Preliminary Draft. Please do not cite, circulate or quote without the authors' permission

Semi-Complete Data Augmentation for Efficient State Space Model Fitting

Agnieszka Borowska[(a,b)] and Ruth King[(c)]

[(a)] School of Business and Economics, Vrije Universiteit Amsterdam and Tinbergen Institute, The Netherlands

[(b)] School of Mathematics, University of Edinburgh, U.K.

November 28, 2018

Abstract

A novel efficient model-fitting algorithm for state space models is proposed. 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, comes at the price of substantially more complicated fitting of such models to data as the associated likelihood is typically analytically intractable. 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, leading to the need for specialist algorithms. The proposed method circumvents the inefficiencies of the previous approaches by combining data augmentation with numerical integration in a Bayesian hybrid approach. This approach permits standard "vanilla" updating algorithms that perform considerably better than the traditional approach in terms of considerably improved mixing and hence lower correlation. A proposed semi-complete data augmentation algorithm is used in different application areas and associated types of models, leading to distinct implementation schemes and demonstrating efficiency gains in empirical studies.

Keywords: Bayesian inference; Markov chain Monte Carlo; data augmentation; numerical integration; effective sample size.

1 Introduction

The task of inference about a latent state governing the dynamics of the system under study given only the observed noisy data is ubiquitous in many contexts, e.g. 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), cf. Durbin and Koopman (2012) for a detailed exposition; see also West and Harrison (1997) for the treatment focused on the Bayesian perspective. They 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 modelling approach.

This flexibility, however, typically comes at the price of substantially more complicated fitting of such models to data. For the general non-linear non-Gaussian SSM the associated likelihood is analytically

intractable so that no closed-form solution is available to the optimal estimation problem. Only in certain circumstances the associated likelihood can be calculated explicitly: for linear Gaussian systems the Kalman filter provides the optimal state estimator; for hidden Markov models specified on a discrete state space the likelihood may be available in a closed-form (but may become infeasible for a large number of states). In this paper we focus on models for which the likelihood is intractable or for which it may be infeasible 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, cf. Tanner and Wong (1987). The former group includes the sequential Monte Carlo (SMC) methods, cf. Doucet et al. (2001) for an extensive review, which can be used for parameter estimation within a standard Markov chain Monte Carlo (MCMC) algorithm (i.e. particle MCMC, cf. Andrieu et al., 2010). Provided the corresponding likelihood estimator is unbiased, the convergence to the correct posterior is guaranteed by the pseudo marginal theory (Beaumont, 2003; Andrieu and Roberts, 2009). In general, numerical integration provides a limited solution, feasible only for very low dimensional systems. The latter DA approach has become a standard method for inference for SSMs within a Bayesian framework, cf. Frühwirth-Schnatter (1994, 2004); Hobert (2011). DA relies on the true unknown states being treated as auxiliary variables and imputed within the MCMC algorithm. However, the general Bayesian DA approach implemented using standard "vanilla" MCMC algorithms may perform very poorly due to high correlation between the imputed states, cf. Hobert et al. (2011) and the references therein. This leads to the need to specialist, model-specific algorithms being developed.

We propose a novel efficient model-fitting algorithm to circumvent these inefficiencies by combining DA with numerical integration in a Bayesian hybrid approach, where the associated standard "vanilla" algorithms perform substantially more efficiently. The underlying idea is to combine the "good" aspects of both methods by minimising the problems that arise for each, i.e. highly correlated latent states for DA and the curse of dimensionality for numerical integration. To this end, we utilise the structure of the unknown states which can be split into two types: auxiliary variables, which are imputed within the MCMC algorithm using DA; and "integrable" states, which are numerically integrated out within the likelihood expression. We specify the unknown states in such a way that the algorithm is efficient where the imputed states have limited/reduced correlation and the numerical integration is over a very low number of dimensions.

The structure of the paper is a follows. Section 2 presents the general SSM specification and discusses the previous approaches to fit these general models to data. Section 3 introduces the proposed semi-complete data augmentation approach while Section 4 develops a general HMM-based approximation to the associated likelihood. We demonstrate the efficiency gains from the new method in Section 5, where we discuss two empirical applications relating to the abundance estimation for the ecological data, and to the estimation of the stochastic volatility (SV) for financial data. Section 6 concludes with a discussion.

2 State space models

Consider a state space model of the form:

$$\boldsymbol{y}_t|\boldsymbol{x}_t,\boldsymbol{\theta} \sim p(\boldsymbol{y}_t|\boldsymbol{x}_t,\boldsymbol{\theta}), \tag{2.1}$$

$$\boldsymbol{x}_{t+1}|\boldsymbol{x}_t,\boldsymbol{\theta} \sim p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t,\boldsymbol{\theta}), \tag{2.2}$$

$$\boldsymbol{x}_0|\boldsymbol{\theta} \sim p(\boldsymbol{\theta}), \tag{2.3}$$

for $t = 1, \ldots, T$, with $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$, $\boldsymbol{y}_t \in \mathcal{Y}$, denoting a time series of observations (potentially multivariate, although in our examples they are univariate), $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ a series of latent states (with $\boldsymbol{x}_t = [x_{1,t}, \ldots, x_{D,t}]^T$ potentially multivariate, $x_{d,t} \in \mathcal{X}_d$) and $\boldsymbol{\theta}$ the model static parameters for which we put a prior $p(\boldsymbol{\theta})$. T denotes the length of the time series and $D < \infty$ the dimension of the state $\boldsymbol{x}_t$. To simplify notation, below 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 (2.2). The observation process which generates $\boldsymbol{y}_t$, the observed data given the true underlying states, is specified by (2.1). 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 2.1 graphically presents the dependencies between states and observations in the SSM. An extensive discussion of SSMs is provided by Durbin and Koopman (2012) and also Cappé et al. (2006), where this class of models is called hidden Markov models (HMM)[1].

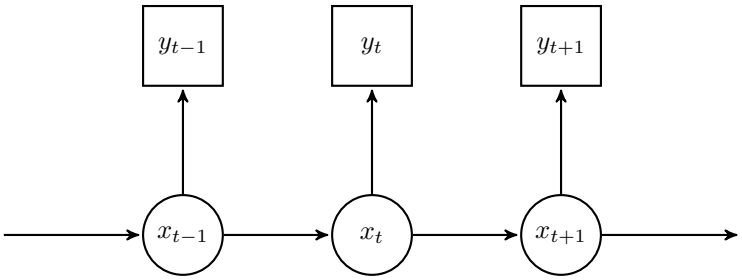


Figure 2.1: A graphical representation of the general first-order SSM.

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 (2.1)–(2.3) is given by

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \int p(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x} = \int p(x_0|\boldsymbol{\theta}) \prod_{t=1}^{T} p(y_t|x_t, \boldsymbol{\theta}) p(x_t|x_{t-1}, \boldsymbol{\theta}) dx_0 dx_1 \ldots dx_T, \tag{2.4}$$

and typically is not available in closed form. This is due to the integration over the latent variables, which is difficult to calculate, despite the tractability of the joint distribution of the data and the auxiliary variables $p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\theta})$. The latter is 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 (2.2). This likelihood can be efficiently calculated using the forward algorithm (see Zucchini et al., 2016). However, for systems with multiple processes there may be a very large number of possible states. This can lead to the approach being infeasible due to the curse of dimensionality. In addition, such an approach becomes infeasible even for simple systems, with e.g. only 2 processes, but with many potential state outcomes (i.e. when $\dim(\mathcal{X}_d)$ is "large").

To overcome the problem of the intractable likelihood, the standard DA technique is commonly adopted, see Tanner and Wong (1987); Frühwirth-Schnatter (1994, 2004); Hobert (2011). In DA the unknown states $\boldsymbol{x}$ are treated as auxiliary variables and imputed[2]. This way one can work with the closed-from complete data likelihood

$$p(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta}) = p(x_0|\boldsymbol{\theta}) \prod_{t=1}^{T} p(y_t|x_t, \boldsymbol{\theta}) p(x_t|x_{t-1}, \boldsymbol{\theta}).$$

[1]The terminology is not fully consistent in this context: the term "HMM" is sometimes used only for SSMs with a finite state space, i.e. $\dim(\mathcal{X}_d) < \infty$. This convention is used by e.g. Zucchini et al. (2016).

[2]A similar idea underlies the Expectation Maximisation algorithm of Dempster et al. (1977) in the classical framework.

In the Bayesian framework, the complete data likelihood is used to construct the joint posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{x}$

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

Then an MCMC algorithm (or other) can be employed to draw from the joint posterior distribution and the generated values of $\boldsymbol{\theta}$ are taken as a sample from the (marginal) posterior distribution of interest $p(\boldsymbol{\theta}|\boldsymbol{y})$. 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, Ch. 4).

DA is a powerful tool for dealing with intractable likelihoods, however it 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 models which impose a strong dependence structure on the latent variables and parameters. Single-update algorithms can perform especially poorly, nevertheless they are often used as they are presumably the most general and the easiest to implement. An alternative approach based on block sampling, i.e. simultaneously updating the target distribution in multiple dimensions, can lead to an improved mixing. However, it requires defining an appropriate partition of the states and parameters into blocks and specifying an efficient proposal distributions for each block. These problems of the standard DA approach often result in specialist algorithms being developed for the purpose of efficient estimation of a given model. Consequently, bespoke codes need to be written dependent on model and data.

3 Semi-complete data augmentation

To improve the efficiency of the standard DA approach, we propose to combine DA with numerical integration within a Bayesian hybrid framework, which we call *Semi-complete data augmentation*. A key idea is to separate the latent state $\boldsymbol{x}$ into two components $\boldsymbol{x} = (\boldsymbol{x_{aug}}, \boldsymbol{x_{int}})$. We will refer to $\boldsymbol{x_{int}}$ and $\boldsymbol{x_{aug}}$ as the "integrated" states and the "augmented" states, respectively. The starting point for our method is to specify the *semi-complete data likelihood* (SCDL) $p(\boldsymbol{y}, \boldsymbol{x_{aug}}|\boldsymbol{\theta})$ as follows:

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{x_{aug}}|\boldsymbol{\theta}) &= \int p(\boldsymbol{y}, \boldsymbol{x_{aug}}, \boldsymbol{x_{int}}|\boldsymbol{\theta}) d\boldsymbol{x_{int}} \\ &= \int p(\boldsymbol{y}|\boldsymbol{x_{aug}}, \boldsymbol{x_{int}}, \boldsymbol{\theta}) p(\boldsymbol{x_{aug}}, \boldsymbol{x_{int}}|\boldsymbol{\theta}) d\boldsymbol{x_{int}}. \end{aligned} \tag{3.1}$$

The joint posterior distribution of the parameters and augmented states can be then expressed as

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{x_{aug}}|\boldsymbol{y}) &\propto p(\boldsymbol{y}, \boldsymbol{x_{aug}}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \\ &= p(\boldsymbol{y}|\boldsymbol{x_{aug}}, \boldsymbol{\theta}) p(\boldsymbol{x_{aug}}|\boldsymbol{\theta}) p(\boldsymbol{\theta}). \end{aligned}$$

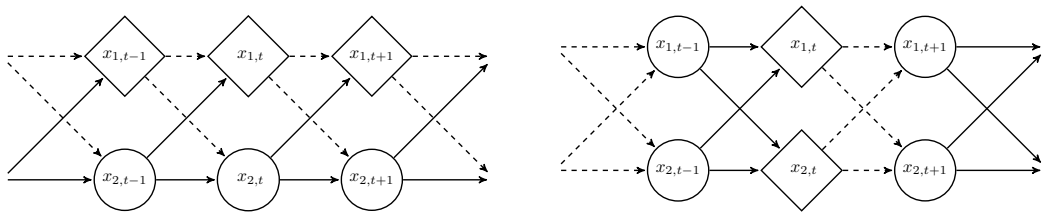
We note that our approach builds upon the work of King et al. (2016), who propose a Bayesian hybrid approach applied to the particular case of capture-recapture data. These authors define the "semi-complete" data likelihood as the product of a complete data likelihood component for the individuals observed within the study (related to $\boldsymbol{x_{aug}}$) and a marginal data likelihood component for the unobserved individuals (related to $\boldsymbol{x_{int}}$). We extend their approach to the general state space models framework and consider different schemes for specifying the semi-complete data likelihood in terms of defining $\boldsymbol{x_{aug}}$ and $\boldsymbol{x_{int}}$.

Specification of the auxiliary variables More precisely, consider a time series $\boldsymbol{x} = \{\boldsymbol{x}_t\}_{t=0}^{T}$ of length $T+1$, where the state at time t is D dimensional: $\boldsymbol{x}_t = \left[x_{1,t}, \ldots, x_{D,t}\right]^T$, for $t = 0, 1, \ldots, T$.

We want to integrate out D_{int} dimensions of the state at time points T_{int}, where $D_{int} \subset \{1, \dots, D\}$ are $T_{int} \subset \{0, 1, \dots, T\}$ are "suitably" chosen subsets of dimension and time indices, respectively. Such a "suitable" specification of subsets D_{int} and T_{int} depends on the dependence structure of the model under consideration so that the implied integral can be efficiently calculated. For instance, it can be low dimensional or it can be reduced to a product of low-dimensional integrals. The compliments of both subsets are denoted D_{aug} and T_{aug}, respectively. We also denote T^+_{int} and T^+_{aug} the corresponding sets without the initial observations, i.e. excluding time $t = 0$. The "integrated" and "augmented" states are then defined as the partition of $\boldsymbol{x}$ into $\boldsymbol{x_{int}} = \{x_{d,t}\}_{d \in D_{int}, t \in T_{int}}$ and $\boldsymbol{x_{aug}} = \{x_{d,t}\}_{d \in D_{aug}, t \in T_{aug}}$, where we denote their corresponding elements at time t by $\boldsymbol{x_{int,t}} = \{x_{d,t}\}_{d \in D_{int}}$ and $\boldsymbol{x_{aug,t}} = \{x_{aug,t}\}_{d \in D_{int}}$, respectively. In particular, we give the following two examples of integration/augmentation schemes.

(a) *"Horizontal" integration:* e.g. for a $D = 2$ dimensional state we integrate out the second state at all time periods, so that $D_{int} = \{2\}$ (and hence $D_{aug} = \{1\}$), and $T_{int} = \{0, 1, \dots, T\}$ (and hence $T_{aug} = T_{int}$), see Figure 3.1a. We use this scheme in the lapwings data application in Section 5.1.

(b) *"Vertical" integration:* e.g. all D states are integrated out at odd time periods $D_{int} = \{1, \dots, D\}$ and $T_{int} = \{2t + 1\}_{t=0}^{\lfloor T/2 \rfloor}$ (and hence $T_{aug} = \{2t\}_{t=0}^{\lfloor T/2 \rfloor}$), see Figure 3.1b. We use this scheme in the SV model application in Section 5.2, for $D = 1$ dimensional state.



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

Figure 3.1: 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.

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 $1, 2, \dots, |T^+_{int}|$. 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. $a(t) = \emptyset$). Then the subsequent integrated and augmented states are given by $\dots, x_{int,\tau(t-1)}, x_{int,\tau(t)}, x_{int,\tau(t+1)}, \dots$ and $\dots, \boldsymbol{x}_{aug,a(t-1)}, \boldsymbol{x}_{aug,a(t)}, \boldsymbol{x}_{aug,a(t+1)}, \dots$, respectively, for $t = 1, 2, \dots, |T^+_{int}|$. 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, i.e. allowing us to iterate over the set of observation indices $\{1, \dots, T\}$ using the same index as to iterate over T_{int} and T_{aug}. Therefore, we want the image of $o(t)$ to cover $\{1, \dots, T\}$, the whole set of indices of y_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 $\{1, \dots, T\}$ to the iterating variable t. To this end, we allow $o(t)$ to take values in the power set of $T_{int} \cup T_{aug}$. For illustration, consider vertical integration (b) together with conditionally independent observations $y_t | \boldsymbol{x}_t \sim p(y_t | \boldsymbol{x}_t)$. Because for $t = 1, 2, \dots, |T^+_{int}|$ we consider states in two different time periods, i.e. at period $\tau(t) = 2t + 1$ for the integrated states and at period $a(t) = 2t$ for the imputed states, for each t we need to account for two different observations, $y_{\tau(t)}$ and $y_{a(t)}$. This means that $o(t) = \{2t, 2t + 1\}$. In the case of horizontal integration (a) $T_{int} = T_{aug}$ so we simply set $o(t) = t$.

In order to identify conditionally independent latent states to "integrate out", one can use the graphical structure of the problem: Figure 2.1 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. (2000a) note that the "tractable structure" of some state space models might by analytically marginalised out given imputed other nodes.

Rao-Blackwellisation We note that integrating out, or "marginalising out", some of the variables is a case of the general technique known as *Rao-Blackwellisation*, which relies on the Rao-Blackwell formula. Suppose that we are interested in a function f of two random vectors z_1 and z_2, and let $\hat{f}$ be an estimator of f. Then

$$\mathrm{Var}[\hat{f}(z_1, z_2)] = \mathrm{Var}[\underbrace{\mathbb{E}[\hat{f}(z_1, z_2)|z_2]}_{=:\hat{f}'}] + \underbrace{\mathbb{E}[\mathrm{Var}[\hat{f}(z_1, z_2)|z_2]]}_{(*)},$$

which implies that $\hat{f}'$ has the same expected value as $\hat{f}$ but a lower variance than $\hat{f}$ by an additive factor of $(*)$. Rao-Blackwellisation was introduced to the MCMC literature by Gelfand and Smith (1990) in their seminal paper on the Gibbs sampler to become a commonly applied tool for variance reduction of integral approximations. In general context of sampling schemes, Rao-Blackwellisation was further analysed by Casella and Robert (1996), whose approach was then used by Douc and Robert (2011) for improving efficiency of the MH algorithm and by Doucet et al. (2000b,a) to enhance particle filters. Durbin and Koopman (2012, Ch. 12) note that in the context of state space models z_2, i.e. the variable being integrated out, is not a *sufficient statistic*, hence the term "Rao-Blackwellisation" is not fully appropriate since the the Rao-Blackwell theorem concerns the case when z_2 is a sufficient statistic for z_1. We employ the Rao-Blackwell principle for DA in the context of state space models.

Approximate marginal likelihood Recall that the joint posterior distribution over $\boldsymbol{\theta}$ and $\boldsymbol{x_{aug}}$ can be expressed in terms of the SCDL

$$p(\boldsymbol{\theta}, \boldsymbol{x_{aug}}|\boldsymbol{y}) \propto p(\boldsymbol{y}, \boldsymbol{x_{aug}}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

However, the SCDL $p(\boldsymbol{y}, \boldsymbol{x_{aug}}|\boldsymbol{\theta})$ may still be analytically intractable so that we need to estimate it using simulation-based techniques. Suppose we have a sample of length N of unknown variables of interest (i.e. $\boldsymbol{\theta}$ and $\boldsymbol{x_{aug}}$). Here, N is a number of points used for integration: for a deterministic integration it is the number of grid points, for a stochastic, i.e. Monte Carlo (MC), integration it is a number of draws. We can use such a sample to compute $\hat{p}_N(\boldsymbol{y}, \boldsymbol{x_{aug}}|\boldsymbol{\theta})$, the N-sample estimate of the SCDL, and consequently to approximate the posterior distribution in the following way

$$\hat{p}_N(\boldsymbol{\theta}, \boldsymbol{x_{aug}}|\boldsymbol{y}) \propto \hat{p}_N(\boldsymbol{y}, \boldsymbol{x_{aug}}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

We set $\hat{p}_N(\boldsymbol{y}, \boldsymbol{x_{aug}}|\boldsymbol{\theta})$ such that

$$\hat{p}_N(\boldsymbol{y}, \boldsymbol{x_{aug}}|\boldsymbol{\theta}) \overset{N\to\infty}{\to} p(\boldsymbol{y}, \boldsymbol{x_{aug}}|\boldsymbol{\theta}),$$

so that

$$\hat{p}_N(\boldsymbol{\theta}, \boldsymbol{x_{aug}}|\boldsymbol{y}) \overset{N\to\infty}{\to} p(\boldsymbol{\theta}, \boldsymbol{x_{aug}}|\boldsymbol{y}).$$

Further properties of the resulting likelihood estimator depend in general on the approximation scheme, which in turn determine the properties of the corresponding MCMC algorithm. If $\mathbb{E}[\hat{p}_N(\boldsymbol{y}, \boldsymbol{x_{aug}}|\boldsymbol{\theta})] = p(\boldsymbol{y}, \boldsymbol{x_{aug}}|\boldsymbol{\theta})$ standard MH algorithms converge to $p(\boldsymbol{\theta}, \boldsymbol{x_{aug}}|\boldsymbol{y})$, which follows from the pseudo-marginal argument. The pseudo-marginal theory, originated by Beaumont (2003), further developed by Andrieu and Roberts (2009) and popularised by Andrieu et al. (2010) (who called their method the particle

MCMC, PMCMC), guarantees that an MCMC scheme based on the unbiased (marginal) likelihood estimator converges to the exact posterior distribution [3]. Such an unbiased likelihood estimator is delivered by e.g. MC integration, in which the integral is evaluated at random points. Hence, whether the resulting MCMC algorithm is "exact approximate" or "just approximate" depends on whether the approximate likelihood is an unbiased estimator of the marginal likelihood.

