Semi-Complete Data Augmentation for Efficient State-Space Model Fitting

Agnieszka Borowska

University of Glasgow

Joint with: Ruth King

03.07.2018

Outline

Motivation and context

Motivation

Overall goal:
to develop a novel model-fitting algorithm for state-space models, to permit standard "vanilla" algorithms to be efficiently applied.

Context

- State space models (SSM): 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;
  - the observation process.
- Price: considerably more complicated fitting to data as the associated likelihood is typically analytically intractable.
- Common approaches: Data Augmentation (DA) and numerical integration ⇒ often inefficient and/or unfeasible.
- "Vanilla" MCMC algorithms may perform very poorly due to high correlation between the imputed states, leading to the need to specialist algorithms being developed.

Context

- State space models (SSM): 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;
  - the observation process.
- Price: considerably more complicated fitting to data as the associated likelihood is typically analytically intractable.
- Common approaches: Data Augmentation (DA) and numerical integration ⇒ often inefficient and/or unfeasible.
- "Vanilla" MCMC algorithms may perform very poorly due to high correlation between the imputed states, leading to the need to specialist algorithms being developed.

Context

- State space models (SSM): 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;
  - the observation process.
- Price: considerably more complicated fitting to data as the associated likelihood is typically analytically intractable.
- Common approaches: Data Augmentation (DA) and numerical integration ⇒ often inefficient and/or unfeasible.
- "Vanilla" MCMC algorithms may perform very poorly due to high correlation between the imputed states, leading to the need to specialist algorithms being developed.

Context

- State space models (SSM): 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;
  - the observation process.
- Price: considerably more complicated fitting to data as the associated likelihood is typically analytically intractable.
- Common approaches: Data Augmentation (DA) and numerical integration $\Rightarrow$ often inefficient and/or unfeasible.
- "Vanilla" MCMC algorithms may perform very poorly due to high correlation between the imputed states, leading to the need to specialist algorithms being developed.

Context

- State space models (SSM): 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;
  - the observation process.
- Price: considerably more complicated fitting to data as the associated likelihood is typically analytically intractable.
- Common approaches: Data Augmentation (DA) and numerical integration ⇒ often inefficient and/or unfeasible.
- "Vanilla" MCMC algorithms may perform very poorly due to high correlation between the imputed states, leading to the need to specialist algorithms being developed.

Contributions

1. **Semi-Complete Data Augmentation:** a Bayesian hybrid approach efficiently combining DA and numerical integration.
2. Extending the specific *semi-complete data likelihood* approach of King et al. (2016) to the the **general class of SSM**.
3. **Improving efficiency** while still using "vanilla" MCMC algorithms.
4. Proposing various **integration schemes** based on Hidden Markov Models (HMM) embedding.
5. Utilising the **graphical structure** of the problem to identify conditionally independent latent states to "integrate out".

Contributions

1. **Semi-Complete Data Augmentation:** a Bayesian hybrid approach efficiently combining DA and numerical integration.
2. Extending the specific *semi-complete data likelihood* approach of King et al. (2016) to the the **general class of SSM**.
3. **Improving efficiency** while still using "vanilla" MCMC algorithms.
4. Proposing various **integration schemes** based on Hidden Markov Models (HMM) embedding.
5. Utilising the **graphical structure** of the problem to identify conditionally independent latent states to "integrate out".

Contributions

1. **Semi-Complete Data Augmentation:** a Bayesian hybrid approach efficiently combining DA and numerical integration.
2. Extending the specific *semi-complete data likelihood* approach of King et al. (2016) to the the **general class of SSM**.
3. **Improving efficiency** while still using “vanilla” MCMC algorithms.
4. Proposing various **integration schemes** based on Hidden Markov Models (HMM) embedding.
5. Utilising the **graphical structure** of the problem to identify conditionally independent latent states to “integrate out”.

Contributions

1. **Semi-Complete Data Augmentation:** a Bayesian hybrid approach efficiently combining DA and numerical integration.
2. Extending the specific *semi-complete data likelihood* approach of King et al. (2016) to the the **general class of SSM**.
3. **Improving efficiency** while still using "vanilla" MCMC algorithms.
4. Proposing various **integration schemes** based on Hidden Markov Models (HMM) embedding.
5. Utilising the **graphical structure** of the problem to identify conditionally independent latent states to "integrate out".

