Extended ensemble Kalman filters for high-dimensional hierarchical state-space models

Matthias Katzfuss

Department of Statistics
Texas A&M University

Joint work with Jon Stroud (Georgetown) and Chris Wikle (Missouri)

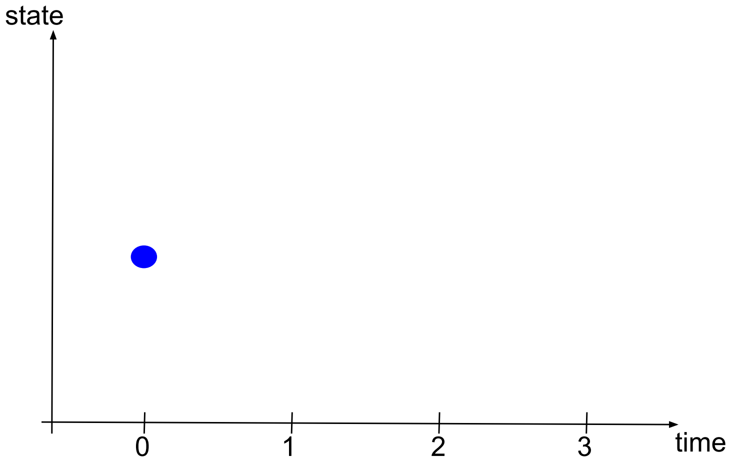
Outline

1. Data assimilation
2. Review of the EnKF
3. Extended EnKFs
4. Numerical examples
5. Conclusions

Outline

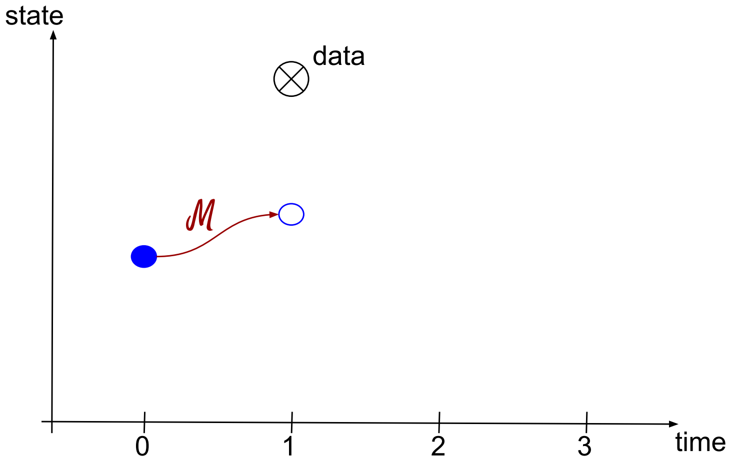
Data assimilation

Data assimilation (DA): sequentially infer the true state of a system, by combining (noisy) observations with an evolution operator $\mathcal{M}$



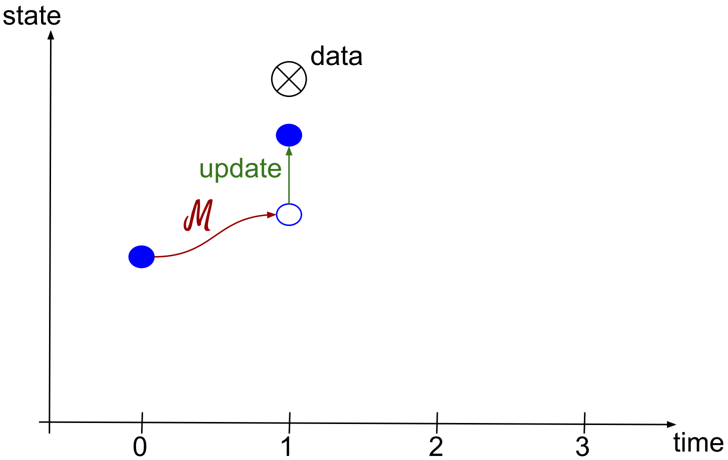
Data assimilation

Data assimilation (DA): sequentially infer the true state of a system, by combining (noisy) observations with an evolution operator $\mathcal{M}$



