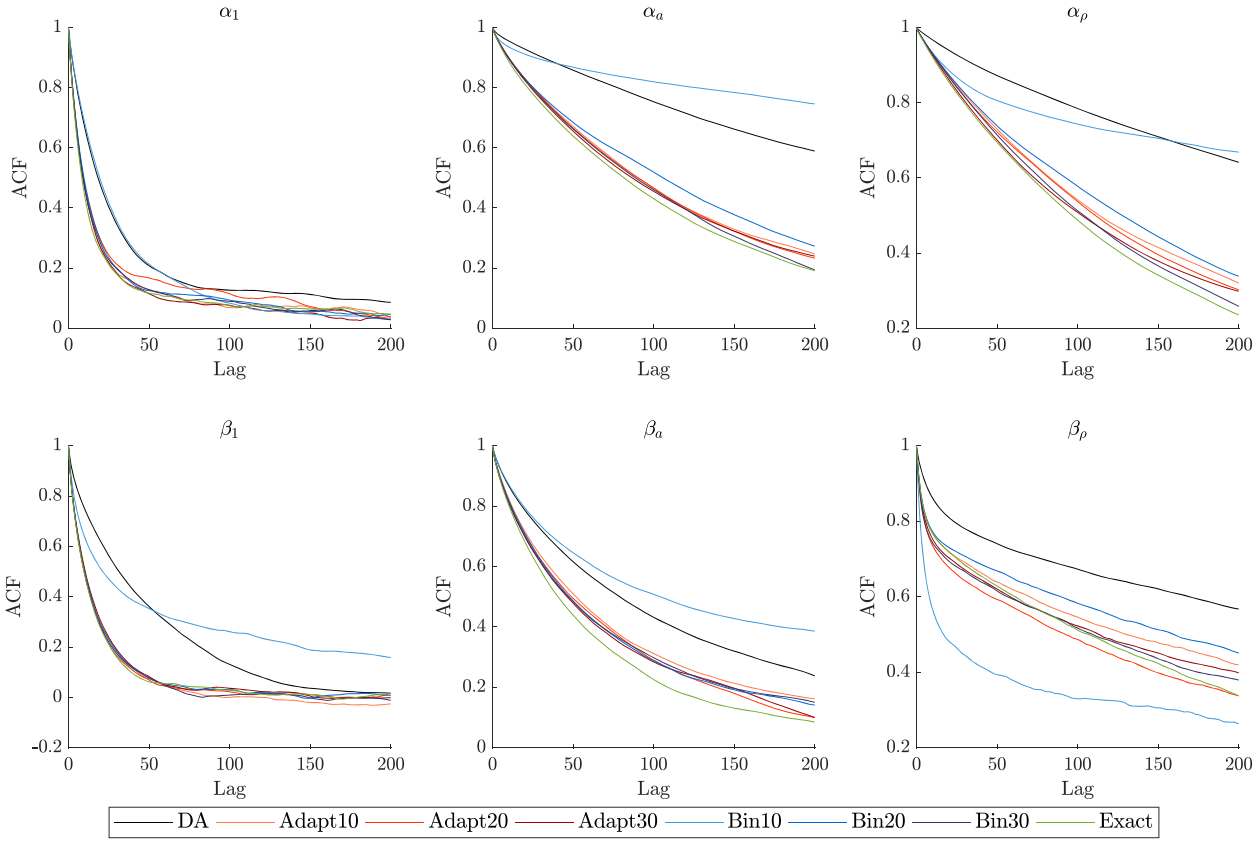


Figure 9. Lapwing data: combining DA and HMM structure. Diamonds—the imputed nodes, squares—the data, circles—the unknown variables. Integrating out $\boldsymbol{N_1}$ leads to a second order HMM on $\boldsymbol{N_a}$. Dashed lines for the relations *from* the imputed states. See Section 4.2 for the general explanation of the notation and Section 5.2.1 for the details specific to the lapwings case.

Table 2. Lapwing data: absolute (in seconds) and relative (wrt DA) computing times.