Contributions

1. **Semi-Complete Data Augmentation:** a Bayesian hybrid approach efficiently combining DA and numerical integration.
2. Extending the specific *semi-complete data likelihood* approach of King et al. (2016) to the the **general class of SSM**.
3. **Improving efficiency** while still using "vanilla" MCMC algorithms.
4. Proposing various **integration schemes** based on Hidden Markov Models (HMM) embedding.
5. Utilising the **graphical structure** of the problem to identify conditionally independent latent states to "integrate out".

State space models

State space model

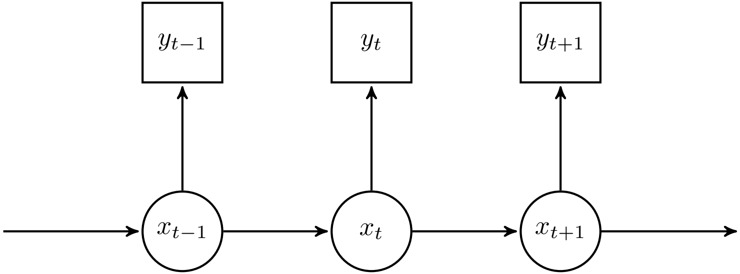
Described via two distinct processes:

$$\boldsymbol{y}_t \sim p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\theta}), \tag{1}$$
$$\boldsymbol{x}_{t+1} \sim p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t, \boldsymbol{\theta}), \tag{2}$$
$$\boldsymbol{x}_0 \sim p(\boldsymbol{\theta}). \tag{3}$$

- $\boldsymbol{y} = (y_1, \dots, y_T)$ – observations;
- $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_T)$ – latent states
(with $\boldsymbol{x}_t = [x_{1,t}, \dots, x_{D,t}]^T$ potentially multivariate);
- $\boldsymbol{\theta}$ – static model parameters with a prior $p(\boldsymbol{\theta})$.

State space model (cont'd)

A graphical representation of the general first-order SSM:
squares – observations, circles – unknown latent states.



Intractable likelihood

The *observed data likelihood* for (1)–(3):

$$
\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{\theta}) &= \int p(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x} \\
&= \int p(x_0|\boldsymbol{\theta}) \prod_{t=1}^{T} p(y_t|x_t, \boldsymbol{\theta}) p(x_t|x_{t-1}, \boldsymbol{\theta}) d\boldsymbol{x},
\end{aligned}
$$

Estimation challenge: observed data likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$ typically not available in closed form.

Intractable likelihood

The *observed data likelihood* for (1)–(3):

$$
\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{\theta}) &= \int p(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x} \\
&= \int p(x_0|\boldsymbol{\theta}) \prod_{t=1}^{T} p(y_t|x_t, \boldsymbol{\theta}) p(x_t|x_{t-1}, \boldsymbol{\theta}) d\boldsymbol{x},
\end{aligned}
$$

Estimation challenge: observed data likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$ typically not available in closed form.

Intractable likelihood – solutions

Two dominant approaches:

1. **Numerical integration:**
   Deterministic (quadrature) or stochastic (Monte Carlo, Particle MCMC).
   Cf.: Andrieu and Roberts (2009), Andrieu et al. (2010).

   Problem: curse of dimensionality – feasible when the integral is of a very low dimension; or tuning required.

2. **Data Augmentation (DA):**
   Impute latent $\boldsymbol{x}$ to form the *complete data likelihood* $p(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta})$ available in closed form and use MCMC to marginalise.
   Cf.: Tanner and Wong (1987), Frühwirth-Schnatter (1994).

   Problem: "vanilla" MCMC algorithms inefficient: high posterior correlation and hence poor mixing.

Intractable likelihood – solutions

Two dominant approaches:

1. **Numerical integration:**
   Deterministic (quadrature) or stochastic (Monte Carlo, Particle MCMC).
   Cf.: Andrieu and Roberts (2009), Andrieu et al. (2010).

   Problem: curse of dimensionality – feasible when the integral is of a very low dimension; or tuning required.