For fixed points, such as in a quadrature, obtaining of an "exact approximate" algorithm is not guaranteed but the resulting approximation converges to the true value as $N \to \infty$. It means that a "just approximate" algorithm can be made arbitrarily close to the true integral by considering sufficiently many samples to construct the estimator. Additionally, we note that unbiased estimators might be characterised by large MC errors, particularly for a small number of samples, see e.g. Korattikara et al. (2014), Jacob and Thiery (2015). The choice between different likelihood approximation methods fits into the traditional discussion on the bias-variance trade-off. However, as pointed out by Robert (2016), especially from a Bayesian perspective unbiasedness is a "second order property" [4].

4 Approximations for MCMC sampling

Below we consider possible ways for obtaining an estimate $\hat{p}_N(\boldsymbol{y}, \boldsymbol{x_{aug}}|\boldsymbol{\theta})$. In particular, we focus on the case when it 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 be made such that this condition is satisfied. There exist several methods to numerically estimate a single one dimensional integral: (1) quadrature with fixed nodes; (2) quadrature with adaptive nodes; (3) stochastic (MC) integration. The two former approaches 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 e.g. by Langrock et al. (2012a,b); Langrock and King (2013). For the latter MC approach the resulting estimator of the complete data likelihood is unbiased $\mathbb{E}(\hat{p}_N(\boldsymbol{y}, \boldsymbol{x_{aug}}|\boldsymbol{\theta})) = p(\boldsymbol{y}, \boldsymbol{x_{aug}}|\boldsymbol{\theta})$. Hence, the pseudo-marginal argument guarantees that the chain generated with a standard MH algorithm (using the estimate $\hat{p}_N(\boldsymbol{y}, \boldsymbol{x_{aug}}|\boldsymbol{\theta})$) converging to the exact posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{x_{aug}}|\boldsymbol{y})$ in this case; see Beaumont (2003), Andrieu and Roberts (2009) and Andrieu et al. (2010).

We note that in low dimensions all of 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 an algorithm.

4.1 Approximation bins as hidden Markov model states

Below we discuss 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 $\boldsymbol{x}_{int,\tau(t)}$ is univariate so we can write $x_{int,\tau(t)}$. For multivariate $\boldsymbol{x}_{int,\tau(t)}$ we may consider separate bins for each integrated state dimension $d \in D_{int}$ at

[3] PMCMC algorithms are thus called "exact approximate". Note that they are the extreme case of our approach with $\boldsymbol{x_{int}} = \boldsymbol{x}$.

[4] There are two obvious reasons for that. First, the concept of a bias is conditional on the true value of a parameter, which is unknown (Gelman, 2011). Second, unbiasedness cannot be achieved for most transformations of the model parameter vector and is not preserved under reparameterisation (Robert, 2016).

time $\tau(t)$. We then 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[5].

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. Discretising of the state space to perform numercial integration dates back to Kitagawa (1987) and was discussed in Zucchini et al. (2016). The associated approximate posterior distribution can be made arbitrarily accurate by increasing the number of bins (quadrature points).

The idea is to split the state space $\mathcal{X}_{int}$ of the state to be integrated out into B bins of length k (for integer-valued variables we assume $k \in \mathbb{N}$) and to consider e.g. the midpoints of the bins 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 et al. (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 as follows:

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

For continuous variables $\mathcal{B}_i$ is simply a continuous interval of length k, while for discrete variables it consists of k subsequent integers, e.g. 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 then define $\{z_t\}$, $t \in 1, \dots, T^*$, as a B-state, discrete-time (not necessarily homogeneous) Markov chain[6] with transition probabilities $\gamma_{jk,t} = \mathbb{P}(z_t = k | z_{t-1} = j)$ defined as

$$\gamma_{jk,t} := \mathbb{P}(x_{int,\tau(t)} \in \mathcal{B}_k | x_{int,\tau(t-1)} \in \mathcal{B}_j, \dots).$$

Then a transition of $z_{t-1} = j$ to $z_t = k$ is equivalent to $x_{int,\tau(t)}$ "falling into" bin k given $x_{int,\tau(t-1)}$ was in bin j and given $\boldsymbol{x_{aug}}$. For computationally intensive probabilities we can further approximate these as $\tilde{\gamma}^*_{jk,t} := p(b_k^* | b_j^*, \dots)$, which for discrete variables means $\mathbb{P}(x_{int,\tau(t)} = b_k^* | x_{int,\tau(t-1)} = b_j^*, \dots)$. To get the valid probability values (i.e. summing up to one) we normalise 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. We can alternatively compute the transition probabilities between bins directly, by integrating with respect to the required ranges as follows

$$\mathbb{P}(x_{int,\tau(t)} \in \mathcal{B}_k | x_{int,\tau(t-1)} \in \mathcal{B}_j, \dots) \propto \int_{\mathcal{B}_k} \int_{\mathcal{B}_j} p(x_{int,\tau(t)} | x_{int,\tau(t-1)}, \dots) dx_{int,\tau(t-1)} dx_{int,\tau(t)}.$$

However, typically such an analytical integration will only be possible in simple cases, e.g. discrete variables. One can visualise this method by considering small squares of a bigger transition matrix instead each its element separately.

[5] We note that from the perspective of the original process $\{\boldsymbol{x}\}$ the process we want to integrate out $\{\boldsymbol{x_{int}}\}$ will not be a Markov chain due to its potential dependence on the imputed states $\{\boldsymbol{x_{aug}}\}$. However, since we know the latter, conditioning on them can be understood as adopting a time-varying transition probabilities for $\{\boldsymbol{x_{int}}\}$, parametrised with relevant $\{\boldsymbol{x_{aug}}\}$.

[6] Even though we hardly refer to $\{z_t\}$ explicitly later in the text, they are useful to understand the introduced construction relating the potentially continuously valued process of interest $x_{int,\tau(t)}$ to a finite state HMM z_t. Such an exposition is inspired by Langrock et al. (2012b, Section 2.2).

Adaptive bins An alternative approach to fixed width binning is to use adaptive intervals which do not require any limiting of the integration range. This can be done by transforming the variable of interest to the $[0,1]$ range by applying a cdf. Then the bins can be specified on the $[0,1]$ interval and their limits or midpoints can be transformed back to obtain the values needed for the approximation of the original variable of interest. In particular, quantiles of the distribution associated with the variable of interest can be used. Then instead of specifying the grid points we fix the probabilities for each bin, which previously needed to be determined. This means a quantile determination problem which are needed e.g. to obtain the midpoint values used in conditioning.

Suppose that $x_{int,\tau(t)} \sim p(\vartheta_{\tau(t)})$, $\tau(t) \in T_{int}$, 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 $\boldsymbol{q} = [q_0, q_2, \ldots, q_B]$. The corresponding B mid-quantiles, denoted $\boldsymbol{q}^* = [q_1^*, q_2^*, \ldots, q_B^*]$, are given by $q_i^* = \frac{q_{i-1}+q_i}{2}$ (for instance, $\boldsymbol{q} = [0.0, 0.1, 0.2, \ldots, 1.0]$ and $\boldsymbol{q}^* = [0.05, 0.15, \ldots, 0.95]$). For $F(\vartheta_t)$ continuous and strictly monotonically increasing (such as a normal cdf) the bin midpoints at time t are determined by the mid-quantiles as follows

$$b_i^* = F^{-1}\left(q_i^* | \vartheta_{\tau(t)}\right).$$

For discrete variables one can either use the generalized inverse distribution function, or use a continuous approximation to the associated discrete distribution. For instance for a Poisson variable with a high enough mean, the normal approximation could be adopted. We note in general, the adaptive approach can be easily implemented in any programming language or software for statistical computing.

4.2 Hidden Markov model likelihood

Having specified the states of the underlying Markov chain in Section 4.1, we aim to use them to approximate the joint SCDL (3.1) 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 SCDL integral (5.15) with respect to the "integration time" and thus allows us to cast it into a likelihood of an HMM. Note that without any form of DA the likelihood can be simply decomposed by making use of the Markov property of the original state process $\boldsymbol{x}_t$, i.e. $p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} = \int \boldsymbol{x}_0 \prod_{t=1}^{T} p(y_t|\boldsymbol{x}_t)p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})d\boldsymbol{x}_0 d\boldsymbol{x}_1 \ldots d\boldsymbol{x}_T$.

Motivating example For illustration, consider the state specification from Figure 3.1a to which we add conditionally independent observations to result in an SSM presented in Figure 4.1. Such a system is representative for e.g. dynamic factor models (linear or nonlinear), with y_t multivariate, broadly applied in macroeconometrics and finance; it was also used by e.g. Abadi et al. (2010) to model population dynamics of little owl.

We specify $\boldsymbol{x_{aug}} = \{x_{1,t}\}_{t=0}^{T} =: \boldsymbol{x_1}$ (state 1) and $\boldsymbol{x_{int}} = \{x_{2,t}\}_{t=0}^{T} =: \boldsymbol{x_2}$ (state 2), which corresponds to the "horizontal" integration. Hence we put $T_{int} = T_{aug} = \{0, 1, \ldots, T\}$, $\tau(t) = t$, $a(t) = t$ and $o(t) = t$. We denote $T^* = |T_{int}^+|$. Using the temporal dependence in this system, the SCDL $p(\boldsymbol{y}, \boldsymbol{x_{aug}})$ can be expressed as

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

which is not tractable without integrating out $\boldsymbol{x_2}$. Hence, we marginalise over $\boldsymbol{x_2}$ and aim at approxi-

mating the resulting integral using a quadrature based on B bins $\mathcal{B}_k$, $k = 1, \ldots, B$, as follows

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{x_{aug}}) = \int \ldots \int p(x_{1,0}) p(x_{2,0}) \prod_{t=1}^{T^*} & \; p(y_{o(t)} | x_{1,a(t)}, x_{2,\tau(t)}) p(x_{1,a(t)} | x_{1,a(t-1)}, x_{2,\tau(t-1)}) \\ & p(x_{2,\tau(t)} | x_{1,a(t-1)}, x_{2,\tau(t-1)}) dx_{2,\tau(T^*)} \ldots dx_{2,\tau(1)} \\ \approx \sum_{k_1=1}^{B} \cdots \sum_{k_{T^*}=1}^{B} p(x_{1,0}) p(x_{2,0}) \prod_{t=1}^{T^*} & p(y_{o(t)} | x_{1,a(t)}, x_{2,\tau(t)} \in \mathcal{B}_{k_t}) p(x_{1,a(t)} | x_{1,a(t-1)}, x_{2,\tau(t-1)} \in \mathcal{B}_{k_{t-1}}) \\ & p(x_{2,\tau(t)} \in \mathcal{B}_{k_t} | x_{1,a(t-1)}, x_{2,\tau(t-1)} \in \mathcal{B}_{k_{t-1}}). \end{aligned} \tag{4.1}$$

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

$$\Gamma_t = \left[\mathbb{P}(x_{2,\tau(t)} \in \mathcal{B}_k^* | x_{1,a(t-1)}, x_{2,\tau(t-1)} \in \mathcal{B}_l^*) \right]_{k,l=1,\ldots,B}, \tag{4.2}$$

for $t \in 1, 2, \ldots, T^*$. Next to the transition matrix, we need to specify two matrices for the "augmented data": one for the augmented states $\boldsymbol{x}_{aug}$ and one for the real observations $\boldsymbol{y}$. This is different compared to standard HMMs in which only the latter is used. We specify the likelihood matrices for the augmented states and the observation as follows

$$P_t = \text{diag} \left(p(x_{1,a(t)} | x_{1,a(t-1)}, x_{2,\tau(t-1)} \in \mathcal{B}_l^*) \right)_{l=1,\ldots,B}, \tag{4.3}$$

$$Q_t = \text{diag} \left(p(y_{o(t)} | x_{1,a(t)}, x_{2,\tau(t)} \in \mathcal{B}_k^*) \right)_{k=1,\ldots,B}. \tag{4.4}$$

Notice that for both the integrated and augmented states the conditioning is with respect to their previous realisations, whilst for the observations it is with respect to the current values of both states. The quadrature based approximation to the SCDL (4.1) can be then approximated as

$$\hat{p}_B(\boldsymbol{y}, \boldsymbol{x_{aug}}) = p(x_{1,0}) \boldsymbol{u}_0 \left(\prod_{t=1}^{T^*} P_t \Gamma_t Q_t \right) \mathbf{1}, \tag{4.5}$$

where $\boldsymbol{u}_0 = \left[\mathbb{P}(x_{2,0} \in \mathcal{B}_1) \quad \ldots \quad \mathbb{P}(x_{2,0} \in \mathcal{B}_B) \right]$ is the initial distribution of the Markov chain. Appendix A.1 presents the underlying SSM and the details of the derivations.

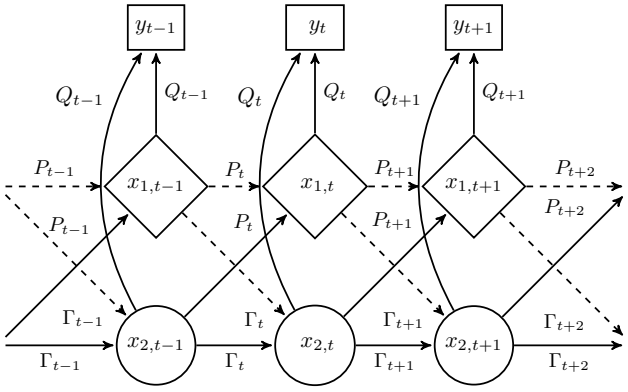


Figure 4.1: Illustration of combining DA and HMM structure. Conditionally independent observations added to the state specification from Figure 3.1a Diamonds represent the imputed states, circles – the integrated states. Dashed lines used for the relations *from* the imputed (known) states..

General formulation The generic matrices of the HMM-based approximation have the form

$$\Gamma_t = \left[\mathbb{P}(x_{int,\tau(t)} \in \mathcal{B}_k | x_{int,\tau(t-1)} \in \mathcal{B}_l, x_{aug,a(t-1)})\right]_{k,l=1,\dots,B},$$
$$P_t = \text{diag}\left(p(x_{aug,a(t)} | x_{int,\tau(t-1)} \in \mathcal{B}_l)\right)_{l=1,\dots,B}$$
$$Q_t = \text{diag}\left(p(\boldsymbol{y}_{o(t)} | x_{int,\tau(t)} \in \mathcal{B}_k, \boldsymbol{x}_{aug,a(t)})\right)_{k=1,\dots,B}$$

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

$$\hat{p}_B(\boldsymbol{y}, \boldsymbol{x_{aug}}) = p(x_{1,0})\boldsymbol{u}_0 Q_0 \left(\prod_{t=1}^{T^*} P_t \Gamma_t Q_t\right) \mathbf{1}, \tag{4.6}$$

which differs from (4.5) by including $Q_0 := \text{diag}\left(p(y_{o(0)} | x_{int,\tau(0)} \in \mathcal{B}_k^*)\right)_{k=1,\dots,B}$, which allows for a dependence of some observations on the initial state of the Markov process[7]. We require $\tau(t) \geq \max\{a(t)\}$ and $o(t) \subset \tau(t) \cup a(t)$, which is natural given the real dependencies in the original SSM (2.1)–(2.3).

5 Applications

In this section we consider applications of the proposed SCDA method and assess their performance. We consider two case studies with distinctively different features resulting in different integration schemes. The first application involves the dataset on the Northern lapwing (*Vanellus vanellus*), which has been extensively analysed in statistical ecology, see Besbeas et al. (2002), Brooks et al. (2004), King et al. (2008), and the references therein. We adopt the integrated population modelling approach of Besbeas et al. (2002), to be explained below, however our main focus is on modelling the abundance of the species via a state space model with discrete states. The second application relates to the well-known stochastic volatility model (SV), which is a popular tool to model time-varying volatility especially for financial time series, see Taylor (1994), Ghysels et al. (1996) or Shephard (1996). Further, we demonstrate how the SCDA framework can be easily adjusted to accommodate more complex properties of financial time series such as SV in the mean of Koopman and Uspensky (2002) or leverage effects, see Jungbacker and Koopman (2007).

Algorithm tuning In each case study we are interested in comparing the performance of the standard DA approach with that of the SCDA. To guarantee the between-method comparability, for each method we perform the estimation using a "vanilla" MH RW (single-update) algorithm. We tune each sampler so that the acceptance rates for each element of the parameter vector $\boldsymbol{\theta}$ and the average acceptance rates for each of the imputed states are "reasonable", i.e. between $20 - 40\%$.

Such a range corresponds to the seminal results of Gelman et al. (1996) and Roberts and Rosenthal (2001). The former authors prove that the asymptotically (as the dimension of the state space diverges to infinity) optimal mean acceptance rate is equal to 0.234 for a target distribution consisting of i.i.d. components and a normal proposal distribution of the same dimension as the target. Hence, they do not consider single-state updates (i.e. one-dimensional increments), for which the later authors deliver the optimal acceptance rate of 0.44 for a normal proposal distribution (see also Rosenthal, 2011). Generally, a mean acceptance rate of 20–40% is considered to deliver a well-performing chain.

Effective Sample Size Since the samples generated by MCMC algorithms are not independent, standard convergence results for independent MC sampling do not apply; in particular, the standard variance

[7]The SV model example in Appendix C demonstrates the role of Q_0.

estimator (i.e. the sample empirical variance) cannot be used to measure the variance of the empirical average delivered by an MCMC algorithm. The stochastic dependence in the (stationary) Markov chain $X_1, X_2, \ldots$ results in the associated asymptotic variance σ^2_{MCMC} taking account of the covariance in the Markov chain:

$$\begin{aligned}\sigma^2_{\text{MCMC}} &= \text{Var}[X_i] + 2\sum_{k=1}^{\infty} \text{Cov}[X_i, X_{i+k}] \\ &= \text{Var}[X_i] \underbrace{\left(1 + 2\sum_{k=1}^{\infty} \rho(k)\right)}_{\text{IF}},\end{aligned} \tag{5.1}$$

where $\rho(k)$ is the kth order serial correlation (Geyer, 2011). The term in the parentheses in (5.1) is referred to as the *autocorrelation function* (Geyer, 2011), *(integrated) autocorrelation time* (Robert and Casella, 2004, Ch. 12.3.5) or *inefficiency factor* (IF, Pitt et al., 2012), which is the name we use. High values of autocorrelation, typically reported for MCMC sampling, lead to the standard variance estimator underestimating the true variance σ^2_{MCMC}.