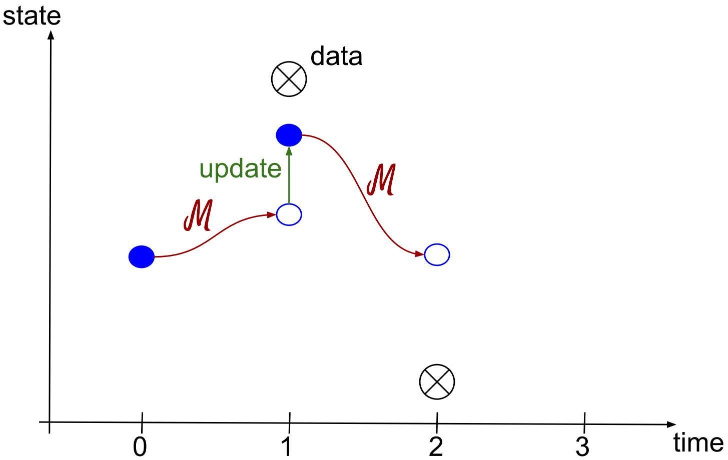
Data assimilation

Data assimilation (DA): sequentially infer the true state of a system, by combining (noisy) observations with an evolution operator $\mathcal{M}$



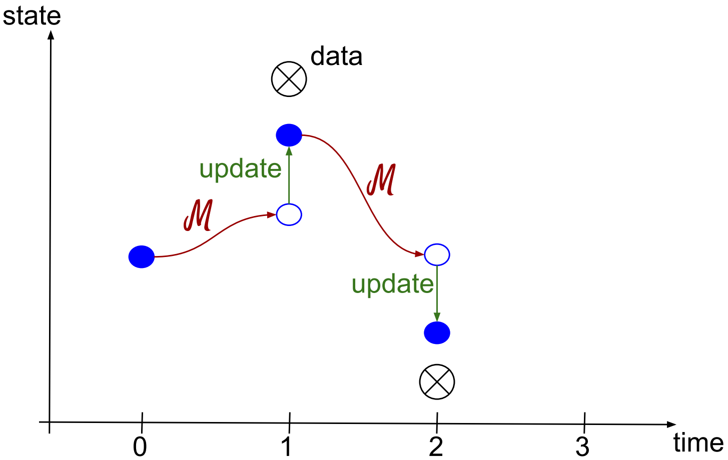
Data assimilation

Data assimilation (DA): sequentially infer the true state of a system, by combining (noisy) observations with an evolution operator $\mathcal{M}$



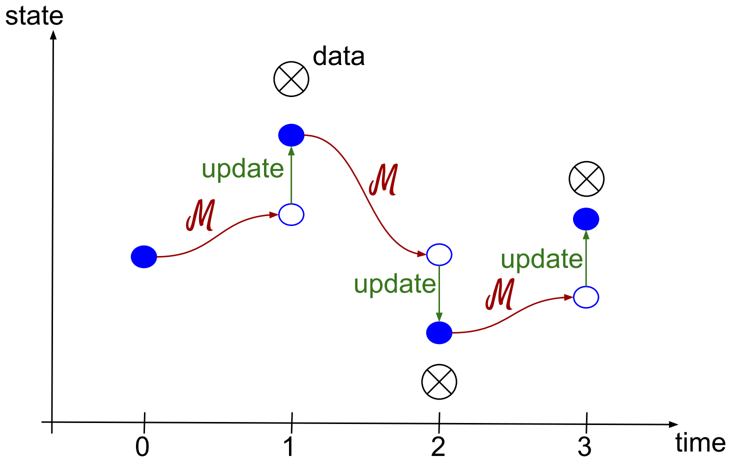
Data assimilation

Data assimilation (DA): sequentially infer the true state of a system, by combining (noisy) observations with an evolution operator $\mathcal{M}$



Data assimilation

Data assimilation (DA): sequentially infer the true state of a system, by combining (noisy) observations with an evolution operator $\mathcal{M}$



Data assimilation: Goals and examples

Goals:

- State inference, initialization for forecasting, model calibration, . . .