2. **Data Augmentation (DA):**
   Impute latent $\boldsymbol{x}$ to form the *complete data likelihood* $p(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta})$ available in closed form and use MCMC to marginalise.
   Cf.: Tanner and Wong (1987), Frühwirth-Schnatter (1994).

   Problem: "vanilla" MCMC algorithms inefficient: high posterior correlation and hence poor mixing.

Intractable likelihood – solutions

Two dominant approaches:

1. **Numerical integration:**
   Deterministic (quadrature) or stochastic (Monte Carlo, Particle MCMC).
   Cf.: Andrieu and Roberts (2009), Andrieu et al. (2010).

   Problem: curse of dimensionality – feasible when the integral is of a very low dimension; or tuning required.

2. **Data Augmentation (DA):**
   Impute latent $\boldsymbol{x}$ to form the *complete data likelihood* $p(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta})$ available in closed form and use MCMC to marginalise.
   Cf.: Tanner and Wong (1987), Frühwirth-Schnatter (1994).

   Problem: "vanilla" MCMC algorithms inefficient: high posterior correlation and hence poor mixing.

Intractable likelihood – solutions

Two dominant approaches:

1. **Numerical integration:**
   Deterministic (quadrature) or stochastic (Monte Carlo, Particle MCMC).
   Cf.: Andrieu and Roberts (2009), Andrieu et al. (2010).

   Problem: curse of dimensionality – feasible when the integral is of a very low dimension; or tuning required.

2. **Data Augmentation (DA):**
   Impute latent $\boldsymbol{x}$ to form the *complete data likelihood* $p(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta})$ available in closed form and use MCMC to marginalise.
   Cf.: Tanner and Wong (1987), Frühwirth-Schnatter (1994).

   Problem: "vanilla" MCMC algorithms inefficient: high posterior correlation and hence poor mixing.

Intractable likelihood – solutions

Two dominant approaches:

1. **Numerical integration:**
   Deterministic (quadrature) or stochastic (Monte Carlo, Particle MCMC).
   Cf.: Andrieu and Roberts (2009), Andrieu et al. (2010).

   Problem: curse of dimensionality – feasible when the integral is of a very low dimension; or tuning required.

2. **Data Augmentation (DA):**
   Impute latent $\boldsymbol{x}$ to form the *complete data likelihood* $p(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta})$ available in closed form and use MCMC to marginalise.
   Cf.: Tanner and Wong (1987), Frühwirth-Schnatter (1994).

   Problem: "vanilla" MCMC algorithms inefficient: high posterior correlation and hence poor mixing.

Semi-Complete Data Augmentation

Semi-Complete Data Augmentation

Bayesian hybrid approach: combining DA and numerical integration.

Key idea: separate the latent state $\boldsymbol{x}$ into two components $\boldsymbol{x} = (\boldsymbol{x}_{ing}, \boldsymbol{x}_{aug})$, the 'integrated' states and the 'augmented' states, respectively.

Semi-Complete Data Augmentation

Bayesian hybrid approach: combining DA and numerical integration.

Key idea: separate the latent state $\boldsymbol{x}$ into two components $\boldsymbol{x} = (\boldsymbol{x_{ing}}, \boldsymbol{x_{aug}})$, the 'integrated' states and the 'augmented' states, respectively.

Semi-Complete Data Likelihood

Define the **semi-complete data likelihood** (SCDL) as $p(\boldsymbol{y}, \boldsymbol{x_{aug}}|\boldsymbol{\theta})$, given by

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

Used to form the **joint posterior** distribution:

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

Semi-Complete Data Likelihood

Define the **semi-complete data likelihood** (SCDL) as $p(\boldsymbol{y}, \boldsymbol{x_{aug}}|\boldsymbol{\theta})$, given by

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

Used to form the **joint posterior** distribution:

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

Auxiliary variables

Specification of the auxiliary variables

Let D_{int} and T_{int} be subsets of dimension and time indices of $\boldsymbol{x}$, respectively, 'suitable' for integration (D_{aug} and T_{aug} – their compliments).