A common measure for assessing the deterioration in the sampling efficiency due to the draws autocorrelation is the *effective sample size* (ESS) defined as

$$\text{ESS} = \frac{M}{IF},$$

where M is the sample size (Robert and Casella, 2004, Ch. 12.3.5). It indicates what the size of an i.i.d. sample would be, had it the same variance as the MCMC sample. Equivalently, the IF gives the factor by which the "nominal" MCMC sample size would need to be increased in order to achieve the same accuracy as i.i.d. sampling.

In practice, one typically cannot compute the IF directly and needs to estimate it instead. As noted by Robert and Casella (2004, Ch. 12.3.5) estimation of IF is a "delicate issue", as it contains an infinite sum. A possible solution to this problem is set a cut-off value K for the autocorrelation terms being summed up: $\widehat{IF} = 1 + 2\sum_{k=1}^{K} \hat{\rho}(k)$. The choice of K of course poses the risk of subjectiveness; setting K to the lowest lag at which $\hat{\rho}(k)$ become insignificant seems to be a reasonable solution suggested by e.g. Kass et al. (1998) or Pitt et al. (2012) and this is the approach we take here.

5.1 Ecological model: lapwings data

We consider a time series of observations relating to census data (abundance index) of adult British lapwings (*Vanellus vanellus*), which we denote by $\boldsymbol{y} = (y_1, \ldots, y_T)$. The lapwings dataset plays an important role in statistical ecology and has served as an illustration in several handbooks (see King, 2011; King et al., 2010) and papers (Besbeas et al., 2002, e.g.) in this field. It was also used as an example of a complex statistical model by e.g. Goudie et al. (2018). We provide the details of this dataset in Appendix B. Figure 5.1 presents the data on the index of lapwings as well as on the normalised frost days, used as a covariate to describe the survival process. The latter is based on the number of days below freezing between April of year t and March of year $t+1$, inclusive and is a proxy for harshness of winter, which can affect the survival probability of wild birds more by lengthy cold periods rather than by low average temperature.

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, which we denote $\boldsymbol{N_1} = (N_{1,1}, \ldots, N_{1,T})$ and $\boldsymbol{N_a} = (N_{a,1}, \ldots, N_{a,T})$, respectively.

Figure 5.1: Lapwings census data and normalised frost days.

Hence, the latent state is given by $\boldsymbol{x} = \{\boldsymbol{N_1}, \boldsymbol{N_a}\}$. Following Besbeas et al. (2002) we model the count data via the following state space model:

$$y_t | N_{a,t}, \boldsymbol{\theta} \sim \mathcal{N}(N_{a,t}, \sigma_y^2), \tag{5.2}$$

$$N_{1,t+1} | N_{a,t}, \boldsymbol{\theta} \sim \mathcal{P}(N_{a,t}\rho_t\phi_{1,t}), \tag{5.3}$$

$$N_{a,t+1} | N_{1,t}, N_{a,t}, \boldsymbol{\theta} \sim \mathcal{B}\big((N_{1,t} + N_{a,t}), \phi_{a,t}\big), \tag{5.4}$$

$$N_{1,0} \sim \mathcal{NB}(r_{1,0}, p_{1,0}), \tag{5.5}$$

$$N_{a,0} \sim \mathcal{NB}(r_{a,0}, p_{a,0}), \tag{5.6}$$

for $t = 1, \ldots, T$, where $\mathcal{N}$, $\mathcal{P}$, $\mathcal{B}$ and $\mathcal{NB}$ stand for normal, Poisson, binomial and negative binomial distributions, respectively. The model is parametrised 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 $N_{i,0}$, $i \in \{1, a\}$.

Following Besbeas et al. (2002), we assume the following functional forms for the model time varying parameters

$$\begin{aligned}
\operatorname{logit} \phi_{1,t} &= \log\left(\frac{\phi_{1,t}}{1-\phi_{1,t}}\right) = \alpha_1 + \beta_1 f_t, \\
\operatorname{logit} \phi_{a,t} &= \log\left(\frac{\phi_{a,t}}{1-\phi_{a,t}}\right) = \alpha_a + \beta_a f_t, \\
\log \rho_t &= \alpha_\rho + \beta_\rho \tilde{t},
\end{aligned}$$

where f_t denotes the normalised value of frost days $fdays$ in year t and $\tilde{t}$ the normalised time index. As explained by King (2011), we introduce normalisation of f_t and $\tilde{t}$ to improve the mixing of the Markov chain and to facilitate the interpretation of the regression parameters.

To improve the estimation, Besbeas et al. (2002) propose using an additional source of information provided by the ring-recovery (RR) data, independent from the count series. The RR model shares with the SSM the survival parameters $\phi_{1,t}$ and $\phi_{a,t}$ but it does not involve the productivity rate ρ_t. Instead, the RR models includes the common time-varying recovery rate λ_t (denoting the probability that a bird which dies in year t is recovered), specified to be of the form

$$\operatorname{logit} \lambda_t = \log\left(\frac{\lambda_t}{1-\lambda_t}\right) = \alpha_\lambda + \beta_\lambda \tilde{t}_t.$$

Combining both models results in the so-called *integrated model*, which is parametrised by the regression parameters and the variance of the observation error. We refer to Besbeas et al. (2002) for a

more detailed description of the integrated model. The model parameters are 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, to complete the Bayesian specification of the model, we set independent vague priors being normal $\mathcal{N}(0, 100)$ for the logistic regression coefficient α_i and β_i, $i \in \{1, a, \rho, \lambda\}$, while for the observation variance σ_y^2 conjugate inverse gamma $\Gamma^{-1}(a_y, b_y)$ with $a_y = 0.001 = b_y$. For the initial states, we set the following values for the hyperparameters: for first-years $r_{1,0} = 4$ and $p_{1,0} = 0.98$ so that the prior mean and variance of 1-years are roughly 200 and $10,000$, respectively; for adults $r_{a,0} = 111$ and $p_{a,0} = 0.9$, so that the prior mean and variance adults are roughly $1,000$ and $10,000$, respectively.

System (5.2)–(5.6) is non-Gaussian and nonlinear with the associated likelihood unavailable in a closed form. It could be analysed using the normal approximation, which has an advantage that the Kalman filtering and smoothing techniques can be employed, see Besbeas et al. (2002). However, we aim at estimation of the original model, in which case the standard approach has been a DA approach. The problem with the standard DA approach is that it may lead to poorly mixing MCMC algorithms as demonstrated by King (2011). To this end, we first analyse the dependence structure in the model to select most promising states to integrate out.

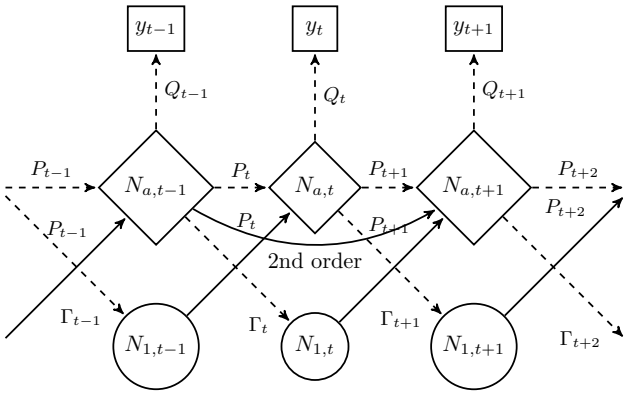


Figure 5.2: Lapwings data: combining DA and HMM structure. Diamonds represent the imputed nodes, squares – the data, circles – the unknown variables. Integrating out $\boldsymbol{N_1}$ leads to a second order HMM on $\boldsymbol{N_a}$. Dashed lines used for the relations *from* the imputed (known) states.

The two-dimensional state $[N_1, N_a]_{t=1}^T$ follows the first-order Markov process with a non-standard transition kernel. We can notice that 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 the observed estimate 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 $\boldsymbol{x_{int}} = \boldsymbol{N_1}$ and $\boldsymbol{x_{aug}} = \boldsymbol{N_a}$. The resulting modified dependence structure is presented in Figure 5.2. Marginalising over $\boldsymbol{N_1}$ allows us to simplify the analysis as we only need to consider $\boldsymbol{N_a}$ which now follows a second-order Markov process. A similar second order structure in this context has also been noted by Besbeas and Morgan (2018).

Hidden Markov Model approximation The resulting SCDL for the augmented data set $(\boldsymbol{y}^T, \boldsymbol{N_a}^T)^T$ is given by

$$p(\boldsymbol{y}, \boldsymbol{N_a}|\boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{N_a}, \boldsymbol{\theta}) p(\boldsymbol{N_a}|\boldsymbol{\theta}), \tag{5.7}$$

which is still intractable. Hence, we employ an HMM-based approximation to (5.7) discussed in Section 4. Because $N_{1,t}$ follows a Poisson distribution, we only need to specify a truncation value N^* for the maximum population size for first-years (i.e. we set $b_B = N^*$, with b_0 naturally being equal to 0). Since

the observations $\boldsymbol{y}$ are conditionally independent from $\boldsymbol{N_1}$ given $\boldsymbol{N_a}$, integrating out of $\boldsymbol{N_1}$ can be done only for the second term on the right hand side (5.7), to obtain the marginal pmf for $\boldsymbol{N_a}$. Below, to ease the notation, we omit $\boldsymbol{\theta}$ in the conditioning. The marginal pmf of $\boldsymbol{N_a}$ is given by

$$\begin{aligned} p(\boldsymbol{N_a}) &= p(N_{a,0}, N_{a,1}, \ldots, N_{a,T}) \\ &= \sum_{\boldsymbol{N_1}} p(N_{a,0}), p(N_{1,0}) p(N_{1,1}|N_{a,0})\, p(N_{a,1}|N_{a,0}, N_{1,0}) \ldots p(N_{1,T}|N_{a,T-1})\, p(N_{a,T}|N_{a,T-1}, N_{1,T-1}) \end{aligned} \tag{5.8}$$

and we can approximate the elements of this multiple sum as (for $t \geq 2$)

$$\begin{aligned} p(N_{a,t}|\boldsymbol{N}_{a,0:t-1}) &= \sum_{k=0}^{N^*} \mathbb{P}(N_{1,t-1} = k|\boldsymbol{N}_{a,0:t-1}) p(N_{a,t}\boldsymbol{N}_{a,0:t-1}, N_{1,t-1} = k), \\ &= \sum_{k=0}^{N^*} \underbrace{\mathbb{P}(N_{1,t-1} = k|N_{a,t-2})}_{=:u_{k,t-1}} \underbrace{p(N_{a,t}|N_{a,t-1}, N_{1,t-1} = k)}_{=:p_{k,t}}. \end{aligned} \tag{5.9}$$

In (5.9) $p_{k,t}$ denotes the conditional pmf of $N_{a,t}$ given $N_{1,t-1} = k$ and $N_{a,t-1}$ for which

$$p_{k,t} = p(N_{a,t}|N_{a,t-1}, N_{1,t-1} = k) \equiv \mathcal{B}\big((N_{a,t-1} + k), \phi_{a,t-1}\big).$$

Further, $u_{k,t}$ denotes the unconditional[8] probability of $N_{1,t} = k$. These unconditional probabilities of the hidden states can be derived as

$$\begin{aligned} u_{k,t} &= \mathbb{P}(N_{1,t} = k|N_{a,t-1}) \\ &= \sum_{l=0}^{N^*} \mathbb{P}(N_{1,t-1} = l|\boldsymbol{N}_{a,0:t-1}) \mathbb{P}(N_{1,t} = k|N_{1,t-1} = l, \boldsymbol{N}_{a,0:t-1}) \\ &= \sum_{l=0}^{N^*} \underbrace{\mathbb{P}(N_{1,t-1} = l|N_{a,t-2})}_{=u_{l,t-1}} \underbrace{\mathbb{P}(N_{1,t} = k|N_{a,t-1})}_{=:\gamma_{lk,t}}, \end{aligned}$$

which we collect in a vector $\boldsymbol{u}_t = \Big[u_{k,t}\Big]_{k=1}^{N^*}$. In general, the unconditional probabilities of an HMM are related to each other via the transition probabilities $\gamma_{lk,t}$ (i.e. conditional probabilities) as $\boldsymbol{u}_t = \boldsymbol{u}_{t-1}\Gamma_t$ with $\Gamma_t = \Big[\gamma_{lk,t}\Big]_{l,k=1}^{N^*}$. Here we have $\gamma_{lk,t} = \mathbb{P}(N_{1,t} = k|N_{1,t-1} = l, \boldsymbol{N}_{a,0:t-1})$, but since in the model $N_{1,t}$'s are mutually independent given $N_{a,t-1}$ we can simplify the transition probabilities to

$$\gamma_{lk,t} = \mathbb{P}(N_{1,t} = k|N_{a,t-1}) \equiv \mathcal{P}(N_{a,t}\rho_t\phi_{1,t})$$

for $k = 0, \ldots, N^* - 1$, while for $k = N^*$ we need $\gamma_{lk,t} = 1 - \sum_{j=0}^{N^*-1} \gamma_{lj,t}$ to ensure a valid probability distribution. This means that the time varying state transition matrix Γ_t takes a simple form

$$\Gamma_t = \Big[\boldsymbol{\gamma}_{1,t} \quad \ldots \quad \boldsymbol{\gamma}_{N^*-1,t} \quad \boldsymbol{\gamma}_{N^*,t}\Big],$$

i.e. with each column equal to $\boldsymbol{\gamma}_{k,t} = \gamma_{lk,t}\mathbf{1}$.

[8] In a sense of the Markov structure, but not in terms of the imputed $\boldsymbol{N_a}$ which we treat as known (see Zucchini et al., 2016, p.16, 32).

Finally, we can conveniently express (5.9) using matrix notation as

$$p(N_{a,t}|\boldsymbol{N}_{a,0:t-1}) = \underbrace{\begin{bmatrix} \gamma_{1,1,t-1} & \dots & \gamma_{1,T^*,t-1} \\ \vdots & \ddots & \vdots \\ \gamma_{1,1,t-1} & \dots & \gamma_{1,T^*,t-1} \end{bmatrix}}_{=\Gamma_{t-1}} \underbrace{\begin{bmatrix} p_{1,t} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & p_{N^*,t} \end{bmatrix}}_{=:P_t} \underbrace{\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}}_{\mathbf{1}} = \Gamma_{t-1} P_t \mathbf{1}.$$

Combining (5.8) and (5.9) yields the HMM form for the joint pmf of the imputed states

$$\begin{aligned} p(\boldsymbol{N_a}) &= \boldsymbol{u}_0 p(N_{a,0}) P_1 \Gamma_1 P_2 \dots \Gamma_{T-1} P_T \Gamma_T \mathbf{1} \\ &= \boldsymbol{u}_0 p(N_{a,0}) \left(\prod_{t=1}^{T} P_t \Gamma_t \right) \mathbf{1} \end{aligned}$$

where $\boldsymbol{u}_0 = \begin{bmatrix} p(N_{1,0}) = 0 & \dots & p(N_{1,0}) = N^* \end{bmatrix}^T$ is the initial state distribution.

As stated above, the real observations y_t, conditionally on $N_{a,t}$, are independent of $N_{1,t}$ so that the observation matrix becomes a scaled identity matrix

$$Q_t = p(y_t|N_{a,t})\mathbb{I} = \mathcal{N}(y_t|N_{a,t}, \sigma_y^2)\mathbb{I}.$$

Finally, the approximation to the SCDL (5.7) can be expressed as

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

State acceptance rate Let the current state of a Markov chain be $\boldsymbol{N_a^{(j)}} = \{N_{a,t}^{(j)}\}_{t=1}^T$ and consider updating of its t'th component. Let the proposed value be $N_{a,t}^{(\bullet)}$, with $\boldsymbol{N_a^{(\bullet)}} = \{N_{a,1}^{(j)}, \dots, N_{a,t-1}^{(j)}, N_{a,t}^{(\bullet)}, N_{a,t+1}^{(j)}, \dots, N_{a,T}^{(j)}\}$. The move is accepted with the probability $1 \wedge a(\boldsymbol{N_a^{(j)}}, \boldsymbol{N_a^{(\bullet)}})$, where $a(\boldsymbol{N_a^{(j)}}, \boldsymbol{N_a^{(\bullet)}})$ is the acceptance rate. Since we use a single update MH-RW with a symmetric (uniform) proposal distribution, the proposal terms required for the ratio cancel in the acceptance rate, which then can be further simplified as follows

$$\begin{aligned} a(\boldsymbol{N_a^{(j)}}, \boldsymbol{N_a^{(\bullet)}}) &= \frac{p(\boldsymbol{y}, \boldsymbol{N_a^{(\bullet)}}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\boldsymbol{y}, \boldsymbol{N_a^{(j)}}|\boldsymbol{\theta}) p(\boldsymbol{\theta})} = \frac{p(\boldsymbol{y}, \boldsymbol{N_a^{(\bullet)}}|\boldsymbol{\theta})}{p(\boldsymbol{y}, \boldsymbol{N_a^{(j)}}|\boldsymbol{\theta})} \\ &= \frac{p(y_t|N_t^{(\bullet)}) \mathbf{1}^T \Gamma_{t-1}^{(\bullet)} P_t^{(\bullet)} \Gamma_t^{(\bullet)} P_{t+1}^{(\bullet)} \Gamma_{t+1}^{(\bullet)} P_{t+2}^{(\bullet)} \mathbf{1}}{p(y_t|N_t^{(j)}) \mathbf{1}^T \Gamma_{t-1}^{(j)} P_t^{(j)} \Gamma_t^{(j)} P_{t+1}^{(j)} \Gamma_{t+1}^{(j)} P_{t+2}^{(j)} \mathbf{1}}, \end{aligned}$$

where the superscripts $^{(j)}$ and $^{(\bullet)}$ refer to values computed based on the current state of the Markov chain, $N_{a,t}^{(j)}$, and on the proposed value $N_{a,t}^{(\bullet)}$, respectively. Hence, due to the second-order structure we need five elements of the vector $\boldsymbol{N_a}$ when updating $N_{a,t}$ (i.e. $N_{a,t-2}$, $N_{a,t-1}$, $N_{a,t}$, $N_{a,t+1}$ and $N_{a,t+2}$), while all other terms cancel in the acceptance probability as they are conditionally independent.

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 $\boldsymbol{N_a}$. As already mentioned above, for comparability we use a "vanilla" MH RW algorithm for the estimation of the integrated model. In particular, we use a discrete uniform Metropolis RW algorithm to perform single-step updates of the states and normal Metropolis RW to sample the logistic regression coefficients. For the observation variance we use a Gibbs

Method	Absolute time	Relative time
DA	1203.67	1.00
Adapt10	978.27	0.81
Adapt20	1067.61	0.89
Adapt30	1024.87	0.85
Bin10	1022.32	0.85
Bin20	1060.41	0.88
Bin30	1135.83	0.94
Exact	2855.16	2.37

Table 1: Lapwings data: absolute (in seconds) and relative (wrt the full DA) computation times for $M = 100,000$ posterior draws after a burn-in of $10,000$.

update with the conditional distribution being of the form

$$\sigma_y^2 | \boldsymbol{N_a} \sim \Gamma{-}1 \left(a_y + \frac{T}{2}, b_y + \frac{1}{2} \sum_{t=1}^{T} (y_t - N_{a,t})^2 \right).$$

For the SCDA next to the "exact" integration, in which the only influence on the posterior is the upper limit of the admissible integration range $b_B = 679$. This choice of the upper bound is based on the results for first-years from previous studies and from preliminary runs of the full DA. 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 as mentioned in Section 4.1. Each time we draw $M = 100,000$ draws after a burn-in of $10,000$.