Method	DA	Adapt10	Adapt20	Adapt30	Fixed10	Fixed20	Fixed30	Exact
Absolute time	1203	978	1067	1024	1022	1060	1135	2855
Relative time	1.00	0.81	0.89	0.85	0.85	0.88	0.94	2.37

5.2.2. Results

We compare the performance of the standard DA approach, in which we impute $\boldsymbol{\theta}$, $\boldsymbol{N_1}$ and $\boldsymbol{N_a}$, with that of the SCDA, in which we impute $\boldsymbol{\theta}$ and $\boldsymbol{N_a}$. For comparability we use a "vanilla" MH RW algorithm for the estimation of the integrated model, with discrete uniform updates for the states and normal updates for the regression coefficients. We use a Gibbs update $\sigma_y^2|\boldsymbol{N_a} \sim \Gamma^{-1}\left(a_y + T/2, b_y + \sum_{t=1}^{T}(y_t - N_{a,t})^2/2\right)$ for the observation variance. For the SCDA we first consider the "exact" integration used in the derivations above, in which the only influence on the posterior is the upper limit which we set $b_B = 679$. This choice of the upper bound is based on the results for first-years from previous studies. We further consider a number of approximate schemes based on fixed and adaptive intervals (with 10, 20, and 30 bins in each case). For adaptive bins we use a normal approximation to the Poisson distribution. Each time we use 100,000 draws after a burn-in of 10,000.

Table 2 summaries computation time for each scheme. As expected, the exact method is the slowest (2.5 times than the full DA approach) as each integration is based on summing 680 elements. All the approximate schemes are faster (10%–20%) than the DA approach due to their efficient implementation based on vectorized computations with relatively few elements to be summed every iteration. Table 3 presents the results for the regression parameters in terms of posterior means and standard deviations as well as ESSs and ESSs per second, for DA and selected SDCA approaches. Appendix F.1, supplementary materials provides a comparison for all elements of $\boldsymbol{\theta}$ and selected elements of $\boldsymbol{N_a}$ between all the methods.

The results demonstrate the efficiency of the proposed SCDA approach: all the SCDA-based schemes, except the one based on 10 fixed bins, outperform the full DA approach by delivering much higher (up to four times) ESSs and ESSs/sec. This is illustrated in Figure 10, which shows the autocorrelation (ACF) plots for the SSM parameters. We refer to Appendix F.1, supplementary materials for the ACF plots for selected elements of $\boldsymbol{N_a}$.

5.3. Practical Issues

Here we briefly discuss some of the practical issues that need to be considered when implementing an SCDA approach.

Table 3. Lapwing data: SSM regression parameter posterior means, standard deviations and ESSs and ESS/s (highest in bold).

Method		α_1	α_a	α_ρ	β_1	β_a	β_ρ
DA	Mean	0.547	1.574	−1.189	−0.164	−0.240	−0.348
	(Std)	(0.068)	(0.071)	(0.091)	(0.062)	(0.039)	(0.043)
	ESS	685	124	112	1050	389	106
[1204 s]	ESS/s	0.57	0.10	0.09	0.87	0.32	0.09
Adapt10	Mean	0.547	1.564	−1.180	−0.163	−0.239	−0.350
	(Std)	(0.068)	(0.070)	(0.092)	(0.061)	(0.040)	(0.040)
	ESS	1490	390	316	2777	**527**	126
[978 s]	ESS/s	1.52	**0.40**	**0.32**	**2.84**	**0.54**	0.13
Fixed10	Mean	0.512	1.441	−1.044	−0.207	−0.205	−0.348
	(Std)	(0.070)	(0.055)	(0.063)	(0.050)	(0.039)	(0.022)
	ESS	942	34	37	181	105	**282**
[1022 s]	ESS/s	0.92	0.03	0.03	0.18	0.10	**0.28**
Fixed30	Mean	0.545	1.562	−1.170	−0.162	−0.240	−0.342
	(Std)	(0.069)	(0.073)	(0.095)	(0.061)	(0.039)	(0.040)
	ESS	**1758**	**439**	**329**	**2873**	502	208
[1136 s]	ESS/s	**1.55**	0.39	0.29	2.53	0.44	0.18

NOTE: Computing times in square brackets.