Then the 'integrated' and 'augmented' states are induced by the partition of $\boldsymbol{x}$ into

$$\boldsymbol{x_{int}} = \{x_{d,t}\}_{d \in D_{int}, t \in T_{int}} \quad \text{and} \quad \boldsymbol{x_{aug}} = \{x_{d,t}\}_{d \in D_{aug}, t \in T_{aug}}.$$

For instance:

- for $D = 2$, $D_{int} = \{d_2\}$, $T_{int} = \{0, \ldots, T\}$ – 'horizontal' integration of the second state at all times;
- $D_{int} = \{1, \ldots, D\}$, $T_{int} = \{2t + 1\}_{t=0}^{T/2}$ – 'vertical' integration of all states at odd time periods.

Auxiliary variables

Specification of the auxiliary variables

Let D_{int} and T_{int} be subsets of dimension and time indices of $\boldsymbol{x}$, respectively, 'suitable' for integration (D_{aug} and T_{aug} – their compliments).

Then the 'integrated' and 'augmented' states are induced by the partition of $\boldsymbol{x}$ into

$$\boldsymbol{x_{int}} = \{x_{d,t}\}_{d \in D_{int}, t \in T_{int}} \quad \text{and} \quad \boldsymbol{x_{aug}} = \{x_{d,t}\}_{d \in D_{aug}, t \in T_{aug}}.$$

For instance:

- for $D = 2$, $D_{int} = \{d_2\}$, $T_{int} = \{0, \dots, T\}$ – 'horizontal' integration of the second state at all times;
- $D_{int} = \{1, \dots, D\}$, $T_{int} = \{2t+1\}_{t=0}^{T/2}$ – 'vertical' integration of all states at odd time periods.

Auxiliary variables

Specification of the auxiliary variables

Let D_{int} and T_{int} be subsets of dimension and time indices of $\boldsymbol{x}$, respectively, 'suitable' for integration (D_{aug} and T_{aug} – their compliments).

Then the 'integrated' and 'augmented' states are induced by the partition of $\boldsymbol{x}$ into

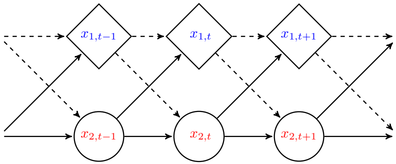
$$\boldsymbol{x_{int}} = \{x_{d,t}\}_{d \in D_{int}, t \in T_{int}} \quad \text{and} \quad \boldsymbol{x_{aug}} = \{x_{d,t}\}_{d \in D_{aug}, t \in T_{aug}}.$$

For instance:

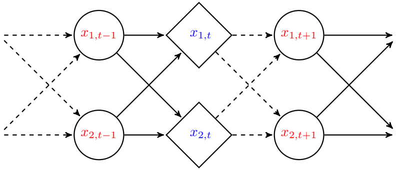
- for $D = 2$, $D_{int} = \{d_2\}$, $T_{int} = \{0, \dots, T\}$ – 'horizontal' integration of the second state at all times;
- $D_{int} = \{1, \dots, D\}$, $T_{int} = \{2t + 1\}_{t=0}^{T/2}$ – 'vertical' integration of all states at odd time periods.

Integration schemes

Two examples of an **integration/augmentation scheme:**
diamonds – the imputed states, circles – the integrated states, dashed lines – the relations *from* the imputed (known) states.



(a) **Horizontal integration**.

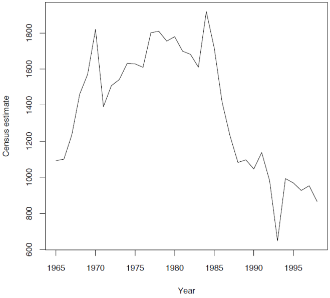


(b) **Vertical integration**

Applications

Lapwings data

$\boldsymbol{y} = (y_1, \ldots, y_T)$ observations on census (count) data on adult population of the British lapwing (*Vanellus vanellus*).
Popular in statistical ecology, cf.: Besbeas et al. (2002), Brooks et al. (2004).