Table 1 summarises computation time for each of the schemes. As expected, the exact method is the slowest (2.5 times than the full DA approach) as each integration is based on summing of 680 elements. All the approximate schemes are faster (10–20%) than the DA approach thanks to their efficient implementation based on vectorised computations with relatively few elements to be summed every iteration. Tables 2 and 3 present the results for $\boldsymbol{\theta}$ and for selected elements of $\boldsymbol{N_a}$, respectively, and we report posterior means and standard deviations as well as ESSs and ESSs per second. Figure 5.3 illustrates the posterior means and 95% credible intervals (CI) for the adult population comparing the accuracy of the full DA with that of the SCDA methods (separately for the adaptive intervals and fixed bins). We can see that all the methods deliver virtually the same posterior means and comparable 95% symmetric CI, with only the fixed bin case with 10 bins deviating slightly from all other methods. Interestingly, 10 adaptive bins work fine in this case, indicating an increased accuracy of the adaptive approach.

Our 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 4 times) ESSs and ESSs per second. This can be also seen in Figures 5.4 and 5.5 which show the autocorrelation (ACF) plots for the SSM parameters (except for Gibbs-updated σ_y^2) and for the selected elements of $\boldsymbol{N_a}$, respectively. In most of the illustrated cases the ACF plots for all the SCDA variants are much flatter that these for the full DA approach.

Method		α_1	α_a	α_ρ	α_λ	β_1	β_a	β_ρ	β_λ	σ_y^2
DA	Mean	0.547	1.574	-1.189	-4.578	-0.164	-0.240	-0.348	-0.364	30180.443
	(Std)	(0.068)	(0.071)	(0.091)	(0.035)	(0.062)	(0.039)	(0.043)	(0.040)	(8890.540)
	ESS	685.018	124.003	111.731	1089.194	1050.268	389.651	105.958	8205.778	1245.440
[1203.67 s]	ESS/sec.	0.569	0.103	0.093	0.905	0.873	0.324	0.088	6.817	1.035
Adapt10	Mean	0.547	1.564	-1.180	-4.580	-0.163	-0.239	-0.350	-0.364	30355.448
	(Std)	(0.068)	(0.070)	(0.092)	(0.035)	(0.061)	(0.040)	(0.040)	(0.040)	(8928.277)
	ESS	1490.019	390.223	316.009	**3035.051**	2777.217	527.023	126.022	7491.584	1852.490
[978.27 s]	ESS/sec.	1.523	**0.399**	**0.323**	**3.102**	**2.839**	0.539	0.129	7.658	1.894
Adapt20	Mean	0.544	1.564	-1.173	-4.581	-0.162	-0.238	-0.342	-0.363	30002.520
	(Std)	(0.069)	(0.072)	(0.094)	(0.035)	(0.060)	(0.039)	(0.039)	(0.040)	(8759.372)
	ESS	1359.221	395.695	324.918	2720.786	2685.188	586.964	243.425	8212.358	2074.915
[1067.62 s]	ESS/sec.	1.273	0.371	0.304	2.548	2.515	**0.550**	0.228	7.692	1.944
Adapt30	Mean	0.542	1.561	-1.166	-4.581	-0.162	-0.241	-0.339	-0.363	30311.546
	(Std)	(0.069)	(0.071)	(0.092)	(0.036)	(0.061)	(0.039)	(0.040)	(0.040)	(8888.626)
	ESS	1438.464	322.169	243.433	2736.107	2471.447	563.625	195.736	7146.030	2129.241
[1024.87 s]	ESS/sec.	1.404	0.314	0.238	2.670	2.411	0.550	0.191	6.973	2.078
Bin10	Mean	0.512	1.441	-1.044	-4.599	-0.207	-0.205	-0.348	-0.353	29992.001
	(Std)	(0.070)	(0.055)	(0.063)	(0.034)	(0.050)	(0.039)	(0.022)	(0.040)	(8837.189)
	ESS	942.247	34.191	37.276	562.131	181.066	104.744	**282.356**	8770.878	1627.515
[1022.32 s]	ESS/sec.	0.922	0.033	0.036	0.550	0.177	0.102	**0.276**	**8.579**	1.592
Bin20	Mean	0.546	1.570	-1.179	-4.579	-0.170	-0.240	-0.343	-0.364	30156.695
	(Std)	(0.069)	(0.069)	(0.090)	(0.035)	(0.061)	(0.039)	(0.040)	(0.040)	(8802.537)
	ESS	1250.068	269.489	210.552	2328.072	2566.054	525.402	139.107	8582.549	2269.829
[1060.41 s]	ESS/sec.	1.179	0.254	0.199	2.195	2.420	0.495	0.131	8.094	2.141
Bin30	Mean	0.545	1.562	-1.170	-4.580	-0.162	-0.240	-0.342	-0.363	30012.541
	(Std)	(0.069)	(0.073)	(0.095)	(0.035)	(0.061)	(0.039)	(0.040)	(0.040)	(8698.313)
	ESS	**1758.336**	**438.518**	**329.308**	2901.919	**2873.035**	501.803	207.643	7613.013	2705.886
[1135.83 s]	ESS/sec.	**1.548**	0.386	0.290	2.555	2.529	0.442	0.183	6.703	**2.382**
Exact	Mean	0.545	1.564	-1.175	-4.580	-0.162	-0.240	-0.345	-0.363	30063.430
	(Std)	(0.068)	(0.069)	(0.090)	(0.035)	(0.060)	(0.039)	(0.042)	(0.040)	(8770.776)
	ESS	1631.604	433.106	361.747	3058.624	2658.813	**720.514**	191.434	**8859.552**	**2734.263**
[2855.16 s]	ESS/sec.	0.571	0.152	0.127	1.071	0.931	0.252	0.067	3.103	0.958

ESS: at lag equal to the lowest order at which sample autocorrelation is not significant.

Computing times (in seconds) in square brackets.

Table 2: Posterior means, standard deviations and effective sample sizes (ESS) of the model parameters for $M = 100,000$ posterior draws after a burn-in of $10,000$ for the lapwings data. The highest ESS and ESS/sec. for each parameter in bold.

Method		Na_4	Na_8	Na_{12}	Na_{16}	Na_{20}	Na_{24}	Na_{28}	Na_{32}	Na_{36}	Na_{min}	Na_{max}
DA	Mean	1083.511	1325.452	1674.379	1935.843	1614.735	1264.606	1174.201	964.850	776.152	1113.758	1083.511
	(Std)	(26.311)	(43.084)	(52.062)	(67.986)	(53.974)	(53.679)	(49.459)	(59.570)	(72.066)	(50.975)	(26.311)
	ESS	459.890	179.262	120.533	134.145	147.491	154.218	59.186	61.888	68.193	55.390	459.890
[1203.67 s]	ESS/sec.	0.382	0.149	0.100	0.111	0.123	0.128	0.049	0.051	0.057	0.046	0.382
Adapt10	Mean	1083.463	1326.451	1681.573	1947.344	1621.635	1268.140	1174.992	962.144	770.630	1113.567	1083.463
	(Std)	(26.252)	(43.505)	(51.511)	(68.960)	(53.635)	(56.490)	(49.993)	(56.982)	(66.910)	(51.062)	(26.252)
	ESS	879.048	**393.504**	317.925	335.006	294.857	234.322	42.972	46.077	51.279	41.776	879.048
[978.27 s]	ESS/sec.	**0.899**	**0.402**	0.325	0.342	0.301	0.240	0.044	0.047	0.052	0.043	**0.899**
Adapt20	Mean	1081.745	1320.731	1670.036	1934.081	1615.289	1268.925	1181.088	973.608	785.422	1121.175	1081.745
	(Std)	(27.448)	(44.994)	(52.415)	(67.959)	(51.419)	(53.695)	(46.369)	(51.319)	(61.671)	(46.408)	(27.448)
	ESS	792.495	300.795	262.727	413.404	298.921	310.999	173.526	150.006	**160.730**	167.407	792.495
[1067.62 s]	ESS/sec.	0.742	0.282	0.246	0.387	0.280	0.291	0.163	0.141	0.151	0.157	0.742
Adapt30	Mean	1081.738	1319.287	1670.942	1938.897	1617.392	1268.431	1184.041	978.162	791.521	1124.325	1081.738
	(Std)	(27.373)	(45.491)	(51.544)	(65.247)	(53.622)	(54.570)	(48.404)	(56.590)	(68.242)	(49.640)	(27.373)
	ESS	596.757	278.204	246.717	434.153	326.282	305.667	180.570	154.889	163.795	**170.926**	596.757
[1024.87 s]	ESS/sec.	0.582	0.271	0.241	0.424	0.318	0.298	**0.176**	0.151	**0.160**	**0.167**	0.582
Bin10	Mean	1075.915	1343.027	1699.239	1979.991	1671.882	1313.004	1194.677	954.958	733.686	1140.264	1075.915
	(Std)	(26.815)	(43.436)	(50.665)	(62.415)	(48.444)	(54.649)	(42.543)	(39.704)	(41.444)	(43.179)	(26.815)
	ESS	868.244	310.552	327.451	243.190	172.939	108.027	97.045	**162.952**	91.937	87.733	868.244
[1022.32 s]	ESS/sec.	0.849	0.304	0.320	0.238	0.169	0.106	0.095	**0.159**	0.090	0.086	0.849
Bin20	Mean	1079.800	1319.959	1671.116	1939.171	1619.882	1270.619	1183.223	976.198	788.919	1123.716	1079.800
	(Std)	(25.975)	(44.138)	(51.979)	(67.014)	(52.053)	(54.001)	(46.782)	(53.316)	(64.386)	(47.502)	(25.975)
	ESS	785.417	279.161	225.312	331.389	344.148	328.106	73.864	57.640	63.944	67.105	785.417
[1060.41 s]	ESS/sec.	0.741	0.263	0.212	0.313	0.325	**0.309**	0.070	0.054	0.060	0.063	0.741
Bin30	Mean	1079.485	1320.178	1671.219	1936.264	1615.844	1268.040	1181.895	975.230	787.549	1121.970	1079.485
	(Std)	(25.719)	(43.238)	(47.997)	(62.926)	(50.327)	(53.921)	(46.464)	(53.458)	(65.120)	(47.102)	(25.719)
	ESS	**911.428**	369.874	**373.588**	**504.546**	346.512	246.825	111.292	89.719	98.283	102.004	**911.428**
[1135.83 s]	ESS/sec.	0.802	0.326	**0.329**	**0.444**	**0.305**	0.217	0.098	0.079	0.087	0.090	0.802
Exact	Mean	1083.134	1324.134	1675.323	1939.629	1615.882	1265.869	1176.687	968.356	780.241	1116.332	1083.134
	(Std)	(27.191)	(44.555)	(52.061)	(67.385)	(53.858)	(54.547)	(46.090)	(54.115)	(66.988)	(47.247)	(27.191)
	ESS	902.234	349.462	293.269	365.968	**402.009**	**418.141**	**196.821**	121.888	116.693	167.980	902.234
[2855.16 s]	ESS/sec.	0.316	0.122	0.103	0.128	0.141	0.146	0.069	0.043	0.041	0.059	0.316

ESS: at lag equal to the lowest order at which sample autocorrelation is not significant.

Na_{min}/Na_{max}: corresponding to the lowest/highest ESS for the DA method.

Computing times (in seconds) in square brackets.

Table 3: Posterior means, standard deviations and effective sample sizes (ESS) of the model parameters for $M = 100,000$ posterior draws after a burn-in of $10,000$ for the lapwings data. The highest ESS and ESS/sec. for each parameter in bold.

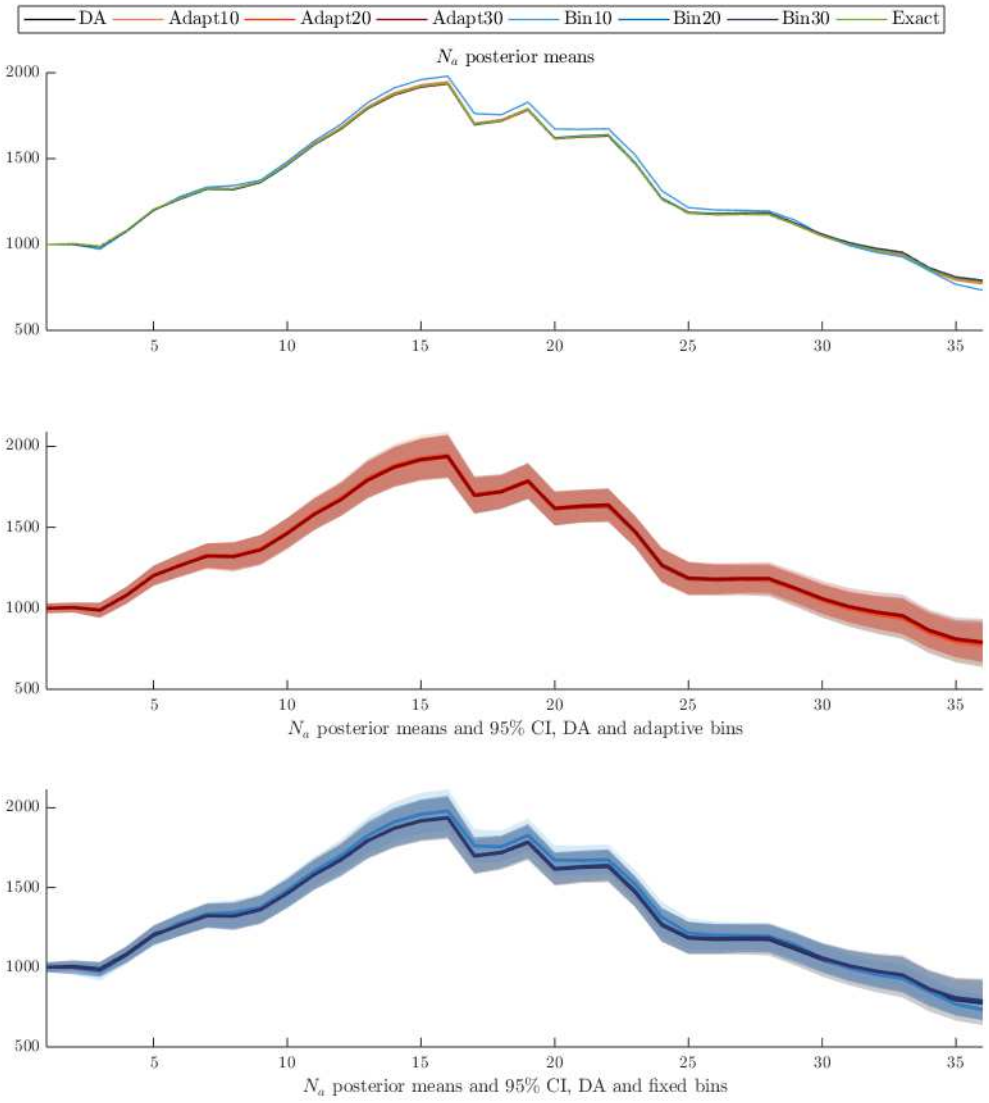


Figure 5.3: Lapwings data: the posterior means and 95% CI for the adult population.

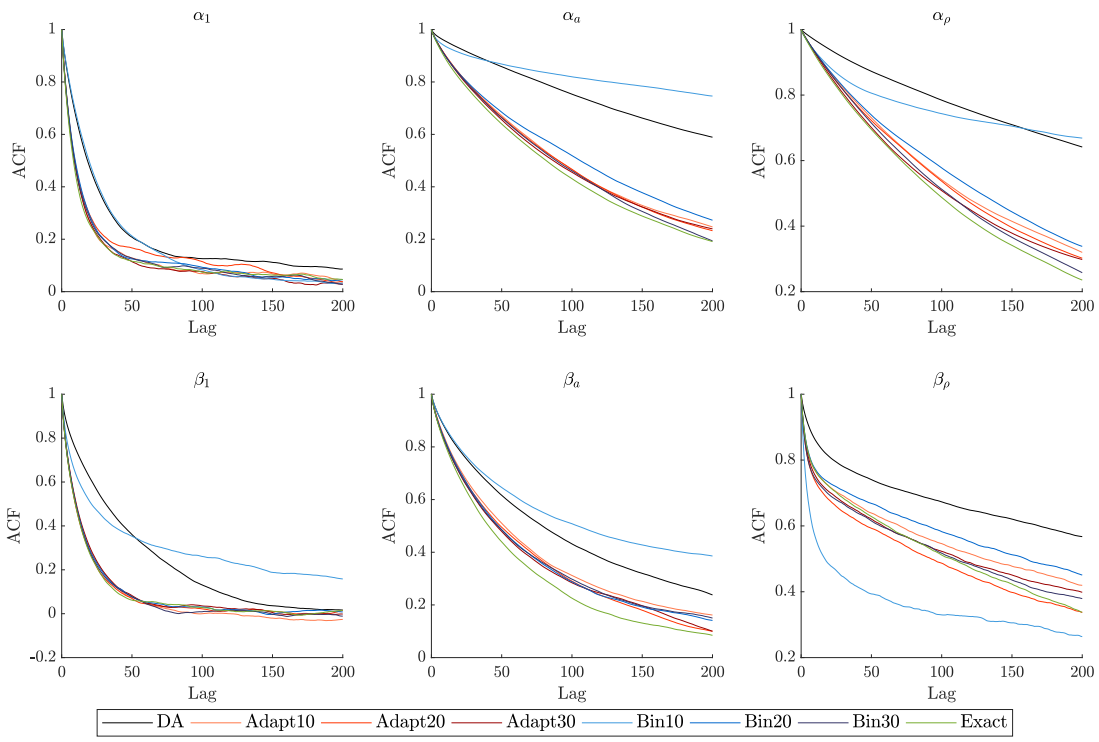


Figure 5.4: Lapwings data: ACF plots for the SSM parameters.

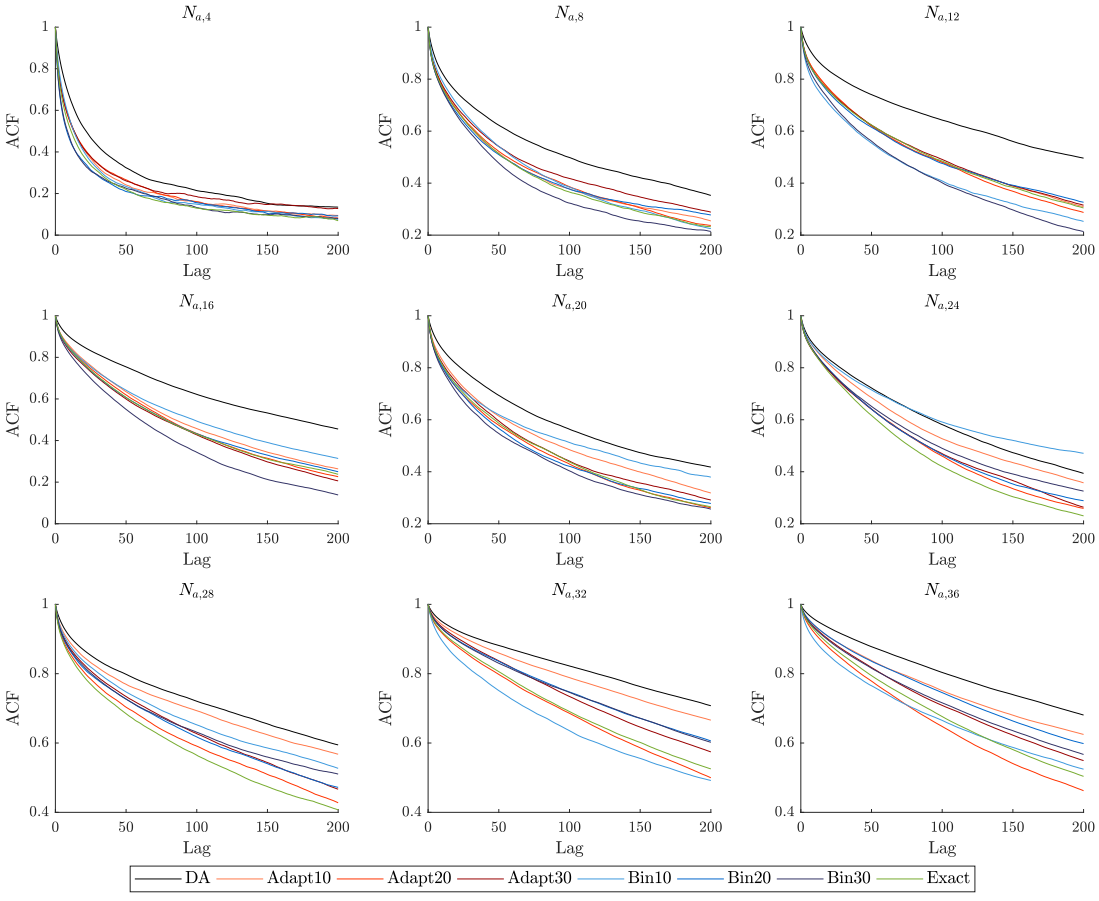


Figure 5.5: Lapwings data: ACF plots for the adult population.