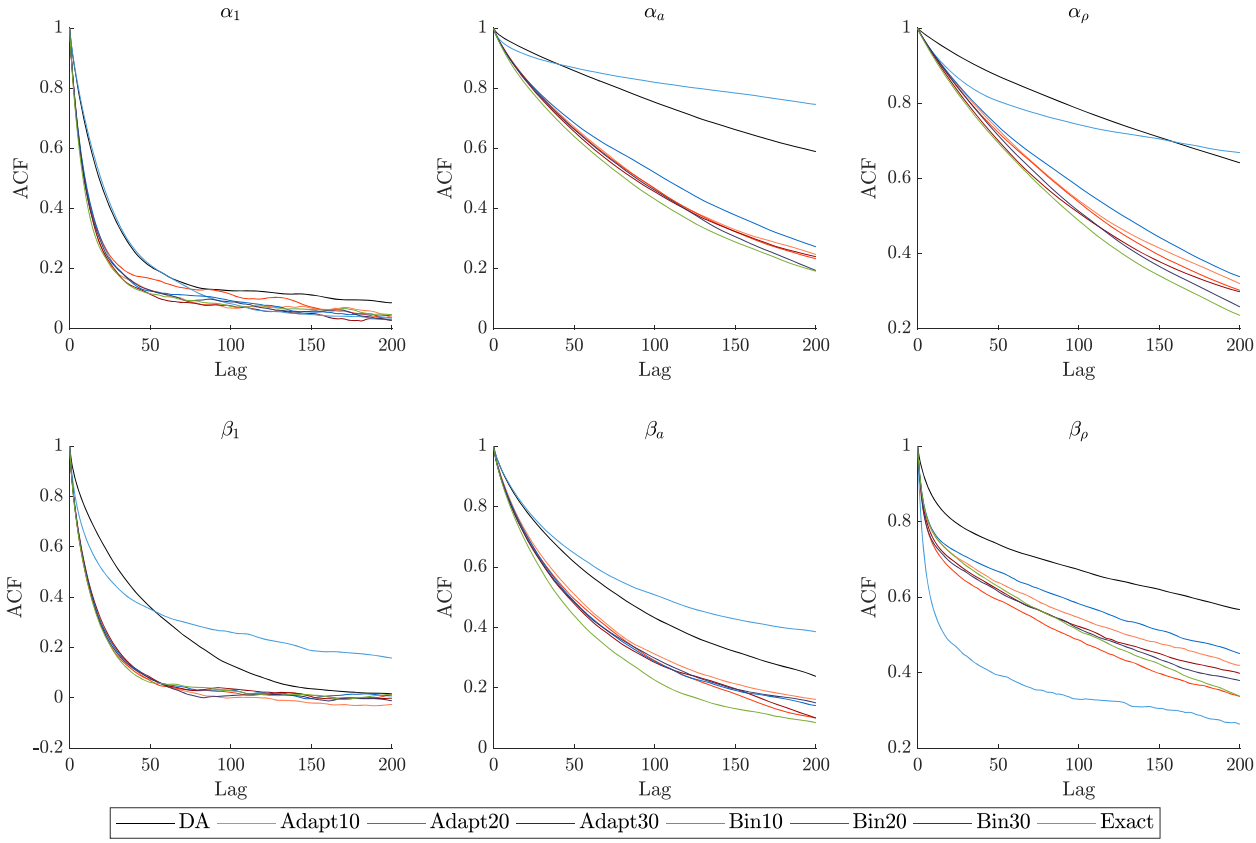


Figure 10. Lapwing data: ACF plots for the SSM regression parameters.