Examples of environmental applications:

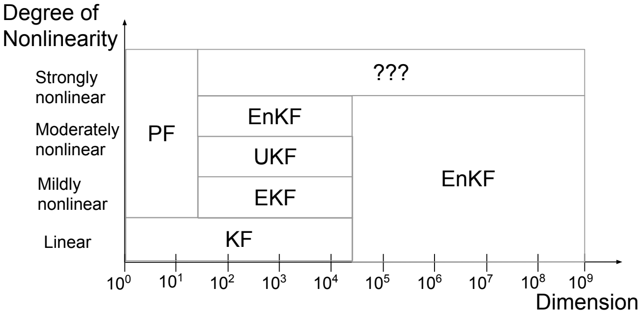
- weather forecasting
- climate studies
- pollution monitoring

In geophysical applications, the state typically consists of one or more (discretized) spatial fields, the state dimension is often massive, and the evolution model is highly complex and nonlinear

Reviews: e.g., Wikle and Berliner, 2007; Evensen, 2009; Nychka and Anderson, 2010; Houtekamer and Zhang, 2016

Overview of data assimilation methods

Best methods (in terms of accuracy and computation time) for data assimilation as a function of state dimension and the degree of nonlinearity:



KF = Kalman filter; PF = particle filter; EKF = extended KF;
UKF = unscented KF; EnKF = ensemble KF

Here we only consider methods for sequential, probabilistic DA, not variational methods (e.g., 4DVAR; Talagrand and Courtier 1987)

Notation: The state-space model (SSM)

From a statistical perspective: DA is filtering in a SSM

SSM with additive Gaussian error in discrete time $t = 1, 2, \ldots$:

$$\begin{aligned} \mathbf{y}_t &= \mathbf{H}_t \mathbf{x}_t + \mathbf{v}_t, & \mathbf{v}_t &\sim \mathcal{N}_{m_t}(\mathbf{0}, \mathbf{R}_t), \\ \mathbf{x}_t &= \mathcal{M}_t(\mathbf{x}_{t-1}) + \mathbf{w}_t, & \mathbf{w}_t &\sim \mathcal{N}_n(\mathbf{0}, \mathbf{Q}_t), \end{aligned}$$

where

- $\mathbf{y}_t$ is the m_t-dimensional vector of observations at time t
- $\mathbf{x}_t$ is the n-dimensional state of primary interest
- $\mathbf{H}_t$ is the observation matrix
- $\mathcal{M}_t(\cdot)$ is the (possibly nonlinear) evolution operator
- errors $\mathbf{v}_t$ and $\mathbf{w}_t$ are mutually and serially independent
- initial state: $\mathbf{x}_0 \sim \mathcal{N}_n(\boldsymbol{\mu}_{0|0}, \boldsymbol{\Sigma}_{0|0})$
- no unknown parameters (for now)

Note that this SSM is very general

The Kalman filter (KF; Kalman, 1960)

Filtering: At each time t, find cond. distrib. of $\mathbf{x}_t$ given $\mathbf{y}_{1:t} = \{\mathbf{y}_1, \ldots, \mathbf{y}_t\}$

If $\mathcal{M}_t$ is linear (i.e., $\mathcal{M}_t(\mathbf{x}_{t-1}) = \mathbf{M}_t \mathbf{x}_{t-1}$), can use KF:

For $t = 1, 2, \ldots$:

- $\mathbf{x}_{t-1} | \mathbf{y}_{1:t-1} \sim \mathcal{N}_n(\boldsymbol{\mu}_{t-1|t-1}, \boldsymbol{\Sigma}_{t-1|t-1})$ computed at $t-1$
- Forecast step: $\mathbf{x}_t | \mathbf{y}_{1:t-1} \sim \mathcal{N}_n(\boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1})$, where

$$\boldsymbol{\mu}_{t|t-1} := \mathbf{M}_t \boldsymbol{\mu}_{t-1|t-1}, \quad \boldsymbol{\Sigma}_{t|t-1} := \mathbf{M}_t \boldsymbol{\Sigma}_{t-1|t-1} \mathbf{M}_t' + \mathbf{Q}_t$$

- Update step: $\mathbf{x}_t | \mathbf{y}_{1:t} \sim \mathcal{N}_n(\boldsymbol{\mu}_{t|t}, \boldsymbol{\Sigma}_{t|t})$, where

$$\boldsymbol{\mu}_{t|t} := \boldsymbol{\mu}_{t|t-1} + \mathbf{K}_t(\mathbf{y}_t - \mathbf{H}_t \boldsymbol{\mu}_{t|t-1}), \quad \boldsymbol{\Sigma}_{t|t} := (\mathbf{I}_n - \mathbf{K}_t \mathbf{H}_t) \boldsymbol{\Sigma}_{t|t-1}$$

  and $\mathbf{K}_t := \boldsymbol{\Sigma}_{t|t-1} \mathbf{H}_t' (\mathbf{H}_t \boldsymbol{\Sigma}_{t|t-1} \mathbf{H}_t' + \mathbf{R}_t)^{-1}$ is the $n \times m_t$ Kalman gain

KF is exact, but infeasible if $\mathcal{M}_t$ nonlinear and/or n and m_t large

The Kalman filter (KF; Kalman, 1960)

Filtering: At each time t, find cond. distrib. of $\mathbf{x}_t$ given $\mathbf{y}_{1:t} = \{\mathbf{y}_1, \ldots, \mathbf{y}_t\}$

If $\mathcal{M}_t$ is linear (i.e., $\mathcal{M}_t(\mathbf{x}_{t-1}) = \mathbf{M}_t \mathbf{x}_{t-1}$), can use KF:

For $t = 1, 2, \ldots$:

- $\mathbf{x}_{t-1} | \mathbf{y}_{1:t-1} \sim \mathcal{N}_n(\boldsymbol{\mu}_{t-1|t-1}, \boldsymbol{\Sigma}_{t-1|t-1})$ computed at $t-1$
- Forecast step: $\mathbf{x}_t | \mathbf{y}_{1:t-1} \sim \mathcal{N}_n(\boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1})$, where

$$\boldsymbol{\mu}_{t|t-1} := \mathbf{M}_t \boldsymbol{\mu}_{t-1|t-1}, \quad \boldsymbol{\Sigma}_{t|t-1} := \mathbf{M}_t \boldsymbol{\Sigma}_{t-1|t-1} \mathbf{M}_t' + \mathbf{Q}_t$$

- Update step: $\mathbf{x}_t | \mathbf{y}_{1:t} \sim \mathcal{N}_n(\boldsymbol{\mu}_{t|t}, \boldsymbol{\Sigma}_{t|t})$, where

$$\boldsymbol{\mu}_{t|t} := \boldsymbol{\mu}_{t|t-1} + \mathbf{K}_t(\mathbf{y}_t - \mathbf{H}_t \boldsymbol{\mu}_{t|t-1}), \quad \boldsymbol{\Sigma}_{t|t} := (\mathbf{I}_n - \mathbf{K}_t \mathbf{H}_t) \boldsymbol{\Sigma}_{t|t-1}$$

and $\mathbf{K}_t := \boldsymbol{\Sigma}_{t|t-1} \mathbf{H}_t' (\mathbf{H}_t \boldsymbol{\Sigma}_{t|t-1} \mathbf{H}_t' + \mathbf{R}_t)^{-1}$ is the $n \times m_t$ Kalman gain