5.2 Financial model: stochastic volatility

As our second illustration we consider the SV model in its basic form given by

$$y_t|h_t, \boldsymbol{\theta} \sim \mathcal{N}\left(0, \exp(h_t)\right), \tag{5.10}$$

$$h_{t+1}|h_t, , \boldsymbol{\theta} \sim \mathcal{N}\left(\mu + \phi(h_t - \mu), \sigma^2\right), \tag{5.11}$$

$$h_0 \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{1-\phi^2}\right), \tag{5.12}$$

for $t = 1, \ldots, T$. We adopt the prior specification of Kim et al. (1998)

$$\begin{aligned} \mu &\sim \mathcal{N}(0, \sigma^2_{\mu 0}), \\ \frac{\phi+1}{2} &\sim \mathcal{B}(\alpha_{\phi 0}, \beta_{\phi 0}), \\ \sigma^2 &\sim \mathcal{IG}(\alpha_{\sigma^2 0}, \beta_{\sigma^2 0}), \end{aligned}$$

with $\sigma^2_{\mu 0} = 10$, $\alpha_{\phi 0} = 20$, $\beta_{\phi 0} = 1.5$, $\alpha_{\sigma^2 0} = 5/2$, $\beta_{\sigma^2 0} = 0.05/2$. Estimation of the SV model has been considered as a challenging problem due to the intractable likelihood

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \int p(\boldsymbol{y}, \boldsymbol{h}) d\boldsymbol{h} = \int p(h_0) \prod_{t=1}^{T} p(y_t|h_t) p(h_t|h_{t-1}) dh_0 dh_1 \ldots dh_T. \tag{5.13}$$

Some of the previous approaches to tackle this issue include standard DA approach, in which the latent volatilities are imputed in an MCMC scheme, see Kim et al. (1998), Omori et al. (2007). Then, the augmented likelihood can be expressed in a closed form as

$$p(\boldsymbol{y}, \boldsymbol{h}|\boldsymbol{\theta}) = p(h_0) \prod_{t=1}^{T} p(y_t|h_t) p(h_t|h_{t-1}).$$

An alternative approach is provided by Fridman and Harris (1998) or Langrock et al. (2012b) who propose numerical integration of the latent states. In particular, Langrock et al. (2012b) approximate (5.13) using an HMM by discretising the state space of the SV model. They consider a form of numerical integration of the latent states based on a grid of B equally sized intervals (bins) $B_i = [b_{i-1}, b_i)$, $i = 1, \ldots, B$, with the corresponding representative points b_i^* (e.g. the midpoints). The range of the admissible values for the demeaned volatility, b_0 and b_B is set e.g. to $\pm 5\sigma_h$, with σ_h being the stationary (unconditional) standard deviation of the logvolatility process. This leads to an approximation of (5.13) via

$$p(\boldsymbol{y}|\boldsymbol{\theta}) \approx \boldsymbol{u}_0 \prod_{t=1}^{T} \Gamma Q_t \mathbf{1},$$

where $\Gamma = \left[\gamma_{i,j}\right]_{i,j=1,\ldots,B}$, with

$$\begin{aligned} \gamma_{i,j} &= \mathbb{P}(h_t - \mu \in B_j | h_{t-1} - \mu = b_i^*) \\ &= \Phi\left(\frac{b_j - \phi b_i^*}{\sigma}\right) - \Phi\left(\frac{b_{j-1} - \phi b_i^*}{\sigma}\right), \\ Q_t &= \operatorname{diag}\left(\varphi\left(\frac{y_t}{\exp((\mu + b_i^*)/2)}\right)\right)_{i=1,\ldots,B}, \end{aligned}$$

where Φ and φ denote the cdf and the pdf of the standard normal density, respectively. Notice that the transition probabilities are time-constant so that the underlying Markov chain is homogeneous. Sandmann

and Koopman (1998) point out that for the SV model such a form of numerical integration might not be always suitable since a fixed grid cannot efficiently capture different scales of volatility (periods of low and high volatility). We address this issue by suggesting an adaptive HMM-based approximation as an alternative to the fixed bins used by Langrock et al. (2012b).

Finally, we note that for μ and σ^2 Gibbs updates can be performed based on full conditional densities, see Kim et al. (1998). Furthermore, numerous enhancements for sampling of the hidden states has been devised, Kim et al. (1998), Omori et al. (2007) and Bos (2011) for an overview. However, our aim is to provide a general framework requiring only "vanilla" type updates (based on an RW–MH algorithm) and hence we consider the standard full DA as a comparison benchmark.

Dependence structure and SCDL The basic SV model specification concerns a single one-dimensional state on the real line, which is saliently different from the lapwings case. The sampling inefficiency in the current case originates from a high persistence of the logvolatility process. In order to break this dependence, we propose to impute $\boldsymbol{h_{2T}}$, the even states and to integrate out $\boldsymbol{h_{2T+1}}$, the odd ones. This corresponds to the *vertical* integration scheme with $\boldsymbol{x_{int}} = \boldsymbol{h_{2T+1}}$ and $\boldsymbol{x_{aug}} = \boldsymbol{h_{2T}}$. Without loss of generality we assume that T is odd so that h_T is integrated out; if T is even then we add one extra integration based on uniformly distributed h_{T+1}. We denote $T^* = \frac{T-1}{2}$ and skip $\boldsymbol{\theta}$ in conditioning for simplicity. The exact SCDL is given by

$$p(y, \boldsymbol{h_{2T}}) = p(h_0) \int p(h_1|h_0)p(y_1|h_0) \left(\prod_{t=1}^{T^*} p(y_{2t+1}|h_{2t+1})p(h_{2t+1}|h_{2t})p(y_{2t}|h_{2t})p(h_{2t}|h_{2t-1}) \right) dh_1 \dots dh_T, \tag{5.14}$$

and conditioning on the even states allows us to split (5.14) into a product $T^* + 1$ of integrals

$$\begin{aligned} p(y, \boldsymbol{h_{2T}}) &= p(h_0) \int p(h_1|h_0)p(y_1|h_0)dh_1 \prod_{t=1}^{T^*} \int p(y_{2t+1}|h_{2t+1})p(h_{2t+1}|h_{2t})p(y_{2t}|h_{2t})p(h_{2t}|h_{2t-1})dh_{2t+1} \\ &= \underbrace{p(h_0)}_{=:C_0} \underbrace{\int p(h_1|h_0)p(y_1|h_0)dh_1}_{=:D_0} \prod_{t=1}^{T^*} \underbrace{p(y_{2t}|h_{2t})}_{=:C_t} \underbrace{\int p(y_{2t+1}|h_{2t+1})p(h_{2t+1}|h_{2t})p(h_{2t}|h_{2t-1})dh_{2t+1}}_{=:D_t}. \end{aligned} \tag{5.15}$$

Since the integrals in (5.15) are conditionally independent, it can be expressed as

$$p(y, \boldsymbol{h_{2T}}) = C_0 D_0 \prod_{t=1}^{T^*} C_t D_t = \prod_{t=0}^{T^*} C_t D_t,$$

which block structure is helpful for visualising the MH update scheme as we present below.

Let us denote the current sequence of the imputed states $\boldsymbol{h_{2T}^{(j)}} = \{h_0^{(j)}, h_2^{(j)}, \dots, h_{2t+2}^{(j)}, \dots, h_T^{(j)}\}$ and suppose that a single RW MH step for of h_{2t+2} results in the proposed sequence $\boldsymbol{h_{2T}^{(\bullet)}}$ with the element $h_{2t+2}^{(j)}$ replaced by the candidate $h_{2t+2}^{(\bullet)}$. Since the proposal distribution is symmetric and thus the proposal terms cancel out, the state acceptance rate is given by

$$a(h_{2t+2}^{(\bullet)}, h_{2t+2}^{(j)}) = \frac{p(y, h_{2T}^{(\bullet)}|\boldsymbol{\theta})}{p(y, h_{2T}^{(j)}|\boldsymbol{\theta})} = \frac{C_t^{[2](\bullet)} D_t^{[1](\bullet)} D_{t+1}^{[1](\bullet)}}{C_t^{[2](j)} D_t^{[1](j)} D_{t+1}^{[1](j)}}, \tag{5.16}$$

where $^{[\cdot](\bullet)}$ and $^{[\cdot](j)}$ refer to the blocks evaluated on the proposed and the current variable, respectively (either the imputed state here or the parameter vector below).

For a single step RW MH update of $\boldsymbol{\theta}$, given $h_{2T}^{(j)}$ and y:

$$a(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}^{(\bullet)}) = \frac{p(y, \boldsymbol{h_{2T}^{(j)}}|\boldsymbol{\theta}^{(\bullet)})p(\boldsymbol{\theta}^{(\bullet)})}{p(y, \boldsymbol{h_{2T}^{(j)}}|\boldsymbol{\theta}^{(j)})p(\boldsymbol{\theta}^{(j)})} = \frac{p(\boldsymbol{\theta}^{(\bullet)})\prod_{t=0}^{T^*} D_t^{(\bullet)}}{p(\boldsymbol{\theta}^{(j)})\prod_{t=0}^{T^*} D_t(j)}. \tag{5.17}$$

Hidden Markov model approximation In practice the integrals D_t cannot be evaluated analytically and a form of numerical approximation needs to be adopted. We first propose to approximate each integral using a B-state HMM structure with fixed bins. This approach follows Langrock et al. (2012b) and consists in relating $z_t = k$, the Markov chain being in state k, to the event of $h_{2t+1} - \mu \in \mathcal{B}_k$, the demeaned volatility in an odd time period $2t+1$ falling into the kth bin B_k. Falling into bin B_k can be specified as e.g. lying in the interval $[b_{k-1}, b_k)$ or being equal to this interval's midpoint $b_k^* = \frac{b_{k-1}+b_k}{2}$. We take equally spaced bins, each of length λ. In particular, we consider approximation of the following form

$$D_t \approx \sum_{k=1}^{B} p(y_{2t+1}|h_{2t+1} - \mu = b_k^*)p(h_{2t+2}|h_{2t+1} - \mu = b_k^*)p(h_{2t+1} - \mu \in B_i|h_{2t}). \tag{5.18}$$

The last term in (5.18) can be approximated as

$$p(h_{2t+1} - \mu \in B_k|h_{2t}) \approx \Phi\left(\frac{b_k - \phi(h_{2t} - \mu)}{\sigma}\right) - \Phi\left(\frac{b_{k-1} - \phi(h_{2t} - \mu)}{\sigma}\right),$$

which is adopted in Langrock et al. (2012b), or using a simpler midpoint approximation

$$p(h_{2t+1} - \mu \in B_k|h_{2t}) \approx \lambda\varphi\left(\frac{b_k^* - \phi(h_{2t} - \mu)}{\sigma}\right),$$

which we adopt in our application due to computing time. Then the state acceptance rate (5.16) is approximated as

$$\begin{aligned} a(h_{2t+2}^{(\bullet)}, h_{2t+2}^{(j)}) \approx & \frac{\varphi\left(\frac{y_{2t+1}}{\exp(h_{2t+2}^{(\bullet)}/2)}\right)\sum_{k=1}^{B}\varphi\left(\frac{y_{2t+1}}{\exp((b_k^*+\mu)/2)}\right)\varphi\left(\frac{h_{2t+2}^{(*)}-\mu-\phi b_k^*}{\sigma}\right)\varphi\left(\frac{b_k^*-\phi(h_{2t}-\mu)}{\sigma}\right)}{\phi\left(\frac{y_{2t+1}}{\exp(h_{2t+2}^{(j)}/2)}\right)\sum_{k=1}^{B}\varphi\left(\frac{y_{2t+1}}{\exp((b_k^*+\mu)/2)}\right)\varphi\left(\frac{h_{2t+2}^{(j)}-\mu-\phi b_k^*}{\sigma}\right)\varphi\left(\frac{b_k^*-\phi(h_{2t}-\mu)}{\sigma}\right)} \\ & \times \frac{\sum_{k=1}^{B}\varphi\left(\frac{y_{2t+3}}{\exp((b_k^*+\mu)/2)}\right)\varphi\left(\frac{h_{2t+4}^{(j)}-\mu-\phi b_k^*}{\sigma}\right)\varphi\left(\frac{b_k^*-\phi(h_{2t+2}^{(*)}-\mu)}{\sigma}\right)}{\sum_{k=1}^{B}\varphi\left(\frac{y_{2t+3}}{\exp((b_k^*+\mu)/2)}\right)\varphi\left(\frac{h_{2t+4}^{(j)}-\mu-\phi b_k^*}{\sigma}\right)\varphi\left(\frac{b_k^*-\phi(h_{2t+2}^{(j)}-\mu)}{\sigma}\right)}, \end{aligned}$$

while for the parameter acceptance rate (5.17) we obtain

$$a(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}^{(\bullet)}) \approx \frac{p(\boldsymbol{\theta}^{(\bullet)})\prod_{t=0}^{T^*}\sum_{k=1}^{B}\varphi\left(\frac{y_{2t+1}}{\exp((b_k^*+\mu^{(\bullet)})/2)}\right)\varphi\left(\frac{b_k^*-\phi^{(\bullet)}(h_{2t}^{(j)}-\mu^{(\bullet)})}{\sigma^{(\bullet)}}\right)\varphi\left(\frac{h_{2t+2}^{(j)}-\mu^{(\bullet)}-\phi^{(\bullet)}b_k^*}{\sigma^{(\bullet)}}\right)}{p(\boldsymbol{\theta}^{(j)})\prod_{t=0}^{T^*}\sum_{k=1}^{B}\varphi\left(\frac{y_{2t+1}}{\exp((b_k^*+\mu^{(j)})/2)}\right)\varphi\left(\frac{b_k^*-\phi(h_{2t}^{(j)}-\mu^{(j)})}{\sigma^{(j)}}\right)\varphi\left(\frac{h_{2t+2}^{(j)}-\mu^{(j)}-\phi^{(j)}b_k^*}{\sigma^{(j)}}\right)}.$$

Adaptive HMM-based approximation An alternative approach to the approximation task is to use adaptive intervals. In particular, quantiles corresponding to intervals of equal probability can be used. Then, instead of specifying the grid points, we fix the probabilities for each bin, which previously needed to be determined. Thus, we face a quantile determination problem, as these are needed to obtain the midpoint values (used in conditioning). Consider a vector of quantiles $\boldsymbol{q} = [q_0, q_1, \ldots, q_B]$ together with their midpoints $\boldsymbol{q}^* = [q_1^*, q_1^*, \ldots, q_B^*]$ given by $q_k^* = \frac{q_{k-1}+q_k}{2}$. Then the bin midpoints at time $2t+1$

determined by the mid-quantiles are given by

$$\beta^*_{k,2t+1} = \phi(h_{2t} - \mu) + \sigma \cdot \Phi^{-1}\left(q^*_k\right), \qquad k = 1, \ldots, B,$$

where h_{2t} the imputed volatility for the previous time period. This means that

$$\gamma_{k,t} = p(h_{2t+1} - \mu \in \mathcal{B}_{k,2t+1} | h_{2t}, \boldsymbol{\theta}) = \frac{1}{B}$$

and we approximate D_t as

$$D_t \approx \sum_{k=1}^{B} \varphi\left(\frac{y_{2t+1}}{\exp((\beta^*_{k,2t+1} + \mu)/2)}\right) \varphi\left(\frac{h_{2t+2} - \mu - \phi\beta^*_{k,2t+1}}{\sigma}\right) \cdot \frac{1}{B},$$

where the constant transition probabilities from an imputed state cancel out in the acceptance ratios.

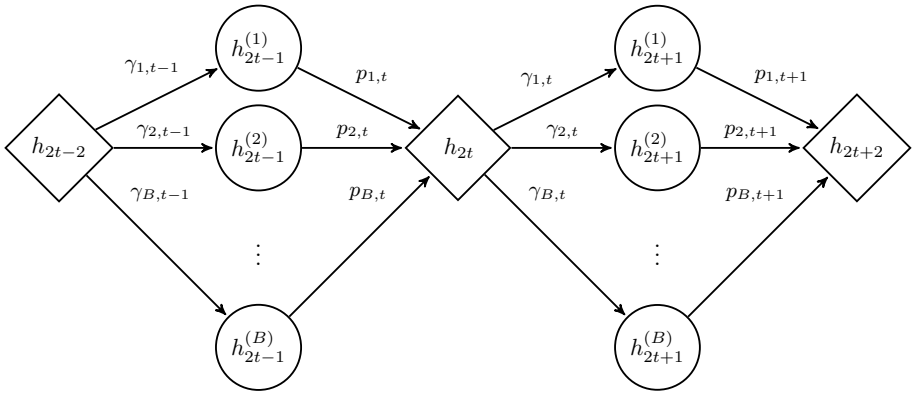


Figure 5.6: SV model: combining DA and the HMM-based integration. Diamonds represent the imputed states, circles – the states being integrated out. $h_t^{(k)}$ denotes $h_t \in \mathcal{B}_k$. The graph presents a single imputation problem of h_{2t} with the associated integrations.

Extensions of the basic SV model

SV in the mean The proposed SCDA scheme easily extends to more complex models, e.g. the popular Stochastic Volatility in the Mean (SVM) model of Koopman and Uspensky (2002) (see also Chan, 2017). Its basic specification is given by

$$y_t | h_t, \boldsymbol{\theta} \sim \mathcal{N}\left(\beta \exp(h_t), \exp(h_t)\right), \tag{5.19}$$

$$h_{t+1} | h_t \sim \mathcal{N}\left(\mu + \phi(h_t - \mu), \sigma^2\right), \tag{5.20}$$

$$h_0 \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{1 - \phi^2}\right), \tag{5.21}$$

for $t = 1, \ldots, T$. Hence, the latent volatility process h_t influences both the conditional variance and the conditional mean of the observation series y_t, which is additionally controlled by a scaling parameter β. For the volatility parameters μ, ϕ and σ^2 we adopt the prior specification as for the standard SV model, while for the mean-scaling parameter we specify $\beta \sim \mathcal{N}(0, \sigma^2_{\beta_0})$, with $\sigma^2_{\beta_0} = 10$.