What states to integrate? In many cases, practitioners may have an awareness of what states are "problematic" in a given application. For instance, it is well known that stochastic volatilities are characterised by high persistence, which adversely affects mixing. In such circumstances such poorly mixing states provide natural candidates for integrating. Alternatively, there might be "double latent" states, such as g_t in the UCSV model or first-years in the Lapwings example, that do not directly feed into the observation process, affecting it only indirectly via other latent states. Thus, integrating these "double latent" states might be more convenient as opposed to the states that the observations are a direct function of. Further, we note that in such circumstances, it may also be that these latter states are of more interest (as observations are made directly on these), and imputing these states provides a direct mechanism for obtaining their associated posterior distribution.

In general, and as discussed in the UCSV example (Section 5.1), simple schemes will often provide a good starting point for developing more complex schemes. For instance, for the TVP case one could start with two independent "naive" schemes with either τ_t's or h_t's being integrated and then combine them into an τ_t-h_t scheme. Further, we note that it will typically be more efficient to implement a scheme with one-dimensional integrals. For a broad class of models with independent states, such as the general TVP model (9)–(10), the suggestions from Section 3.3 also apply.

Number and type of bins. In theory, and also observed in practice for the examples considered, adaptive bins are superior to fixed bins, with the latter also requiring the user to specify the integration range and are more challenging to efficiently accommodate states with nonstationary features (e.g., periods of high and low volatility). However, for certain problems it might be easier to specify fixed bins due to the underlying process specification. For instance, the first-years in the lapwings example follow a Poisson distribution, bounded from below at zero; in addition the upper bound on the integration range might be deduced from previous population studies. Thus, the specific state space model should be taken into consideration, with the general tradeoff between a larger number of fixed bins (if the range of values is not too large) or a smaller number of adaptive bins. For our applications, relatively few bins were generally required to obtain accurate results, (e.g., 10–20 bins sufficing), and thus, we recommend initially starting with a small number of bins and to check the sensitivity by increasing these to check for consistency.

6. Discussion

We have presented a new estimation method for state space models using semi-complete data augmentation. The idea can, in general, be combined with different updating schemes to provide more efficient MCMC sampling algorithms, although for the purposes of this article we simply focus on "vanilla" updating algorithms. The main idea behind the approach is to combine data augmentation with numerical integration, where the latter aims at reducing the dependence between the imputed auxiliary variables. The concept relates to general Rao-Blackwellization methods, however, we do not require the resulting conditional distribution (given the imputed states) to be analytically integrable, nor the imputed auxiliary variables to be sufficient statistics for the marginalized states.

We assume that the system states are ordered, but may be discrete or continuous in nature. The corresponding integration schemes for these states are based on the insights from hidden Markov models in the sense that we specify new transition probabilities between redefined states, to be numerically integrated out, conditionally on the auxiliary variables. Further efficiency gains can be obtained by "binning." This results in an approximation to the semi-complete data likelihood and we note that for continuous states such an approximation is a natural starting point for our approach (as in principle for any MC based analysis). We consider two types of "binning": "fixed bins" based on a prespecified grid and "adaptive bins" based on for example, quantiles of the relevant distribution. The latter removes the problem of specifying the "essential domain" required for fixed bins (see Kitagawa 1987; Langrock, MacDonald, and Zucchini 2012b). Adaptive bins are also more suited for problems with highly varying integration ranges, such as the class of SV models, for which fixed bins are unlikely to be efficient (see Sandmann and Koopman 1998). In our examples a similar accuracy was achieved by using fewer adaptive bins than fixed bins. A more general case would allow for K unordered discrete states. However, we would expect in these cases K would be most likely be small, allowing us to simply calculate the explicit observed data likelihood. If this is not the case, then, an additional step of how to combine the states into bins is necessary. We note that in this case some multivariate classification approach could be used to cluster the groups via some training data, such as k-means clustering. However, this could raise the question of the interpretability of the clustered states.