State space model

$$
\begin{aligned}
y_t &\sim \mathcal{N}(N_{a,t}, \sigma_y^2), \\
N_{1,t+1} &\sim \mathcal{P}(N_{a,t}\rho_t\phi_{1,t}), \\
N_{a,t+1} &\sim \mathcal{B}\big((N_{1,t} + N_{a,t}), \phi_{a,t}\big), \\
N_{1,0} &\sim \mathcal{NB}(r_{1,0}, p_{1,0}), \\
N_{a,0} &\sim \mathcal{NB}(r_{a,0}, p_{a,0}).
\end{aligned}
$$

The latent state: $\boldsymbol{x} = \{\boldsymbol{N_1}, \boldsymbol{N_a}\}$ with $\boldsymbol{N_1} = (N_{1,1}, \dots, N_{1,T})$ and $\boldsymbol{N_a} = (N_{a,1}, \dots, N_{a,T})$, the population sizes of 1-years and adults, respectively.

Time varying parameters:

$$
\operatorname{logit} \phi_{i,t} = \alpha_i + \beta_i f_t,\ i \in \{1, a\}, \qquad \log \rho_t = \alpha_\rho + \beta_\rho \tilde{t}.
$$

(Static) parameters: $\theta = (\alpha_1, \alpha_a, \alpha_\rho, \beta_1, \beta_a, \beta_\rho, \sigma_y^2)^T$.

State space model

$$
\begin{aligned}
y_t &\sim \mathcal{N}(N_{a,t}, \sigma_y^2), \\
N_{1,t+1} &\sim \mathcal{P}(N_{a,t}\rho_t\phi_{1,t}), \\
N_{a,t+1} &\sim \mathcal{B}\big((N_{1,t} + N_{a,t}), \phi_{a,t}\big), \\
N_{1,0} &\sim \mathcal{NB}(r_{1,0}, p_{1,0}), \\
N_{a,0} &\sim \mathcal{NB}(r_{a,0}, p_{a,0}).
\end{aligned}
$$

The latent state: $\boldsymbol{x} = \{\boldsymbol{N_1}, \boldsymbol{N_a}\}$ with $\boldsymbol{N_1} = (N_{1,1}, \dots, N_{1,T})$ and $\boldsymbol{N_a} = (N_{a,1}, \dots, N_{a,T})$, the population sizes of 1-years and adults, respectively.

Time varying parameters:

$$
\operatorname{logit} \phi_{i,t} = \alpha_i + \beta_i f_t,\ i \in \{1, a\}, \qquad \log \rho_t = \alpha_\rho + \beta_\rho \tilde{t}.
$$

(Static) parameters: $\theta = (\alpha_1, \alpha_a, \alpha_\rho, \beta_1, \beta_a, \beta_\rho, \sigma_y^2)^T$.

Integration scheme

SCDL:
integrate out $N_{1,t}$ given the imputed value of $N_{a,t}$ and $\boldsymbol{\theta}$;
use the Markov structure of the model to simplify:

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{N_a}|\theta) &= p(\boldsymbol{y}|\boldsymbol{N_a}, \theta)p(\boldsymbol{N_a}|\theta) \\ &= \sum_{\boldsymbol{N_1}} p_0 \left(\prod_{t=1}^{T} p(y_t|N_{a,t}, N_{1,t})p(N_{a,t}, N_{1,t}) \right). \end{aligned}$$

Idea: write the above marginal pmf as an HMM
(exact result possible, up to the upper bound of the integration).

Integration scheme

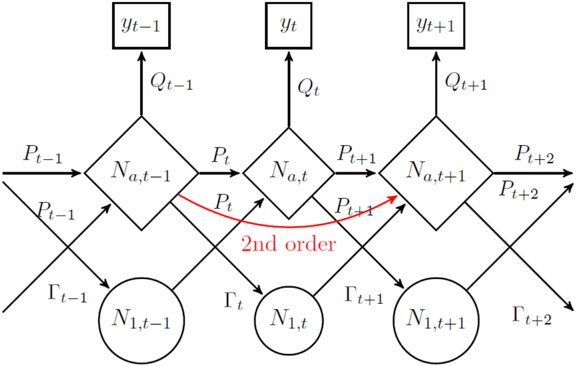
SCDL:
integrate out $N_{1,t}$ given the imputed value of $N_{a,t}$ and $\boldsymbol{\theta}$;
use the Markov structure of the model to simplify:

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{N_a}|\theta) &= p(\boldsymbol{y}|\boldsymbol{N_a}, \theta)p(\boldsymbol{N_a}|\theta) \\ &= \sum_{\boldsymbol{N_1}} p_0 \left(\prod_{t=1}^{T} p(y_t|N_{a,t}, N_{1,t})p(N_{a,t}, N_{1,t}) \right). \end{aligned}$$

Idea: write the above marginal pmf as an HMM
(exact result possible, up to the upper bound of the integration).