SV with leverage The basic SV or SVM models can be extended to allow for *leverage* effects, i.e. a feedback from past logreturns to the current value of the volatility process. This effect is typically modelled as a negative correlation between the last period logreturns and the current value of volatility. The motivation behind the leverage effect is that the volatility in financial markets may adapt differently to positive and negative shocks/news (affecting logreturns), where large negative shocks are likely to increase the volatility. The SV model with leverage (SVL) has been frequently analysed in the literature, see Jungbacker and Koopman (2007), Meyer and Yu (2000), Yu (2005), Durbin and Koopman (2012, Section 9.5.5.) or Zucchini et al. (2016, Section 20.2.3). For convenience, we rewrite the basic SV model (5.10)–(5.12) as

$$\begin{aligned} y_t &= \exp(h_t/2)\varepsilon_t, & \varepsilon_t &\sim \mathcal{N}(0,1), \\ h_{t+1} &= \mu + \phi(h_t - \mu) + \eta_t, & \eta_t &\sim \mathcal{N}\left(0, \sigma^2\right), \\ h_1 &\sim \mathcal{N}\left(\mu, \frac{\sigma^2}{1-\phi^2}\right), \end{aligned}$$

for $t = 1, \ldots, T$. The only difference between the SVL model and the basic specification of the SV model is that now the error terms ε_t and η_t are assumed to be correlated: $\text{corr}[\varepsilon_t, \eta_t] = \rho \neq 0$, with ρ typically estimated to be negative[9]. This apparently slight modification has, however, substantial effect on the dependence structure in the model (see Figure 5.7) and hence the conditional distribution of h_t. To derive the latter several reformulations of the model has been proposed (Jungbacker and Koopman, 2007 or Meyer and Yu, 2000), however we will use the treatment provided by Zucchini et al. (2016, Section 20.2.3). These authors use the basic regression lemma for normal variables to show that

$$h_t | h_{t-1}, y_{t-1}, \mu, \phi, \sigma^2, \rho \sim \mathcal{N}\left(\mu + \phi(h_{t-1} - \mu) + \frac{\rho\sigma y_{t-1}}{\exp(h_{t-1}/2)}, \sigma^2(1-\rho^2)\right) \tag{5.22}$$

(Appendix C provides the details of the derivation). Formulation (5.22) is particularly convenient for "reusing" the derived integration scheme for the basic SV model, as we only need to adjust the transition probabilities in the approximation to C_t.

Modifications to the HMM-based approximation The proposed HMM-based approximation to SCDL can be easily adapted to allow for both extension by simply modifying the components of the matrices Γ_t, P_t and Q_t specified in (4.2)–(4.4). Notice that for the SVM model the dependence structure of the state is the same as for the basic SV model, hence the core of the integration/imputation scheme remains unchanged. What needs to be adjusted is the observation density, which can be done is a straightforward manner. The modification for the SVL model requires adjusting of the transition probabilities and the pdfs of the augmented states. Appendix (A.3) presents the required modifications for the largest model, allowing for both SV in the mean and for the leverage effect (which we refer to as the SVML model).

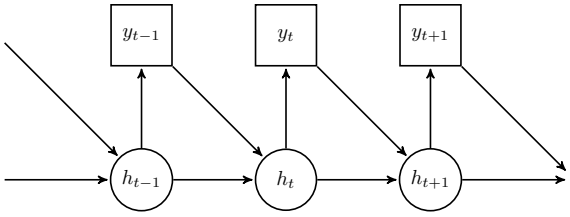


Figure 5.7: SV model with leverage: modified dependence structure due to feedback from the logreturns y_{t-1} to logvolatilities h_t.

[9]For this reason we also need to initialise the state vector one period later, i.e. at h_1. This can be easily understood from (5.22), where h_t is conditioned on y_{t-1}, among others.

Application To illustrate the SCDA approach based on vertical integration we consider daily log-returns of the IBM stock from 4th January 2000 to 29th December 2017 (4527 observations). The data are illustrated in Figure 5.8. We consider the basic SV model as well as its extended version, i.e. the SVML model. For both models we use adaptive intervals based on 10, 20 and 30 quantiles, while for the SV model we also consider fixed bins based on 20 and 30 intervals. The reason for the latter is that fixed bins turned out to be infeasible for the SVML model while for the SV model we needed to specify minimum 20 intervals to obtain stable results. For fixed bin we set the integration range to ± 4 (i.e. $b_0 = -4$ and $b_B = 4$). The obtained posterior means for the imputed volatilities suggest that this choice was sufficient, as the estimated mean of the state ranges roughly from -1 to 3 (Figure 5.9). For each model and method we simulate 50,000 draws after a burn-in of 10,000.

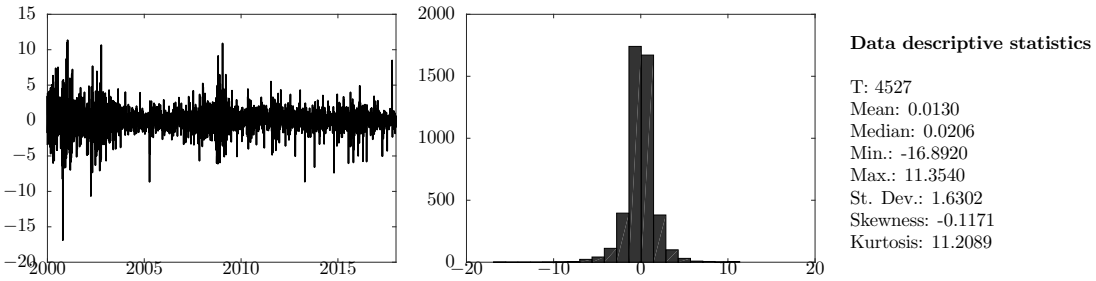


Figure 5.8: SV model: IBM series, 4527 observations from 4th January 2000 to 29th December 2017.

Tables 4 and 5 present the parameter estimation results for the SV model and SVML model, respectively. Tables 6 and 7 report the results for selected volatilities for SV and SVML, repetitively. We can see that for both models all the methods deliver comparable posterior means and standard deviations of parameters and standard deviations. A good agreement of the HMM-based schemes with the benchmark DA approach demonstrates that the developed methods provide a close approximation to the exact semi-complete data posterior. Interestingly, as few as 10 *adaptive* bins suffice to provide accurate estimates, which contrasts with *minimum* 50 fixed bins considered by Langrock et al. (2012b). Figure 5.9 illustrates that the estimates (posterior means) of the volatilities from the SV model obtained by the methods considered are very close to each other.

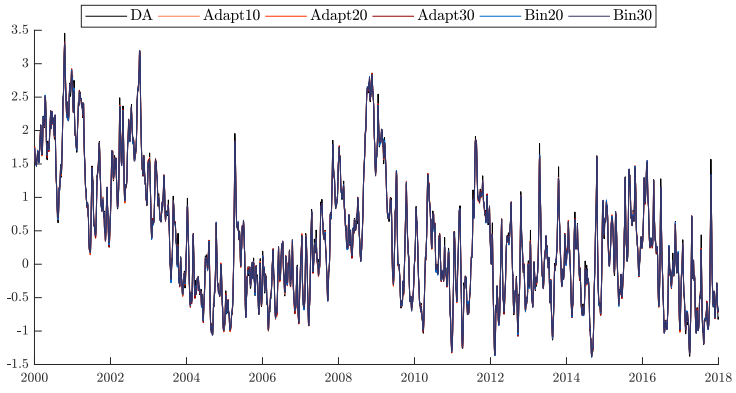


Figure 5.9: SV model: posterior means for the imputed volatilities. For illustration, for the SCDA methods the volatilities at odd time period are intrapolated between even time points.

Tables 4–7 further reveal that the proposed vertical integration scheme breaks the strong dependence between subsequent states to achieve the desired improvement in mixing. The ESS for model parameters obtained with the SCDA methods are typically higher than for the full DA approach. The only exception is the β parameter of the SVML for which all the methods exhibit excellent mixing with the DA approach slightly outperforming the HMM-based approximations. This high efficiency in the estimations of β is

related to the presence of this parameter only in the observation equation hence being less affected by the high autocorrelation of the state process. On the other hand, the second extra parameter of the SVML model, i.e. the leverage parameter ρ, is hard to estimate efficiently. For this parameter the SCDA turns out particularly useful in improving the mixing with the corresponding ESS values being up to 4.5 higher than for the benchmark DA. Figure 5.10 display the ACF plots for the parameters for the SV and SVML model, while Figure 5.11 for the selected volatilities. As suggested by the ESS reported in Tables 4–7, in the majority of the cases we observe much quicker decays in the autocorrelations for the SCDA algorithm compared to the "vanilla" DA approach.

Finally, we note that the computing times are higher for the SCDA approaches, with the computations for the adaptive case based on 10 bins taking roughly 17 times and 7 times longer than for full DA for the basic SV model and the SVML model, respectively. This suggests that the resulting gains in mixing may not necessarily be worth the extra computational cost. However, given the very simple structure of the basic SV model and not much more complex one of the SVML model, this is hardly surprising. We expect the SCDA approach to be more beneficial for more complex models, with even more involved dependence structure and relatively slower computation time for the benchmark DA approach. This can be already partly seen from shorter *relative* (to DA) computing times for the SCDA methods for the SVML compared to these for the SV model. For instance, the prosed integration scheme for the SV model could be particularly useful for a dynamic factor model with double stochastic volatility (where both the observation and the factor disturbances are subject to stochastic volatility). Due to the complex dependence structure as well matrix computations involved, the standard DA can be expected to perform relatively poorly and be time consuming to run. Then, there are several possibilities how to specify the augmentation-integration scheme, e.g. to fully integrate one of the SV processes; or interweave between every-second state of both SV processes (e.g. to integrate odd states for one SV process and even states for another SV process).

Method		μ	ϕ	σ^2
DA	Mean	0.376	0.962	0.081
	(Std)	(0.115)	(0.006)	(0.011)
[111.83 s]	ESS	3201.695	114.259	63.427
Adapt10	Mean	0.382	0.962	0.086
	(Std)	(0.116)	(0.006)	(0.013)
[1980.48 s]	ESS	5398.200	249.402	135.917
Adapt20	Mean	0.379	0.961	0.085
	(Std)	(0.115)	(0.006)	(0.013)
[2290.29 s]	ESS	5440.155	279.273	143.247
Adapt30	Mean	0.376	0.961	0.084
	(Std)	(0.115)	(0.006)	(0.012)
[2566.66 s]	ESS	5784.093	120.823	60.656
Bin20	Mean	0.381	0.962	0.082
	(Std)	(0.116)	(0.006)	(0.012)
[1683.02 s]	ESS	4504.012	239.698	147.409
Bin30	Mean	0.378	0.962	0.081
	(Std)	(0.115)	(0.006)	(0.012)
[2056.75 s]	ESS	5484.200	301.272	167.720

ESS: at lag equal to the lowest order at which sample autocorrelation is not significant.

Computing times (in seconds) in square brackets.

Table 4: SV model: posterior means, standard deviations and effective sample sizes (ESS) of the model parameters for $M = 50,000$ posterior draws after a burn-in of $10,000$.

Method		μ	ϕ	σ^2	β	ρ
DA	Mean	0.389	0.960	0.085	0.005	-0.286
	(Std)	(0.115)	(0.006)	(0.011)	(0.009)	(0.047)
[202.57 s]	ESS	2510.153	119.283	49.404	7375.378	147.507
Adapt10	Mean	0.377	0.963	0.081	0.006	-0.289
	(Std)	(0.112)	(0.006)	(0.011)	(0.009)	(0.044)
[1511.49 s]	ESS	5879.322	340.407	171.999	6969.391	681.943
Adapt20	Mean	0.375	0.961	0.084	0.006	-0.293
	(Std)	(0.115)	(0.006)	(0.012)	(0.009)	(0.045)
[2039.16 s]	ESS	4835.772	432.012	187.508	6692.492	513.298
Adapt30	Mean	0.373	0.960	0.084	0.006	-0.292
	(Std)	(0.113)	(0.006)	(0.013)	(0.009)	(0.048)
[2497.22 s]	ESS	5445.668	239.0242	149.680	7185.875	552.179

ESS: at lag equal to the lowest order at which sample autocorrelation is not significant.

Computing times (in seconds) in square brackets.

Table 5: SVML model: posterior means, standard deviations and effective sample sizes (ESS) of the model parameters for $M = 50,000$ posterior draws after a burn-in of $10,000$.

Method		h_{450}	h_{950}	h_{1450}	h_{1950}	h_{2450}	h_{2950}	h_{3450}	h_{3950}	h_{4450}
DA	Mean	0.974	0.259	-0.009	-0.006	0.257	1.083	0.011	0.733	-0.782
	(Std)	(0.444)	(0.417)	(0.404)	(0.492)	(0.466)	(0.449)	(0.485)	(0.457)	(0.454)
[111.83 s]	ESS	458.198	487.795	573.470	301.690	394.748	501.153	300.990	376.820	428.370
Adapt10	Mean	1.002	0.213	0.0104	-0.051	0.243	1.064	-0.022	0.753	-0.781
	(Std)	(0.434)	(0.431)	(0.432)	(0.492)	(0.469)	(0.445)	(0.479)	(0.453)	(0.482)
[1980.48 s]	ESS	1251.076	1489.877	1438.211	1245.904	1045.616	1509.202	1425.722	1338.145	1299.627
Adapt20	Mean	0.978	0.228	0.020	-0.040	0.241	1.059	-0.014	0.742	-0.805
	(Std)	(0.434)	(0.447)	(0.4358)	(0.504)	(0.467)	(0.442)	(0.489)	(0.448)	(0.467)
[2290.29 s]	ESS	1666.989	1223.563	1435.324	1330.146	1285.580	1458.985	1157.446	1288.562	1402.586
Adapt30	Mean	0.961	0.225	0.014	-0.052	0.240	1.101	0.014	0.745	-0.782
	(Std)	(0.439)	(0.438)	(0.431)	(0.499)	(0.457)	(0.442)	(0.482)	(0.436)	(0.474)
[2566.66 s]	ESS	1583.096	1638.283	1521.382	1270.313	1410.302	1607.990	1359.351	1457.938	1331.350
Bin20	Mean	0.979	0.207	0.018	-0.048	0.251	1.069	0.010	0.756	-0.808
	(Std)	(0.437)	(0.428)	(0.428)	(0.490)	(0.471)	(0.441)	(0.480)	(0.448)	(0.482)
[1683.02 s]	ESS	1419.504	1243.523	1632.920	1223.331	1179.196	1455.774	1395.415	1217.547	1171.629
Bin30	Mean	0.961	0.228	0.008	-0.048	0.238	1.075	-0.011	0.731	-0.819
	(Std)	(0.424)	(0.424)	(0.428)	(0.501)	(0.464)	(0.445)	(0.4866)	(0.4366)	(0.4644)
[2056.75 s]	ESS	1152.304	1759.833	1366.974	1150.885	1395.970	1422.304	1292.074	1222.696	1539.078

ESS: at lag equal to the lowest order at which sample autocorrelation is not significant.

Computing times (in seconds) in square brackets.

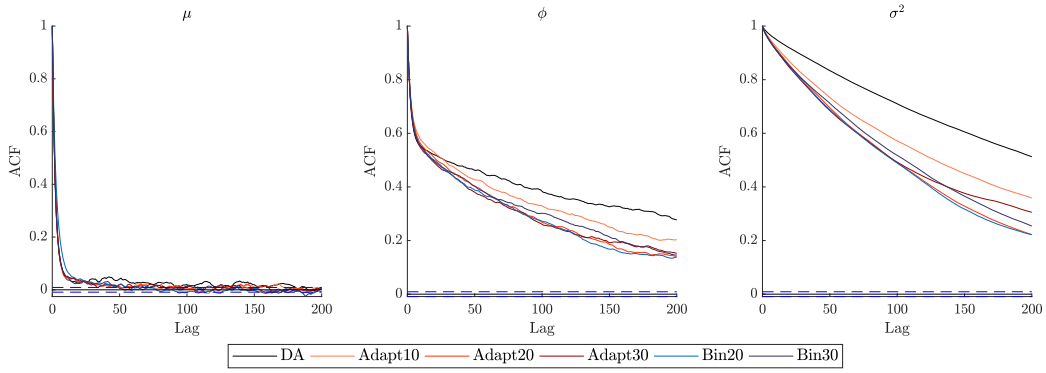
Table 6: SV model: posterior means, standard deviations and effective sample sizes (ESS) of the latent volatilities for $M = 50,000$ posterior draws after a burn-in of $10,000$.

Method		h_{450}	h_{950}	h_{1450}	h_{1950}	h_{2450}	h_{2950}	h_{3450}	h_{3950}	h_{4450}
DA	Mean	0.961	0.478	-0.163	0.011	0.448	1.026	-0.010	0.659	-0.963
	(Std)	(0.429)	(0.405)	(0.449)	(0.517)	(0.407)	(0.439)	(0.479)	(0.431)	(0.506)
[202.57 s]	ESS	650.696	420.645	540.354	276.558	621.079	633.797	506.926	348.977	346.512
Adapt10	Mean	0.955	0.454	-0.198	-0.022	0.413	1.024	-0.004	0.682	-0.984
	(Std)	(0.416)	(0.402)	(0.423)	(0.481)	(0.408)	(0.408)	(0.435)	(0.447)	(0.468)
[1511.49 s]	ESS	1550.904	1192.579	1183.693	1199.115	1548.657	1860.399	1277.721	1139.830	1089.324
Adapt20	Mean	0.920	0.470	-0.127	0.008	0.419	1.016	-0.055	0.677	-0.929
	(Std)	(0.400)	(0.405)	(0.450)	(0.489)	(0.419)	(0.421)	(0.463)	(0.435)	(0.493)
[2039.16 s]	ESS	1708.410	1280.039	1152.046	1008.572	1493.396	1621.910	1352.884	1231.148	970.933
Adapt30	Mean	0.915	0.409	-0.136	-0.034	0.402	1.030	-0.053	0.697	-0.948
	(Std)	(0.404)	(0.398)	(0.439)	(0.486)	(0.403)	(0.430)	(0.438)	(0.451)	(0.483)
[2497.22 s]	ESS	1860.153	1333.922	1846.106	1226.611	1486.479	1549.930	1353.398	1322.118	1205.346

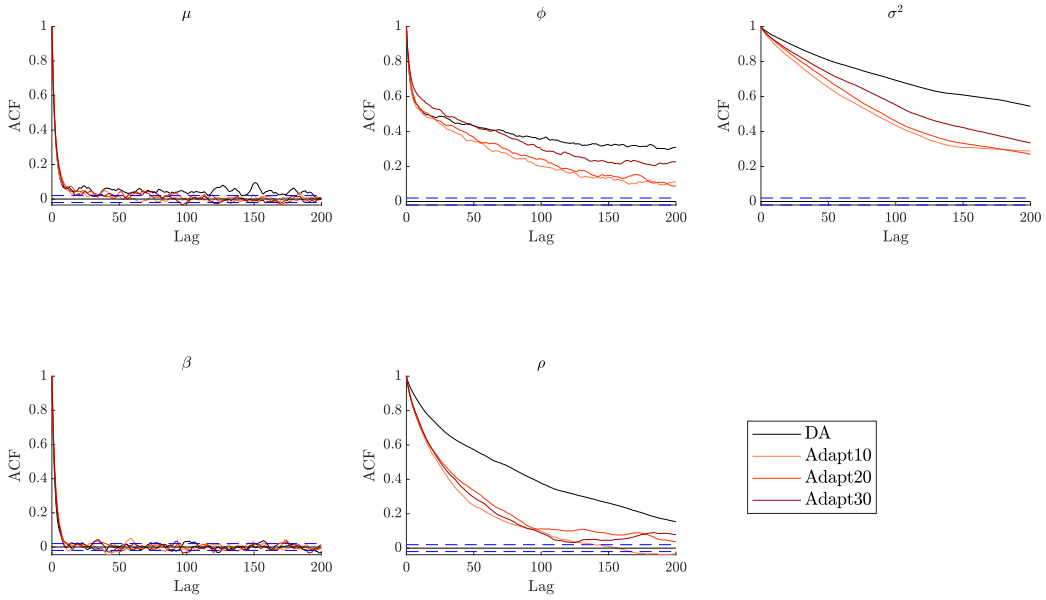
ESS: at lag equal to the lowest order at which sample autocorrelation is not significant.

Computing times (in seconds) in square brackets.

Table 7: SVML model: posterior means, standard deviations and effective sample sizes (ESS) of the latent volatilities for $M = 50,000$ posterior draws after a burn-in of $10,000$.