We describe two empirical studies applying the SCDA approach compared to the general "vanilla" MCMC algorithm, before comparing a basic SCDA algorithm with a tailored R package for a TPV model. For the lapwings data model efficiency gains are observed in both the higher effective sample sizes compared to the standard DA technique and computational speed. For the UCSV models, SCDA does again boost the mixing, but at the cost of an increased computing time. However, there are a number of issues that will have a major impact on computing times (and in particular ESS per second) including for example, advanced programming techniques such as parallelization or coding in C/C++ (but this has not been the main focus of the article). In addition, and as expected, bespoke algorithms designed for specific models will generally be optimal for the designed set of models. However, the SCDA approach has potential for being incorporated into such algorithms, and using the bespoke updating algorithms on the model parameters to improve performance.

The performance of the algorithm will, in general, be dependent on many factors, including, for example, the posterior correlation/structure of the parameters and system states, complexity of the numerical integration scheme, distributions specified within the model and MCMC proposal distributions. In general, efficient algorithms become more important when standard techniques become infeasible or cumbersome. We note that the coding of our SCDA approach is adaptable from the full DA approach, which is a natural first approach taken for many state-space-type models. Even moderate computational savings (on the order of say, 2–4 times faster) will be magnified when the models are required to be fitted to many multiple datasets, which is becoming increasingly common due to the availability of data, for example, ecological organisations and charities (e.g., the British Trust for Ornithology and Butterfly Conservation Trust) collect data on many hundreds of species. Further, increasing the number of latent states and/or the length of the time series of these datasets will also likely increase the correlation between the model parameters and auto-correlation of the MCMC algorithm, which influence the potential computational savings.

The split of the latent states into "auxiliary" and "integrated" variables is model-dependent and should be specified in such a way that the algorithm is efficient. This choice is not unique and multiple approaches may be applied—the efficiency of these will dependent on both the model and data. On the one hand, the imputed states aim to have reduced correlation, to improve mixing of MCMC algorithms; on the other hand, the numerical integration is over a very low number of dimensions, which in many cases is feasible due to conditional independence of the integration problems. To identify such conditionally independent latent states it can be useful to investigate the underlying graphical structure of the model. In general, high dimensional integration remains a challenging problem, which we leave for further research, noting that quasi Monte Carlo could be useful in this context.

Finally, and as previously mentioned, application-specific, off-the-shelf packages such as `shrinkTVP` (Knaus et al. 2021) are likely to provide, at least in the first instance, an efficient model-fitting choice. However, such packages are available only for a number of specific predefined models. Deviations from the built-in specifications often require either shoe-horning the model into such a package (often to the detriment of the application), or resorting to general sampling techniques, typically involving full data augmentation. In contrast, general model-fitting algorithms such as SCDA are easily adaptable between different model specifications, and such an adaptation becomes essentially trivial when the underlying dependence structure is preserved. Further, the methodological framework presented within this approach can incorporate additional tools or techniques that have been developed for other model-specific algorithms, in terms of updating strategies applied to the (static) model parameters and/or auxiliary variables, leading to more tailored and efficient updating algorithms.

The proposed methodology naturally leads to several further topics for further research. First, we aim to investigate bounds of approximation errors in order to quantify the demonstrated higher usefulness of adaptive bins compared to fixed bins. Second, adopting automated methods to identify the correlation structure would make applying the SCDA approach to new models easier and potentially more efficient, especially if the model at hand is complex and/or there are no "natural candidates" for the integrated states. Third, we expect parallelization methods to reduce the increased computing times, since many of the imputed states will be conditionally independent of each other. Finally, providing an accompanying SCDA general package is an appealing idea, but beyond the scope of this article.

Supplementary Materials

Online Appendix (SCDA_OnlineSupplement.pdf, pdf file) provides additional examples and results, technical details and a description of the lapwings dataset.

Codes (SCDA_Codes.zip, zip file) contain `MATLAB` and `R` scripts, functions and data used for the empirical applications (explained in the corresponding `README.txt` files).

Funding

AB acknowledges funding from the UK Engineering and Physical Sciences Research Council (EPSRC), grant numbers EP/N014642/1, EP/R018634/1 and EP/T017899/1. RK was supported by the Leverhulme research fellowship RF-2019-299.