Integration scheme

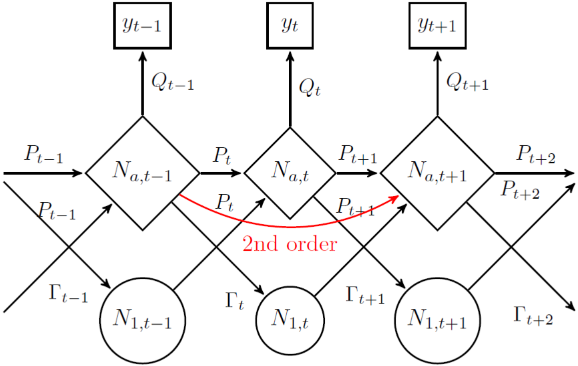
Combining DA and HMM structure. Diamonds – the imputed nodes, squares – the data, circles – the unknown variables.



Removing of the dependence of N_a on N_1 via integration lead to a second order HMM on N_a.

Integration scheme

Combining DA and HMM structure. Diamonds – the imputed nodes, squares – the data, circles – the unknown variables.



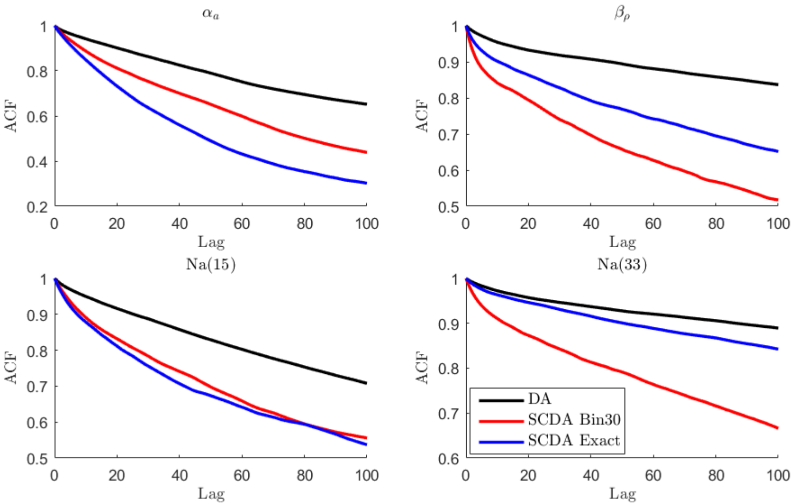
Removing of the dependence of $\boldsymbol{N_a}$ on $\boldsymbol{N_1}$ via integration lead to a second order HMM on $\boldsymbol{N_a}$.

Results

Effective sample sizes (ESS) for $M = 10,000$ draws:

Method		α_1	α_a	α_ρ	β_1	β_a	β_ρ
DA	ESS	49.071	26.675	20.703	94.289	60.003	18.810
[619.76 s]	ESS/sec.	0.079	0.043	0.033	0.152	0.097	0.030
SCDA Exact	ESS	229.047	22.130	11.331	245.528	98.708	14.136
[948.12 s]	ESS/sec.	0.242	0.023	0.012	0.259	0.104	0.015
SCDA Bin30	ESS	246.576	62.439	41.000	259.054	67.991	21.828
[526.24 s]	ESS/sec.	0.469	0.119	0.078	0.492	0.129	0.041

Results (cont'd)



Stochastic Volatility model

The state space model:

$$
\begin{aligned}
y_t &= \exp(h_t/2)\varepsilon_t \\
h_{t+1} &= \mu + \phi(h_t - \mu) + \sigma\eta_t, \\
\varepsilon_t, \eta_t &\overset{i.i.d.}{\sim} \mathcal{N}(0,1), \\
h_0 &\sim \mathcal{N}\left(\mu, \frac{\sigma^2}{1-\phi^2}\right), \\
\boldsymbol{\theta} &= (\mu, \phi, \sigma^2)^T.
\end{aligned}
$$