KF is exact, but infeasible if $\mathcal{M}_t$ nonlinear and/or n and m_t large

The Kalman filter (KF; Kalman, 1960)

Filtering: At each time t, find cond. distrib. of $\mathbf{x}_t$ given $\mathbf{y}_{1:t} = \{\mathbf{y}_1, \dots, \mathbf{y}_t\}$

If $\mathcal{M}_t$ is linear (i.e., $\mathcal{M}_t(\mathbf{x}_{t-1}) = \mathbf{M}_t \mathbf{x}_{t-1}$), can use KF:

For $t = 1, 2, \dots$:

- $\mathbf{x}_{t-1} | \mathbf{y}_{1:t-1} \sim \mathcal{N}_n(\boldsymbol{\mu}_{t-1|t-1}, \boldsymbol{\Sigma}_{t-1|t-1})$ computed at $t-1$
- Forecast step: $\mathbf{x}_t | \mathbf{y}_{1:t-1} \sim \mathcal{N}_n(\boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1})$, where

$$\boldsymbol{\mu}_{t|t-1} := \mathbf{M}_t \boldsymbol{\mu}_{t-1|t-1}, \quad \boldsymbol{\Sigma}_{t|t-1} := \mathbf{M}_t \boldsymbol{\Sigma}_{t-1|t-1} \mathbf{M}_t' + \mathbf{Q}_t$$

- Update step: $\mathbf{x}_t | \mathbf{y}_{1:t} \sim \mathcal{N}_n(\boldsymbol{\mu}_{t|t}, \boldsymbol{\Sigma}_{t|t})$, where

$$\boldsymbol{\mu}_{t|t} := \boldsymbol{\mu}_{t|t-1} + \mathbf{K}_t(\mathbf{y}_t - \mathbf{H}_t \boldsymbol{\mu}_{t|t-1}), \quad \boldsymbol{\Sigma}_{t|t} := (\mathbf{I}_n - \mathbf{K}_t \mathbf{H}_t) \boldsymbol{\Sigma}_{t|t-1}$$

  and $\mathbf{K}_t := \boldsymbol{\Sigma}_{t|t-1} \mathbf{H}_t' (\mathbf{H}_t \boldsymbol{\Sigma}_{t|t-1} \mathbf{H}_t' + \mathbf{R}_t)^{-1}$ is the $n \times m_t$ Kalman gain

KF is exact, but infeasible if $\mathcal{M}_t$ nonlinear and/or n and m_t large

The Kalman filter (KF; Kalman, 1960)

Filtering: At each time t, find cond. distrib. of $\mathbf{x}_t$ given $\mathbf{y}_{1:t} = \{\mathbf{y}_1, \ldots, \mathbf{y}_t\}$

If $\mathcal{M}_t$ is linear (i.e., $\mathcal{M}_t(\mathbf{x}_{t-1}) = \mathbf{M}_t \mathbf{x}_{t-1}$), can use KF:

For $t = 1, 2, \ldots$:

- $\mathbf{x}_{t-1} | \mathbf{y}_{1:t-1} \sim \mathcal{N}_n(\boldsymbol{\mu}_{t-1|t-1}, \boldsymbol{\Sigma}_{t-1|t-1})$ computed at $t-1$
- Forecast step: $\mathbf{x}_t | \mathbf{y}_{1:t-1} \sim \mathcal{N}_n(\boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1})$, where

$$\boldsymbol{\mu}_{t|t-1} := \mathbf{M}_t \boldsymbol{\mu}_{t-1|t-1}, \quad \boldsymbol{\Sigma}_{t|t-1} := \mathbf{M}_t \boldsymbol{\Sigma}_{t-1|t-1} \mathbf{M}_t' + \mathbf{Q}_t$$