ORCID

Agnieszka Borowska http://orcid.org/0000-0001-9123-6227
Ruth King http://orcid.org/0000-0002-5174-8727

References

Andrieu, C., Doucet, A., and Holenstein, R. (2010), "Particle Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society*, Series B, 72, 269–342. [1,4]

Andrieu, C., and Roberts, G. (2009), "The Pseudo-Marginal Approach for Efficient Monte Carlo Computations," *The Annals of Statistics*, 37, 697–725. [4]

Bardenet, R., Doucet, A., and Holmes, C. C. (2017), "On Markov Chain Monte Carlo Methods for Tall Data," *Journal of Machine Learning Research*, 18, 1–43. [2]

Beaumont, M. (2003), "Estimation of Population Growth or Decline in Genetically Monitored Populations," *Genetics*, 164, 1139–1160. [4]

Belmonte, M. A., Koop, G., and Korobilis, D. (2014), "Hierarchical Shrinkage in Time-Varying Parameter Models," *Journal of Forecasting*, 33, 80–94. [5]

Besbeas, P., Freeman, S. N., Morgan, B. J. T., and Catchpole, E. A. (2002), "Integrating Mark–Recapture–Recovery and Census Data to Estimate Animal Abundance and Demographic Parameters," *Biometrics*, 58, 540–547. [7,11,12]

Besbeas, P., and Morgan, B. J. T. (2019), "Exact Inference for Integrated Population Modelling," *Biometrics*, 75, 475–484. [12]

Bitto, A., and Frühwirth-Schnatter, S. (2019), "Achieving Shrinkage in a Time-Varying Parameter Model Framework," *Journal of Econometrics*, 210, 75–97. [2,5]

Brooks, S. P., King, R., and Morgan, B. J. T. (2004), "A Bayesian Approach to Combining Animal Abundance and Demographic Data," *Animal Biodiversity and Conservation*, 27, 515–529. [7]

Cadonna, A., Frühwirth-Schnatter, S., and Knaus, P. (2020), "Triple the Gamma—a Unifying Shrinkage Prior for Variance and Variable Selection in Sparse State Space and TVP Models," *Econometrics*, 8, 20. [10]

Cappé, O., Moulines, E., and Ryden, T. (2006), *Inference in Hidden Markov Models*, Springer Series in Statistics, New York: Springer. [3]

Chan, J. C. (2013), "Moving Average Stochastic Volatility Models with Application to Inflation Forecast," *Journal of Econometrics*, 176, 162–172. [7,8]

——— (2018), "Specification Tests for Time-Varying Parameter Models with Stochastic Volatility," *Econometric Reviews*, 37, 807–823. [7,8]

Doucet, A., de Freitas, N., and Gordon, N. (eds.) (2001), *Sequential Monte Carlo Methods in Practice*, New York: Springer. [1]

Doucet, A., Freitas, N. D., Murphy, K., and Russell, S. (2000), "Rao-Blackwellised Particle Filtering for Dynamic Bayesian Networks," in *Proceedings of the Sixteenth conference on Uncertainty in Artificial Intelligence*, pp. 176–183. [4]

Durbin, J., and Koopman, S. J. (2012), *Time Series Analysis by State Space Methods*, Oxford Statistical Science Series (2nd ed.), Oxford: Oxford University Press. [1,3,5]

Frühwirth-Schnatter, S. (1994a), "Applied State Space Modelling of non-Gaussian Time Series Using Integration-based Kalman Filtering," *Statistics and Computing*, 4, 259–269. [2]

——— (1994b), "Data Augmentation and Dynamic Linear Models," *Journal of Time Series Analysis*, 15, 183–202. [1,3]

——— (2004), "Efficient Bayesian Parameter Estimation," in *State Space and Unobserved Component Models: Theory and Applications*, eds. A. C. Harvey, S. J. Koopman, and N. Shephard, pp. 123–151, Cambridge: Cambridge University Press. [1,3]