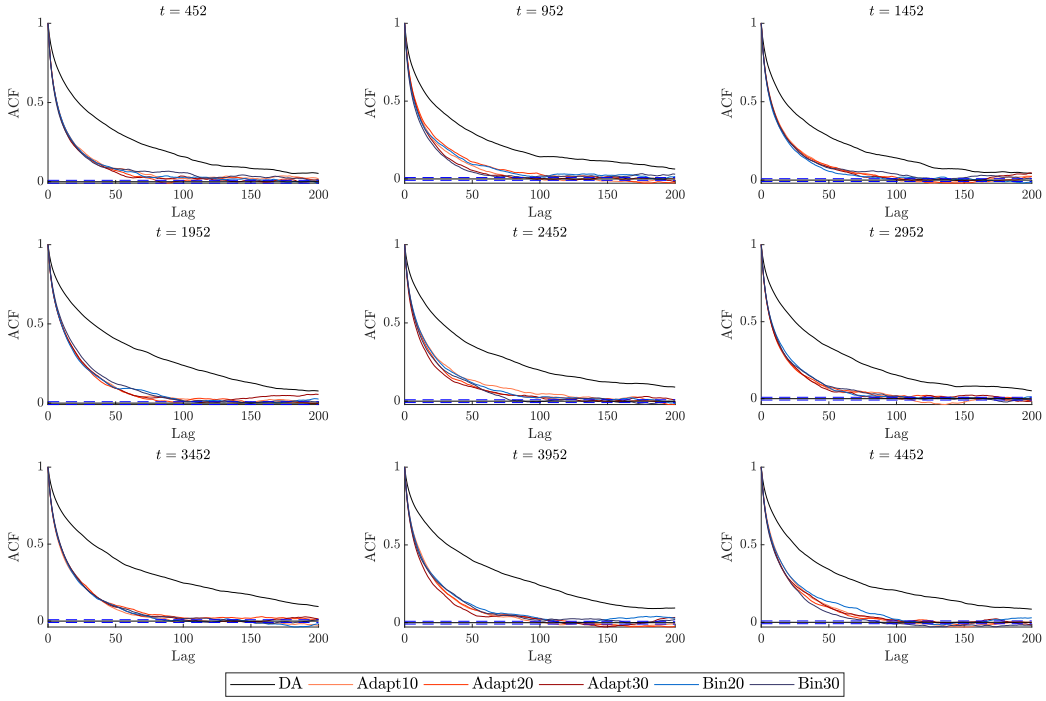


(a) SV model.

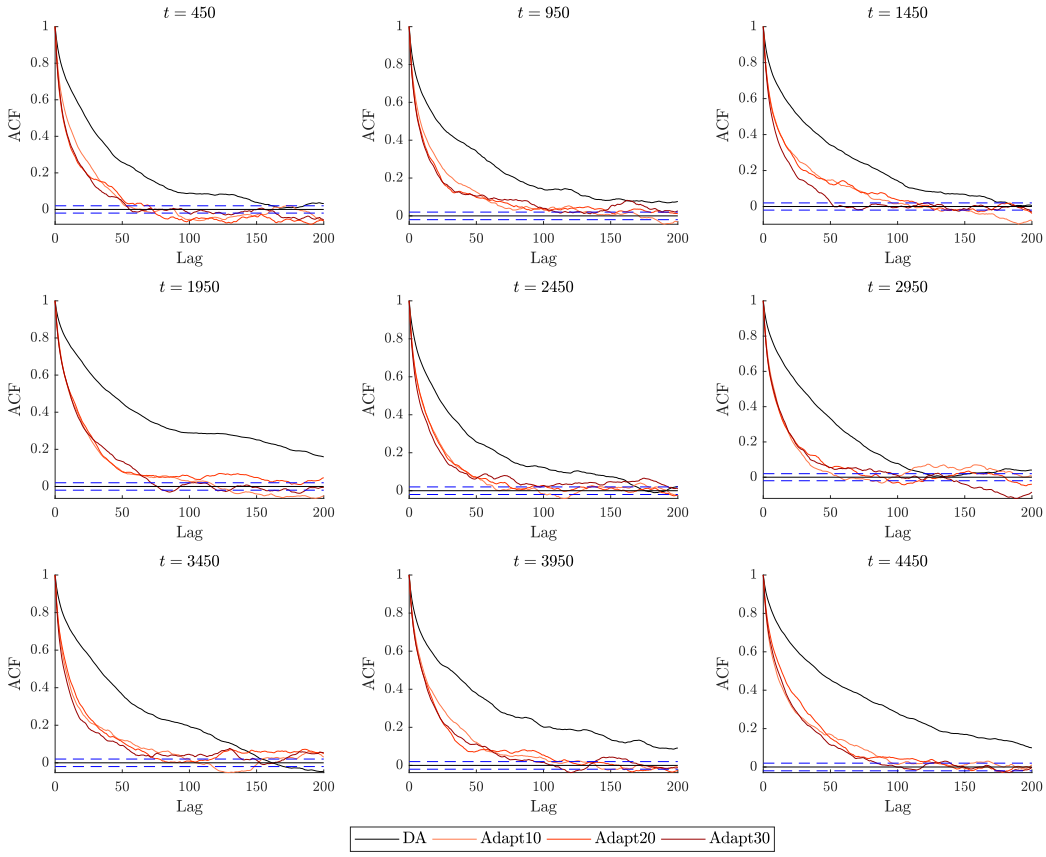


(b) SVML model

Figure 5.10: SV and SVML model: ACF plots for parameters.



(a) SV model.



(b) SVML model

Figure 5.11: SV and SVML model: ACF plots for selected volatilities.

6 Discussion

We have presented a new estimation method for state space models, called semi-complete data augmentation, designed to increase the efficiency of "vanilla" MCMC algorithms. The main idea behind the introduced approach is to combine Data Augmentation with numerical integration, where the latter aims at reducing the dependence between the imputed auxiliary variables. This concept relates to general Rao-Blackwellisation methods, however we do not require the resulting conditional distribution (given the imputed states) to be available in a closed-form (i.e. to be analytically integrable), nor the imputed auxiliary variables to be sufficient statistics for the marginalised variables.

We have proposed to base integration schemes 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", i.e. approximating similar values of the marginalised state with e.g. a single mid-value. 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 pre-specified grid and "adaptive bins" based on e.g. quantiles of the relevant distribution. The latter remove the problem of specifying the "essential domain" required for fixed bins, considered by e.g. Kitagawa (1987) and Langrock et al. (2012b). Adaptive bins are also more suited for problems with highly varying integration ranges such as in the SV model with leverage and SV in the mean (SVML), for which fixed bins are unlikely to be efficient (see Sandmann and Koopman, 1998). Moreover, a specific approximation accuracy typically can be achieved by using fewer adaptive bins than fixed bins, which – given similar computing times for both approaches based on the same number of bins – means the adaptive bins require less computing time to attain an appropriate precision.

The split of the latent states into "auxiliary" and "integrated" variables is model-dependent and is specified in such a way that the algorithm is efficient. 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 investigating of the underlying graphical structure of an SSM can be useful (cf. the concept of *d-separation* in Bayesian Networks). In general, high dimensional integration remains a challenging problem, which we leave for further research, noting that insights from the SMC samplers (Del Moral et al., 2006) could be useful in this context.

Bibliography

Abadi, F., O. Gimenez, B. Ullrich, R. Arlettaz, and M. Schaub (2010), "Estimation of Immigration Rate using Integrated Population Models." *Journal of Applied Ecology*, 393–400.

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

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

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

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

Besbeas, P. and B. J. T. Morgan (2018), "Exact Inference for Integrated Population Modelling." Technical report.

Bos, C. (2011), "Relating Stochastic Volatility Estimation Methods." Technical Report 11-049/4, Tinbergen Institute.

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

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

Casella, G. and C. P. Robert (1996), "Rao-Blackwellisation of Sampling Schemes." *Biometrika*, 83, 81–94.

Chan, J. C. C. (2017), "The Stochastic Volatility in Mean Model with Time-Varying Parameters: An Application to Inflation Modeling." *Journal of Business & Economic Statistics*, 35, 17–28.

Del Moral, P., A. Doucet, and A. Jasra (2006), "Sequential Monte Carlo Samplers." *Journal of the Royal Statistical Society: Series B*, 68, 411–436.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society. Series B*, 39, 1–38.

Douc, R. and C. P. Robert (2011), "A Vanilla Rao–Blackwellization of Metropolis–Hastings Algorithms." *The Annals of Statistics*, 39, 261–277.

Doucet, A., N. de Freitas, and N. Gordon, eds. (2001), *Sequential Monte Carlo Methods in Practice*. Springer.

Doucet, A., N. De Freitas, K. Murphy, and S. Russell (2000a), "Rao-Blackwellised Particle Filtering for Dynamic Bayesian Networks." In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, 176–183.

Doucet, A., S. Godsill, and C. Andrieu (2000b), "On Sequential Monte Carlo Sampling Methods for Bayesian Filtering." *Statistics and Computing*, 10, 197–208.

Durbin, J. and S. J. Koopman (2012), *Time Series Analysis by State Space Methods: Second Edition*. Oxford Statistical Science Series, OUP Oxford.

Fridman, M. and L. Harris (1998), "A Maximum Likelihood Approach for non-Gaussian Stochastic Volatility Models." *Journal of Business & Economic Statistics*, 87, 284—-291.

Frühwirth-Schnatter, S. (1994), "Data Augmentation and Dynamic Linear Models." *Journal of Time Series Analysis*, 15, 183–202.

Frühwirth-Schnatter, S. (2004), "Efficient Bayesian Parameter Estimation." In *State Space and Unobserved Component Models: Theory and Applications* (A. C. Harvey, S. J. Koopman, and N. Shephard, eds.), chapter 7, 123–151, Cambridge University Press.

Gelfand, A. and A. Smith (1990), "Sampling Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A. (2011), "The Bias-Variance Tradeoff." `https://andrewgelman.com/2011/10/15/the-bias-variance-tradeoff/`. [accessed 17 October 2018].

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

Geyer, C. J. (2011), "Introduction to Markov Chain Monte Carlo." In *Handbook of Markov Chain Monte Carlo* (S. Brooks, J. Galin A. Gelman, and X. L. Meng, eds.), chapter 4, 3–48, Chapman and Hall/CRC.

Ghysels, E., A. C. Harvey, and E. Renault (1996), "Stochastic volatility." In *Handbook of Statistics* (G. S. Maddala and C. R. Rao, eds.), volume 14, 119–191, Elsevier.

Goudie, R. J. B., A. M. Presanis, D. Lunn, D. De Angelis, and L. Wernisch (2018), "Joining and Splitting Models with Markov Melding." *Bayesian Analysis*.

Hobert, J. P. (2011), "The Data Augmentation Algorithm: Theory and Methodology." In *Handbook of Markov Chain Monte Carlo* (S. Brooks, J. Galin A. Gelman, and X. L. Meng, eds.), chapter 10, 253–294, CRC Press.

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

International Union for Conservation of Nature (2018), "The IUCN Red List of Threatened Species: lapwing." `http://www.iucnredlist.org/details/22693949/0`. Accessed: 2018-08-05.

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

Jungbacker, B. and S. J. Koopman (2007), "Monte Carlo Estimation for Nonlinear Non-Gaussian State Space Models." *Biometrika*, 94, 827–839.

Kass, R. E., B. P. Carlin, A. Gelman, and R. M. Neal (1998), "Markov Chain Monte Carlo in Practice: A Roundtable Discussion." *The American Statistician*, 52, 93–100.

Kim, S., N. Shephard, and S. Chib (1998), "Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models." *The Review of Economic Studies*, 65, 361–393.

King, R. (2011), "Statistical Ecology." In *Handbook of Markov Chain Monte Carlo* (S. Brooks, J. Galin A. Gelman, and X. L. Meng, eds.), chapter 17, 419–447, Chapman and Hall/CRC.

King, R., S. P. Brooks, C. Mazzetta, S. N. Freeman, and B. J. T. Morgan (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.

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

King, R., B. Morgan, O. Gimenez, and S. Brooks (2010), *Bayesian Analysis for Population Ecology*. Chapman and Hall/CRC.

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

Koopman, S. J. and E. Hol Uspensky (2002), "The Stochastic Volatility in Mean Model: Empirical Evidence from International Stock Markets." *Journal of Applied Econometrics*, 17, 667–689.

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

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

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

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

Marin, J. M. and C. Robert (2007), *Bayesian Core: a Practical Approach to Computational Bayesian Statistics*. Springer Science & Business Media.

Meyer, R. and J. Yu (2000), "BUGS for a Bayesian Analysis of Stochastic Volatility Models." *Econometrics Journal*, 3, 198–215.

Murphy, K. P. (2002), *Dynamic Bayesian Networks: Representation, Inference and Learning*. Ph.D. thesis, University of California, Berkeley.

Omori, Y., S. Chib, N. Shephard, and J. Nakajima (2007), "Stochastic Volatility with Leverage: Fast and Efficient Likelihood Inference." *Journal of Econometrics*, 140, 425–449.

Pitt, M. K., R. S. Silva, P. Giordani, and R. Kohn (2012), "On Some Properties of Markov Chain Monte Carlo Simulation Methods Based on the Particle Filter." *Journal of Econometrics*, 171, 134–151.

Robert, C. P. (2016), "Exact, unbiased, what else?!" `https://xianblog.wordpress.com/2016/04/13/exact-unbiased-what-else/`. [accessed 17 October 2018].

Robert, C. P. and G. Casella (2004), *Monte Carlo Statistical Methods: Second Edition*. Springer Texts in Statistics.

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

Rosenthal, J. S. (2011), "Optimal Proposal Distributions and Adaptive MCMC." In *Handbook of Markov Chain Monte Carlo* (S. Brooks, J. Galin A. Gelman, and X. L. Meng, eds.), chapter 4, 93–111, Chapman and Hall/CRC.

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

Shephard, N. (1996), "Statistical Aspects of ARCH and Stochastic Volatility." In *Time Series Models in Econometrics, Finance and Other Fields* (D. R. Cox, D. V. Hinkley, and O. E. Barndorff-Neilsen, eds.), 1–67, Chapman & Hall.

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

Taylor, S. J. (1994), "Modeling Stochastic Volatility: A Review and Comparative Study." *Mathematical Finance*, 4, 183–204.

The Royal Society for the Protection of Birds (2018), "The Red List of Conservation Concern: lapwing." `https://www.rspb.org.uk/birds-and-wildlife/wildlife-guides/bird-a-z/lapwing/`. Accessed: 2018-08-05.

West, M. and J. Harrison (1997), *Bayesian Forecasting and Dynamic Models: Second Edition*. Springer-Verlag.

Yu, J. (2005), "On Leverage in a Stochastic Volatility Models." *Journal of Econometrics*, 127, 165–178.

Zucchini, W., I. L. MacDonald, and R. Langrock (2016), *Hidden Markov Models for Time Series: An Introduction Using R, Second Edition*. Monographs on Statistics and Applied Probability 150, CRC Press.

A Specification details of the HMM approximations

In this section we present how the general formulation of the HMM-based approximation to the SCDL can be applied for the examples discussed in Sections 4 and 5.

A.1 Motivating example from Section 4.2

The SSM from Figure 4.1 is given by

$$\begin{aligned} y_t|x_{1,t}, x_{2,t} &\sim p(x_{1,t}, x_{2,t}), \\ x_{1,t+1}|x_{1,t}, x_{2,t} &\sim p(x_{1,t}, x_{2,t}), \\ x_{2,t+1}|x_{1,t}, x_{2,t} &\sim p(x_{1,t}, x_{2,t}), \\ x_{i,0} &\sim p(x_{i,0}), i = 1, 2 \end{aligned}$$

and we aim at imputing $x_{1,t}$ and integrating out $x_{2,t}$. Since in this model $T_{int} = T_{aug} = \{0, 1, \dots, T\}$, so that the index functions $\tau(t)$, $a(t)$ and $o(t)$ are simply identities, we skip them below to simplify the exposition. The marginal distribution[10] of the imputed state $x_{1,t}$ can be approximated as

$$\begin{aligned} p(x_{1,t}|\boldsymbol{x}_{1,0:t-1}) &\approx \sum_{j=1}^{B} \mathbb{P}(x_{2,t-1} \in \mathcal{B}_j|\boldsymbol{x}_{1,0:t-1}) p(x_{1,t}|\boldsymbol{x}_{1,0:t-1}, x_{2,t-1} \in \mathcal{B}_j), \\ &= \sum_{j=1}^{B} \underbrace{\mathbb{P}(x_{2,t-1} \in \mathcal{B}_j|\boldsymbol{x}_{1,0:t-2})}_{=:u_{j,t-1}} \underbrace{p(x_{j,t}|x_{1,t-1}, x_{2,t-1} \in \mathcal{B}_j)}_{=:p_{j,t}}, \end{aligned} \tag{A.1}$$

where $p_{j,t}$ is the likelihood of the augmented state at t given the imputed state at $t-1$ was in the jth bin (and previous realisations of x_{aug} but these are treated as known) and $u_{j,t-1}$ is the unconditional probability of the hidden process $x_{2,t}$ falling into the jth bin at $t-1$. This unconditional probability can be expressed as

$$u_{k,t} = \mathbb{P}(x_{2,t} \in \mathcal{B}_k|\boldsymbol{x}_{1,0:t-1}) = \sum_{l=j}^{B} \underbrace{\mathbb{P}(x_{2,t-1} \in \mathcal{B}_j|\boldsymbol{x}_{1,0:t-2})}_{=u_{j,t-1}} \underbrace{\mathbb{P}(x_{2,t} \in \mathcal{B}_k|\boldsymbol{x}_{1,0:t-1}, x_{2,t-1} \in \mathcal{B}_j)}_{=:\gamma_{jk,t}}, \tag{A.2}$$

which leads us to the standard result in HMM that the unconditional distributions in subsequent periods are related via the transition matrix $\Gamma_t = [\gamma_{jk,t}]_{k,j=1,\dots,T}$ as follows (cf. Zucchini et al., 2016, p.16, 32)

$$\boldsymbol{u}_t = \boldsymbol{u}_{t-1}\Gamma_t.$$

Next, the observations are conditionally independent, hence we have

$$p(y_t|\boldsymbol{x}_{1,0:t}) \approx \sum_{k=1}^{B} u_{k,t} \underbrace{p(y_t|x_{1,t}, x_{2,t} \in \mathcal{B}_k)}_{=:q_{k,t}}, \tag{A.3}$$

with $q_{k,t}$ denoting the likelihood of the observation at t given the hidden state in the same period t falling into bin k.