Extensions easy to incorporate:

- SV in the mean of Koopman and Uspensky (2002):

$$y_t = \beta \exp(h_t) + \exp(h_t/2)\varepsilon_t;$$

- SV with leverage Jungbacker and Koopman (2007):

$$\operatorname{corr}(\varepsilon_t, \eta_t) = \rho \neq 0.$$

Stochastic Volatility model

The state space model:

$$\begin{aligned} y_t &= \exp(h_t/2)\varepsilon_t \\ h_{t+1} &= \mu + \phi(h_t - \mu) + \sigma\eta_t, \\ \varepsilon_t, \eta_t &\overset{i.i.d.}{\sim} \mathcal{N}(0,1), \\ h_0 &\sim \mathcal{N}\left(\mu, \frac{\sigma^2}{1-\phi^2}\right), \\ \boldsymbol{\theta} &= (\mu, \phi, \sigma^2)^T. \end{aligned}$$

Extensions easy to incorporate:

- SV in the mean of Koopman and Uspensky (2002):

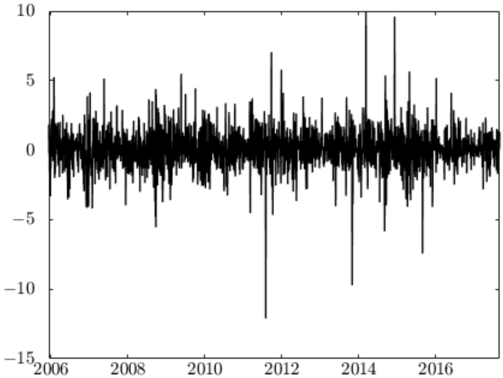
$$y_t = \beta \exp(h_t) + \exp(h_t/2)\varepsilon_t;$$

- SV with leverage Jungbacker and Koopman (2007):

$$\text{corr}(\varepsilon_t, \eta_t) = \rho \neq 0.$$

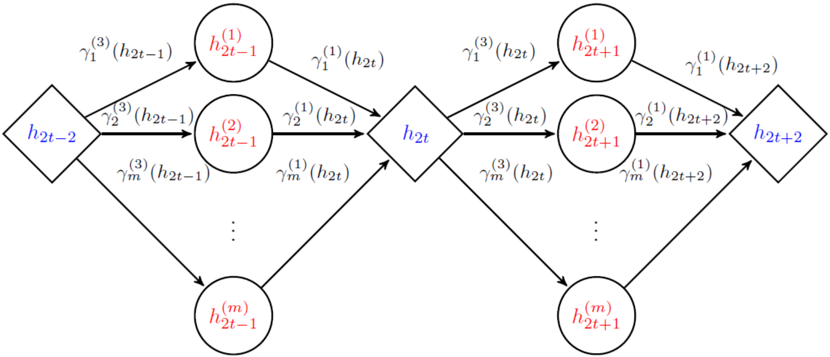
Data

Data: $T = 2000$ MSFT stock returns to 31 Aug 2017.



Integration scheme

Combining DA and the HMM-based integration: a *single imputation problem* of h_{2t} with the associated integrations. Diamonds – the imputed states, circles – the integrated states.



Results

Effective Sample Sizes for $M = 10,000$ draws:

Method	μ	ϕ	σ^2	h_{600}	h_{1000}	h_{1800}
DA	17.890	5.347	5.146	138.882	178.258	298.147
SCDA fix	6.403	358.743	5.011	276.54	77.321	18.834
SCDA adapt	205.907	16.490	16.246	521.829	727.313	782.701

Conclusions

Conclusions

- **Semi-Complete Data Augmentation:** a novel efficient estimation method for state space models, combining **Data Augmentation** with **numerical integration**.
- Integration: to reduce the dependence between the imputed auxiliary variables (cf. Rao-Blackwellisation).
- Integration schemes based on the insights from **Hidden Markov Models:** specify new transition probabilities between the redefined states, to be numerically integrated out, conditionally on the auxiliary variables.
- **"Binning"** for further efficiency gains: approximating similar values of a state with e.g. a single mid-value.
  (a natural starting point for any MC based analysis for continuous states).
- The split of the latent states into "auxiliary" and "integrated" variables: **model-dependent** and specified in such a way that the algorithm is efficient.