Gelman, A., Roberts, G. O., and Gilks, W. R. (1996), "Efficient Metropolis Jumping Rule," in *Bayesian Statistics 5: Proceedings of the Fifth Valencia International Meeting*, eds. S. Brooks, J. G. A. Gelman, and X. L. Meng, pp. 599–607, Oxford: Oxford University Press. [7]

Harrison, P. J., and Stevens, C. F. (1976), "Bayesian Forecasting," *Journal of the Royal Statistical Society*, Series B, 38, 205–247. [2]

Hobert, J. P. (2011), "The Data Augmentation Algorithm: Theory and Methodology," in *Handbook of Markov Chain Monte Carlo*, eds. S. Brooks, A. Gelman, J. Galin, and X. L. Meng, , pp. 253–294, Boca Raton, FL: CRC Press. [3]

Hobert, J. P., Royand, V., and Robert, C. P. (2011), "Improving the Convergence Properties of the Data Augmentation Algorithm with an Application to Bayesian Mixture Modeling," *Statistical Science*, 26, 332–351. [1]

Hoffman, M. D., and Gelman, A. (2014), "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo," *The Journal of Machine Learning Research*, 15, 1593–1623. [2,7]

Jacob, P. E., and Thiery, A. H. (2015), "On Nonnegative Unbiased Estimators," *The Annals of Statistics*, 43, 769–784. [4]

Kastner, G., and Frühwirth-Schnatter, S. (2014), "Ancillarity-Sufficiency Interweaving Strategy (ASIS) for Boosting MCMC Estimation of

Stochastic Volatility Models," *Computational Statistics & Data Analysis*, 76, 408–423. [2,10,11]

King, R. (2011), "Statistical Ecology," in *Handbook of Markov Chain Monte Carlo*, eds. S. Brooks, J. G. A. Gelman, and X. L. Meng, pp. 419–447, Boca Raton, FL: CRC Press. [11,12]

King, R., Brooks, S. P., Mazzetta, C., Freeman, S. N., and Morgan, B. J. T. (2008), "Identifying and Diagnosing Population Declines: a Bayesian Assessment of Lapwings in the UK," *Journal of the Royal Statistical Society*, Series C, 57, 609–632. [7]

King, R., McClintock, B. T., Kidney, D., and Borchers, D. (2016), "Capture–Rrecapture Abundance Estimation using a Semi-complete Data Likelihood Approach," *Annals of Applied Statistics*, 10, 264–285. [2,3]

Kitagawa, G. (1987), "Non-Gaussian State-Space Modeling of Nonstationary Time Series," *Journal of the American Statistical Association*, 82, 1032–1041. [5,15]

Knaus, P., Bitto-Nemling, A., Cadonna, A., and Frühwirth-Schnatter, S. (2021), "Shrinkage in the Time-Varying Parameter Model Framework using the R package shrinkTVP," *Journal of Statistical Software*, 100, 1–32. [7,10,16]

Korattikara, A., Chen, Y., and Welling, M. (2014), "Austerity in MCMC Land: Cutting the Metropolis-Hastings Budget," in *Proceedings of the 31st International Conference on International Conference on Machine Learning* (Vol. 32), pp. 181–189. [4]

Kroese, D., and Chan, J. (2013), *Statistical Modeling and Computation*, New York: Springer. [7]

Langrock, R., and King, R. (2013), "Maximum Likelihood Estimation of Mark–Recapture–Recovery Models in the Presence of Continuous Covariates," *The Annals of Applied Statistics*, 7, 1709–1732. [5]

Langrock, R., King, R., Matthiopoulos, J., Thomas, L., Fortin, D., and Morales, J. M. (2012a), "Flexible and Practical Modeling of Animal Telemetry Data: Hidden Markov Models and Extensions," *Ecology*, 93, 2336–2342. [5]