Comparing (A.2) and (A.3) shows that the distributions of the same period t augmented states and "real" observations are conditioned on the latent states from different periods, i.e. $t-1$ and t, respectively. This is a consequence of the general dependence structure in SSMs. The transition matrix at t captures this

[10] Marginal in the sense of the Markov structure, not the augmented states which we treat as known.

change of the underlying state so that combining of all there parts (A.1), (A.2) and (A.3) results in

$$p(y_t, x_{1,t}|\boldsymbol{x}_{1,0:t-1}) \approx \sum_{j=1}^{B}\sum_{k=1}^{B} u_{j,t-1}p_{j,t}\gamma_{jk,t}q_{k,t}.$$

To compute the HMM-based approximation to the SCDL we consider *forward probabilities* $\alpha_{\mathbf{t}}$ of the imputed states $x_{1,t}$ and observations y_t (cf. Zucchini et al., 2016, Sec. 2.3.2) defined as

$$\begin{aligned}\boldsymbol{\alpha}_t &= p(x_{1,0})\boldsymbol{u}_0\prod_{s=1}^{t}P_s\Gamma_s Q_s, \qquad t = 1, 2, \dots T,\\ \boldsymbol{\alpha}_0 &= p(x_{1,0})\boldsymbol{u}_0 Q_0,\end{aligned}$$

with $\boldsymbol{u}_0 = \begin{bmatrix}\mathbb{P}(x_{2,0} \in \mathcal{B}_1) & \dots & \mathbb{P}(x_{2,0} \in \mathcal{B}_B)\end{bmatrix}$ being the initial distribution of the latent state and $Q_0 = \mathbb{I}$. It follows from this definition that the forward probabilities can be expressed recursively as

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}P_t\Gamma_t Q_t$$

so that the required approximation to the SCDL being given by

$$\hat{p}_B(\boldsymbol{y}, \boldsymbol{x_1}) = p(x_{1,0})\boldsymbol{u}_0\boldsymbol{\alpha}_t\mathbb{I}.$$

Notice that the transition matrix Γ_t is a full matrix, however in some cases, e.g. the lapwing population model, the transition matrix can take a simpler form e.g. it is "column-wise constant": $\gamma_{lk,t} = \mathbb{P}(x_{2,t} = k|\boldsymbol{x}_{1,0:t-1}),\ \forall l$ (each row is the same). On the other hand, the augmented observation matrix and the "real" observation matrix have a diagonal forms $P_t = \text{diag}\left(p_{j,t}\right)_{j=1,\dots,B}$ and $Q_t = \text{diag}\left(q_{k,y}\right)_{j=k,\dots,B}$, respectively. Using the notation introduced in Section 4.2 we can write

$$\hat{p}_B(\boldsymbol{y}, \boldsymbol{x_1}) = p(x_{1,0})\boldsymbol{u}_0 Q_0\prod_{t=1}^{T^*}\left(P_t\Gamma_t Q_t\right)\mathbf{1}.$$

We can verify the above results be explicitly calculating

$$\begin{aligned}\boldsymbol{\alpha}_{t-1}P_t\Gamma_t Q_t\mathbf{1} &= \begin{bmatrix}\alpha_{1,t-1} & \dots & \alpha_{B,t-1}\end{bmatrix}\begin{bmatrix}p_{1,t} & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & p_{B,t}\end{bmatrix}\begin{bmatrix}\gamma_{11,t} & \dots & \gamma_{1B,t}\\ \vdots & \ddots & \vdots\\ \gamma_{B1,t} & \dots & \gamma_{BB,t}\end{bmatrix}\begin{bmatrix}q_{11,t} & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & q_{BB,t}\end{bmatrix}\begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix}\\ &= \begin{bmatrix}\alpha_{1,t-1} & \dots & \alpha_{B,t-1}\end{bmatrix}\begin{bmatrix}p_1\gamma_{11,t} & \dots & p_1\gamma_{1B,t}\\ \vdots & \ddots & \vdots\\ p_B\gamma_{B1,t} & \dots & p_B\gamma_{BB,t}\end{bmatrix}\begin{bmatrix}q_{1,t} & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & q_{B,t}\end{bmatrix}\begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix}\\ &= \begin{bmatrix}\alpha_{1,t-1} & \dots & \alpha_{B,t-1}\end{bmatrix}\begin{bmatrix}p_1\gamma_{11,t}q_{1,t} & \dots & p_1\gamma_{1B,t}q_{B,t}\\ \vdots & \ddots & \vdots\\ p_B\gamma_{B1,t}q_{1,t} & \dots & p_B\gamma_{BB,t}q_{B,t}\end{bmatrix}\begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix}\\ &= \begin{bmatrix}\underbrace{\sum_{j=1}^{B}\alpha_{l,t-1}p_j\gamma_{j1,t}q_{11,t}}_{=\alpha_{1,t}} & \dots & \underbrace{\sum_{j=1}^{B}\alpha_{l,t-1}p_j\gamma_{jB,t}q_{B,t}}_{=\alpha_{B,t}}\end{bmatrix}\begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix} = \boldsymbol{\alpha}_t\mathbf{1}\end{aligned}$$

and expressing

$$\hat{p}_B(\boldsymbol{y}, \boldsymbol{x_1}) = \sum_{k_1}^{B} \cdots \sum_{k_T}^{B} p(x_{1,0}) \boldsymbol{u}_0 \prod_{t=1}^{T} p_{k_{t-1},t} \gamma_{k_{t-1}k_t,t} q_{k_t,t}.$$

A.2 Lapwing population model

The approximation for the lapwings model is a special case of scheme used for the general model discussed in the Section A.1, with the transition matrix Γ_t having equal rows. The hidden Markov chain is here given as $\{z_t\} = \{N_{1,t}\}$ for $t = 0, \ldots, T$ and we again set $T_{int} = T_{aug} = \{0, 1, \ldots, T\}$, so that the index functions $\tau(t)$, $a(t)$ and $o(t)$ are simply identities, The transition matrix has the form

$$\begin{aligned}
\Gamma_t &= \begin{bmatrix} \mathbb{P}(N_{1,t} = b_1^* | N_{1,t-1} = b_1^*, \boldsymbol{N}_{a,0:t-1}) & \ldots & \mathbb{P}(N_{1,t} = b_B^* | N_{1,t-1} = b_1^*, \boldsymbol{N}_{a,0:t-1}) \\ \vdots & \ddots & \vdots \\ \mathbb{P}(N_{1,t} = b_1^* | N_{1,t-1} = b_B^*, \boldsymbol{N}_{a,0:t-1}) & \ldots & \mathbb{P}(N_{1,t} = b_B^* | N_{1,t-1} = b_B^*, \boldsymbol{N}_{a,0:t-1}) \end{bmatrix} \\
&= \begin{bmatrix} \mathbb{P}(N_{1,t} = b_1^* | N_{a,t-1}) & \ldots & \mathbb{P}(N_{1,t} = b_B^* | N_{a,t-1}) \\ \vdots & \ddots & \vdots \\ \mathbb{P}(N_{1,t} = b_1^* | N_{a,t-1}) & \ldots & \mathbb{P}(N_{1,t} = b_B^* | N_{a,t-1}) \end{bmatrix},
\end{aligned}$$

with $b_k^* = k$, for $k = 0, \ldots, N^*$. We can see that for each column of Γ_t its elements are the same. For the augmented observation matrix P_t we have

$$P_t = \text{diag} \left(p(N_{a,t} | \boldsymbol{N}_{a,0:t-1}, N_{1,t-1} = b_j^*) \right)_{j=1,\ldots,B},$$

so P_t and Γ_t condition on the same hidden state. The observation matrix has a simple form

$$Q_t = p(y_t | N_{a,t}) \mathbb{I}.$$

Inserting Q_t, P_t and Q_t in (4.6) leads to

$$\hat{p}_B(y, \boldsymbol{N_a}) = p(N_{a,0}) \mathbf{u}_0 Q_0 \prod_{t=1}^{T^*} P_t \Gamma_t Q_t \mathbf{1}, \tag{A.4}$$

where $\mathbf{u}_0 = \begin{bmatrix} \mathbb{P}(N_{1,0} \in \mathcal{B}_k) & \ldots & \mathbb{P}(N_{1,0} \in \mathcal{B}_B) \end{bmatrix}$ and $Q_0 = \mathbb{I}$. Then (A.4) is an HMM-based approximation to (5.7) converging to its true value in $B \to \infty$ and $b_B \to \infty$.

A.3 SV model

Basic SV model The SCDL for the basic SV model can be expressed as

$$p(y, \boldsymbol{h_T} | \boldsymbol{\theta}) = p(h_0) \int p(h_1 | h_0) p(y_1 | h_0) \left(\prod_{t=1}^{T^*} p(h_{2t+1} | h_{2t}) p(y_{2t+1} | h_{2t+1}) p(h_{2t} | h_{2t-1}) p(y_{2t} | h_{2t}) \right) dh_1 \ldots dh_{T^*}, \tag{A.5}$$

where $T^* = \frac{T-1}{2}$ (we assume T being odd). Since we impute volatilities at even time periods the Markov chain is given by $\{z_t\} = \{h_{2t+1}\}$ for $t = 1, \dots, T^*$ and its transition matrix has the form

$$\begin{aligned}\Gamma_t &= \begin{bmatrix} \mathbb{P}(h_{2t+1} \in \mathcal{B}_1 | h_{2t-1} \in \mathcal{B}_1, h_{2t}) & \dots & \mathbb{P}(h_{2t+1} \in \mathcal{B}_B | h_{2t-1} \in \mathcal{B}_1, h_{2t}) \\ \vdots & \ddots & \vdots \\ \mathbb{P}(h_{2t+1} \in \mathcal{B}_1 | h_{2t-1} \in \mathcal{B}_B, h_{2t}) & \dots & \mathbb{P}(h_{2t+1} \in \mathcal{B}_B | h_{2t-1} \in \mathcal{B}_B, h_{2t}) \end{bmatrix} \\ &= \begin{bmatrix} \mathbb{P}(h_{2t+1} \in \mathcal{B}_1 | h_{2t}) & \dots & \mathbb{P}(h_{2t+1} \in \mathcal{B}_B | h_{2t}) \\ \vdots & \ddots & \vdots \\ \mathbb{P}(h_{2t+1} \in \mathcal{B}_1 | h_{2t}) & \dots & \mathbb{P}(h_{2t+1} \in \mathcal{B}_B | h_{2t}) \end{bmatrix}.\end{aligned}$$

We can see that the rows of Γ_t are the same, which means that the hidden states are conditionally independent given the imputed states. For the augmented observation matrix P_t we have

$$P_t = \text{diag}\left(p(h_{2t} | h_{2t-1} \in \mathcal{B}_j)\right)_{j=1,\dots,B},$$

so P_t and Γ_t condition on the same hidden state. The observation matrix has the form

$$Q_t = \text{diag}\left(p(y_{2t}, y_{2t+1} | h_{2t+1} \in \mathcal{B}_j, h_{2t})\right)_{j=1,\dots,B}.$$

Inserting Q_t, P_t and Q_t in (4.6) with $\tau(t) = 2t+1$, $a(t) = 2t$ and $o(t) = \{2t, 2t+1\}$ leads to

$$\hat{p}_B(y, \boldsymbol{h_{2T}}) = p(h_0)\mathbf{u}_0 Q_0 \prod_{t=1}^{T^*} P_t \Gamma_t Q_t \mathbf{1}, \tag{A.6}$$

where $\mathbf{u}_0 = \begin{bmatrix} \mathbb{P}(h_1 \in \mathcal{B}_k | h_0) & \dots & \mathbb{P}(h_1 \in \mathcal{B}_B | h_0) \end{bmatrix}$ and $Q_0 = \text{diag}\left(y_1 | h_1 \in \mathcal{B}_k\right)_{k=1,\dots,B}$. Then (A.6) is an HMM-based approximation to (A.5) converging to its true value in $B \to \infty$ and $b_0 \to -\infty$, $b_B \to \infty$.

SVML model For the SVML model we only need to adjust the matrices P_t and Q_t as the dependence structure of the observations remains unchanged

$$\begin{aligned}\Gamma_t &= \begin{bmatrix} \mathbb{P}(h_{2t+1} \in \mathcal{B}_1 | h_{2t}, y_{2t}) & \dots & \mathbb{P}(h_{2t+1} \in \mathcal{B}_B | h_{2t}, y_{2t}) \\ \vdots & \ddots & \vdots \\ \mathbb{P}(h_{2t+1} \in \mathcal{B}_1 | h_{2t}, y_{2t}) & \dots & \mathbb{P}(h_{2t+1} \in \mathcal{B}_B | h_{2t}, y_{2t}) \end{bmatrix}, \\ P_t &= \text{diag}\left(p(h_{2t} | h_{2t-1} \in \mathcal{B}_j, y_{2t-1})\right)_{j=1,\dots,B}.\end{aligned}$$

B Lapwings dataset

The lapwings dataset plays an important role in statistical ecology and has served as an illustration in several handbooks (see King, 2011; King et al., 2010) and papers (Besbeas et al., 2002, e.g.) in this field. It was also used as an example of a complex statistical model by e.g. Goudie et al. (2018). One of the main reasons for such a particular interest in this species is a sharp decline in its population in recent years: its European population is considered as *near threatened* by International Union for Conservation of Nature (2018), while in Britain in particular is has been moved to the *red list* of species of conservation concern, see The Royal Society for the Protection of Birds (2018) (i.e. of the highest conservation priority, with species needing urgent action) from the *amber list* (mentioned by previous literature, see Besbeas et al., 2002; Brooks et al., 2004). The state of its population is crucial as it serves as an indicator species for other farmland birds, giving us an insight into the dynamics of similar bird species.

We follow the approach of Besbeas et al. (2002) and use three datasets for the lapwings application: the count census data for the population index, the weather data on the number of frost days, and the ring-recovery data. Combining of independent sources of data underlies the integrated population modelling (IPM) framework and allows for a more precise parameter estimation. This is due to the survival parameters $\alpha_i, \beta_i, i \in 1, a$, being common to the state space model for the census data and to the ring-recovery model

Census data The census data are derived from the Common Birds Census (CBC) of the British Trust for Ornithology, which recently has been replaced by the Breeding Bird Survey. The dataset is constructed as annual estimates of the number of breeding female lapwings based on annual counts made at a number of sites around the UK. Since only a small fraction of sites are surveyed each year, the index can be seen as a proxy for the total population size. For comparability, we use the same time span as Brooks et al. (2004) and King (2011), i.e. from 1965 to 1998. The choice of the starting year is there motivated by the fact that in earlier years the index protocol was being standardised. Finally we note that year 1965 is associated with time index $t = 3$, for consistency with the ring-recovery data (to be discussed below) which start in 1963.

Weather data For bird species there is a natural relationship between the survival probabilities and the weather conditions, most importantly winter severity. Following Besbeas et al. (2002) we measure this factor for year t by the number of days between April of year t and March of year $(t+1)$ inclusive in which the temperature in Central England fell below freezing and denote it by $fdays_t$. We further normalise $fdays_t$ to obtain f_t which we use as a regressor in the logistic regression for the survival probabilities. As noted by King (2011), normalisation of covariates is done to improve the mixing of the sampling scheme and to facilitate the interpretation of the parameters of the logistic regression (intercept and slope).

Ring–recovery data Ring–recovery studies aim at estimating demographic parameters of the population under consideration including first-year survival probabilities, adult survival probabilities and mortality probabilities (referred to as 'recovery' probabilities). These studies consist in marking individuals (e.g. with a ring or a tag) at the beginning of period t and then releasing them. In subsequent periods $t + 1, t + 2, \ldots$ the number of dead animals is recorded, where it is assumed that any recovery of a dead animal is immediate. For lapwings, the ringed birds are chicks ("fist-years") and a "period" corresponds to a "bird year" i.e. 12 months from April to March. We analyse the ring-recovery data for the releases from 1963 to 1997, with the recoveries up to 1998.

Ring-recovery data are stored in an array, an example of which is provided in Table 8. The first column corresponds to the number of ringed animals in a given year R_t, $t = 1, \ldots, T$, and the subsequent columns report the number of recovered rings $m_{t,s}$ (i.e. animals found dead) in each following year s, $s = 1, \ldots, S$. Obviously, $m_{t,s} = 0$ for $t > s$. Finally, there is an additional $(S + 1)$th column, with the entries $m_{t,S+1}$ providing the number of individuals ringed in year t but never seen again (their rings are not recovered), $m_{t,S+1} = R_t - \sum_{s=1}^{S} m_{t,s}$.

The parameters of interest are $\phi_{1,s}$, $\phi_{a,s}$ and λ_s. The former two are the conditional probabilities of survival until year $s+1$ of a first-year and an adult, respectively, given such an individual is alive in year s. The latter one is the conditional probability of ring recovery in year s given an individual dies in year s. Let $\boldsymbol{v} = \{v_s\}_{s=1}^{S-1}$ denote a vector of a variable $v_s \in \{\phi_{1,s}, \phi_{a,s}, \lambda_s\}$. Then each row $\boldsymbol{m}_t$ of the m-array is multinomially distributed: $\boldsymbol{m}_t \sim \mathcal{MN}(R_i, \boldsymbol{q}_i)$ ($\mathcal{MN}$ denotes the multinomial distribution), where $\boldsymbol{q}_i$

Year of Ringing	Number Ringed	Year of Recovery										
		1964	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974
1963	1147	14	4	1	2	1	0	1	1	0	0	0
1964	1285		20	3	4	0	1	1	0	0	0	0
1965	1106			10	1	2	2	0	2	2	1	1
1966	1615				9	7	4	2	1	1	0	0
1967	1618					12	1	6	2	0	0	1
1968	2120						9	6	4	0	2	2
1969	2003							10	8	5	3	1
1970	1963								8	3	2	0
1971	2463									4	1	1
1972	3092										7	2
1973	3442											15

Table 8: A fragment of Ring-Recovery Data for lapwings for the years 1963-1973, table from King (2011).

are the multinomial cell probabilities specified for $s = 1, \ldots, S$ as[11]

$$q_{t,s} = \begin{cases} 0, & t > s, \\ (1 - \phi_{1,t})\lambda_s & t = s, \\ \phi_{1,t}\lambda_s(1 - \phi_{a,s-1}) \prod_{k=1}^{j-2} \phi_{a,k}, & t > s \end{cases}$$

ans for $s = S + 1$ as $q_{t,s} = 1 - \sum_{s=1}^{S} q_{t,s}$.

The likelihood of the m-array is then given by

$$p(\mathbf{m}|\boldsymbol{\phi_1}, \boldsymbol{\phi_a}, \boldsymbol{\lambda}) \propto \prod_{\mathbf{t=1}}^{\mathbf{T}} \prod_{\mathbf{s=1}}^{\mathbf{S+1}} \mathbf{q}_{\mathbf{t,s}}^{\mathbf{m_{t,s}}}.$$

The array $\boldsymbol{m} = [m_{t,s}]_{t=1,\ldots,T}^{s=1,\ldots,S+1}$ is a sufficient statistic for ring-recovery data.

C Conditional state distribution for the SV model with leverage

Following (Zucchini et al., 2016), we aim at deriving the conditional distribution of h_{t+1} given $\boldsymbol{\theta}$, h_t and y_t. Below, we skip $\boldsymbol{\theta}$ in the conditioning to simplify notation. Since $y_t = \exp(h_t/2)\varepsilon_t$, after demeaning (by $\mu + \phi(h_t - \mu)$), this is the distribution of η_t given h_t and ε_t.

The distribution of $\eta_t|\varepsilon_t$ can be obtained using the basic result from multivariate normal regression, which we recall below for convenience:

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} \right) \Rightarrow x|y \sim \mathcal{N}\left(\mu_x + \frac{\sigma_{xy}}{\sigma_y^2}(y - \mu_y), \sigma_x^2 - \frac{\sigma_{xy}^2}{\sigma_y^2} \right).$$

Hence, we obtain

$$\eta_t|\varepsilon_t \sim \mathcal{N}\left(0 + \frac{\rho\sigma}{1}(y - 0), \sigma^2 - \frac{\rho^2\sigma^2}{1} \right) = \mathcal{N}(\rho\sigma\varepsilon_t, \sigma^2(1 - \rho^2))$$

so that

$$h_{t+1}|h_t, \varepsilon_t \sim \mathcal{N}\big(\mu + \rho(h_t - \mu) + \rho\sigma\varepsilon_t, \sigma^2(1 - \rho^2)\big).$$

Finally, we can express the latter in terms of the actual observation y_t rather than the unobserved

[11] For $j - 2 < t$ we put $\prod_{k=1}^{j-2} := 1$.

disturbance ε_t. For the basic SV model this becomes

$$h_{t+1}|h_t, \varepsilon_t \sim \mathcal{N}\left(\mu + \rho(h_t - \mu) + \rho\sigma\frac{y_t}{\exp(h_t/2)}, \sigma^2(1-\rho^2)\right),$$

which is the result reported in Section 5.2, while for the SVM we have

$$h_{t+1}|h_t, \varepsilon_t \sim \mathcal{N}\left(\mu + \rho(h_t - \mu) + \rho\sigma\frac{y_t - \beta\exp(h_t)}{\exp(h_t/2)}, \sigma^2(1-\rho^2)\right).$$