Conclusions

- **Semi-Complete Data Augmentation:** a novel efficient estimation method for state space models, combining **Data Augmentation** with **numerical integration**.
- Integration: to reduce the dependence between the imputed auxiliary variables (cf. Rao-Blackwellisation).
- Integration schemes based on the insights from **Hidden Markov Models:** specify new transition probabilities between the redefined states, to be numerically integrated out, conditionally on the auxiliary variables.
- **"Binning"** for further efficiency gains: approximating similar values of a state with e.g. a single mid-value.
  (a natural starting point for any MC based analysis for continuous states).
- The split of the latent states into "auxiliary" and "integrated" variables: **model-dependent** and specified in such a way that the algorithm is efficient.

Conclusions

- **Semi-Complete Data Augmentation:** a novel efficient estimation method for state space models, combining **Data Augmentation** with **numerical integration**.
- Integration: to reduce the dependence between the imputed auxiliary variables (cf. Rao-Blackwellisation).
- Integration schemes based on the insights from **Hidden Markov Models:** specify new transition probabilities between the redefined states, to be numerically integrated out, conditionally on the auxiliary variables.
- **"Binning"** for further efficiency gains: approximating similar values of a state with e.g. a single mid-value.
  (a natural starting point for any MC based analysis for continuous states).
- The split of the latent states into "auxiliary" and "integrated" variables: **model-dependent** and specified in such a way that the algorithm is efficient.

Conclusions

- **Semi-Complete Data Augmentation:** a novel efficient estimation method for state space models, combining **Data Augmentation** with **numerical integration**.
- Integration: to reduce the dependence between the imputed auxiliary variables (cf. Rao-Blackwellisation).
- Integration schemes based on the insights from **Hidden Markov Models:** specify new transition probabilities between the redefined states, to be numerically integrated out, conditionally on the auxiliary variables.
- **"Binning"** for further efficiency gains: approximating similar values of a state with e.g. a single mid-value.
(a natural starting point for any MC based analysis for continuous states).
- The split of the latent states into "auxiliary" and "integrated" variables: **model-dependent** and specified in such a way that the algorithm is efficient.

Conclusions

- **Semi-Complete Data Augmentation:** a novel efficient estimation method for state space models, combining **Data Augmentation** with **numerical integration**.
- Integration: to reduce the dependence between the imputed auxiliary variables (cf. Rao-Blackwellisation).
- Integration schemes based on the insights from **Hidden Markov Models:** specify new transition probabilities between the redefined states, to be numerically integrated out, conditionally on the auxiliary variables.
- **"Binning"** for further efficiency gains: approximating similar values of a state with e.g. a single mid-value.
  (a natural starting point for any MC based analysis for continuous states).
- The split of the latent states into "auxiliary" and "integrated" variables: **model-dependent** and specified in such a way that the algorithm is efficient.

Further research

- Replacing a deterministic integration with a stochastic one: importance sampling $\Rightarrow$ what importance distribution?
- Adopting insights from Bayesian Networks (e.g. *d-separation*) to identify conditionally independent latent states in the general case.
- High dimensional integration remains a challenging problem $\Rightarrow$ SMC samplers (Del Moral et al., 2006)?

Further research

- Replacing a deterministic integration with a stochastic one: importance sampling ⇒ what importance distribution?
- Adopting insights from Bayesian Networks (e.g. *d-separation*) to identify conditionally independent latent states in the general case.
- High dimensional integration remains a challenging problem ⇒ SMC samplers (Del Moral et al., 2006)?

Further research

- Replacing a deterministic integration with a stochastic one: importance sampling $\Rightarrow$ what importance distribution?
- Adopting insights from Bayesian Networks (e.g. *d-separation*) to identify conditionally independent latent states in the general case.
- High dimensional integration remains a challenging problem $\Rightarrow$ SMC samplers (Del Moral et al., 2006)?

References I

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.

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.

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.

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

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

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

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.

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

References II

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

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.

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.

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.