Langrock, R., MacDonald, I. L., and Zucchini, W. (2012b), "Some Nonstandard Stochastic Volatility Models and their Estimation using Structured Hidden Markov Models," *Journal of Empirical Finance*, 19, 147–161. [2,5,6,15]

Li, M., and Koopman, S. J. (2021), "Unobserved Components with Stochastic Volatility in U.S. Inflation: Estimation and Signal Extraction," *Journal of Applied Econometrics*, 36, forthcoming. [7,8]

Maclaurin, D., and Adams, R. P. (2014), "Firefly Monte Carlo: Exact MCMC with Subsets of Data," *in Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI 2014)*, pp. 543–552. [2]

Marin, J. M., and Robert, C. (2007), *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*, New York: Springer. [2,3]

McCausland, W. J., Miller, S., and Pelletier, D. (2011), "Simulation Smoothing for State–Space Models: A Computational Efficiency Analysis," *Computational Statistics & Data Analysis*, 55, 199–212. [11]

Murphy, K. P. (2002), "Dynamic Bayesian Networks: Representation, Inference and Learning," PhD thesis, University of California, Berkeley. [4]

Neal, R. (2011), "MCMC Using Hamiltonian Dynamics," in *Handbook of Markov Chain Monte Carlo*, eds. S. Brooks, A. Gelman, J.Galin, and X. L. Meng, pp. 116–162, Boca Raton, FL: CRC Press. [2,7]

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2007), "A General Framework for the Parametrization of Hierarchical Models," *Statistical Science*, 22, 59–73. [1]

Plummer, M. (2017), "JAGS Version 4.3.0 user manual," Technical report, Available at *http://mcmc-jags.sourceforge.net/*. [2,7,11]

Primiceri, G. E. (2005), "Time Varying Structural Vector Autoregressions and Monetary Policy," *The Review of Economic Studies*, 72, 821–852. [5]

Quiroz, M., Kohn, R., Villani, M., and Tran, M.-N. (2019), "Speeding Up MCMC by Efficient Data Subsampling," *Journal of the American Statistical Association*, 114, 831–843. [2]

Roberts, G. O., and Rosenthal, J. S. (2001), "Optimal Scaling for Various Metropolis-Hastings Algorithms," *Statistical Science*, 16, 351–367. [7]

Sandmann, G., and Koopman, S. J. (1998), "Estimation of Stochastic Volatility Models via Monte Carlo Maximum Likelihood," *Journal of Econometrics*, 87, 271–301. [15]

Shephard, N. (2015), *Martingale Unobserved Component Models*, pp. 218–249, Oxford: Oxford University Press. [8]

Stan Development Team (2019), "Stan Modeling Language Users Guide and Reference Manual, Version 2.25," Technical report, Available at *https://mc-stan.org*. [2,7]

Stock, J. H., and Watson, M. W. (2007), "Why has U.S. Inflation Become Harder to Forecast?" *Journal of Money, Credit and Banking*, 39, 3–33. [7,8]

Strickland, C. M., Martin, G. M., and Forbes, C. S. (2008), "Parameterisation and Efficient MCMC Estimation of non-Gaussian State Space Models," *Computational Statistics & Data Analysis*, 52, 2911–2930. [1]

Strickland, C. M., Turner, I. W., Denham, R., and Mengersen, K. L. (2009), "Efficient Bayesian Estimation of Multivariate State Space Models," *Computational Statistics & Data Analysis*, 53, 4116–4125. [2]

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–540. [1,3]

West, M., and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models* (2nd ed.), New York: Springer-Verlag. [1]

Yu, Y., and Meng, X.-L. (2011), "To Center or Not to Center: That is not the Question-An Ancillarity-Sufficiency Interweaving Strategy (ASIS) for Boosting MCMC Efficiency," *Journal of Computational and Graphical Statistics*, 20, 531–570. [2]

Zucchini, W., MacDonald, I. L., and Langrock, R. (2016), *Hidden Markov Models for Time Series: An Introduction Using R*, Monographs on Statistics and Applied Probability 150 (2nd ed.), Boca Raton, FL: CRC Press. [3,6]