- Update step: $\mathbf{x}_t | \mathbf{y}_{1:t} \sim \mathcal{N}_n(\boldsymbol{\mu}_{t|t}, \boldsymbol{\Sigma}_{t|t})$, where

$$\boldsymbol{\mu}_{t|t} := \boldsymbol{\mu}_{t|t-1} + \mathbf{K}_t(\mathbf{y}_t - \mathbf{H}_t \boldsymbol{\mu}_{t|t-1}), \quad \boldsymbol{\Sigma}_{t|t} := (\mathbf{I}_n - \mathbf{K}_t \mathbf{H}_t) \boldsymbol{\Sigma}_{t|t-1}$$

and $\mathbf{K}_t := \boldsymbol{\Sigma}_{t|t-1} \mathbf{H}_t' (\mathbf{H}_t \boldsymbol{\Sigma}_{t|t-1} \mathbf{H}_t' + \mathbf{R}_t)^{-1}$ is the $n \times m_t$ Kalman gain

KF is exact, but infeasible if $\mathcal{M}_t$ nonlinear and/or n and m_t large

Outline

The ensemble Kalman filter (EnKF; Evensen 1994)

Approximate version of the KF for large or nonlinear SSMs, in which the state distribution is represented by a sample or "ensemble"

Assume ensemble $\mathbf{x}_{t-1|t-1}^{(1)}, \ldots, \mathbf{x}_{t-1|t-1}^{(N)}$ is sample from filtering distribution at time $t-1$.
For $i = 1, \ldots, N$:

1. Forecast Step: Apply the evolution model:

$$\mathbf{x}_{t|t-1}^{(i)} = \mathcal{M}_t(\mathbf{x}_{t-1|t-1}^{(i)}) + \mathbf{w}_t^{(i)}, \quad \mathbf{w}_t^{(i)} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{Q}_t)$$

2. Update Step: Update by applying a linear "shift":

$$\mathbf{x}_{t|t}^{(i)} = \mathbf{x}_{t|t-1}^{(i)} + \widehat{\mathbf{K}}_t(\widetilde{\mathbf{y}}_t^{(i)} - \mathbf{H}_t\mathbf{x}_{t|t-1}^{(i)})$$

where $\widehat{\mathbf{K}}_t$ is an approximation of the Kalman gain, and $\widetilde{\mathbf{y}}_t^{(i)} = \mathbf{y}_t + \mathbf{v}_t^{(i)}$ is a perturbed observation

The ensemble Kalman filter (EnKF; Evensen 1994)

Approximate version of the KF for large or nonlinear SSMs, in which the state distribution is represented by a sample or "ensemble"

Assume ensemble $\mathbf{x}_{t-1|t-1}^{(1)}, \ldots, \mathbf{x}_{t-1|t-1}^{(N)}$ is sample from filtering distribution at time $t-1$.
For $i = 1, \ldots, N$:

1. Forecast Step: Apply the evolution model:

$$\mathbf{x}_{t|t-1}^{(i)} = \mathcal{M}_t(\mathbf{x}_{t-1|t-1}^{(i)}) + \mathbf{w}_t^{(i)}, \quad \mathbf{w}_t^{(i)} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{Q}_t)$$

2. Update Step: Update by applying a linear "shift":

$$\mathbf{x}_{t|t}^{(i)} = \mathbf{x}_{t|t-1}^{(i)} + \widehat{\mathbf{K}}_t(\widetilde{\mathbf{y}}_t^{(i)} - \mathbf{H}_t \mathbf{x}_{t|t-1}^{(i)})$$

where $\widehat{\mathbf{K}}_t$ is an approximation of the Kalman gain, and $\widetilde{\mathbf{y}}_t^{(i)} = \mathbf{y}_t + \mathbf{v}_t^{(i)}$ is a perturbed observation

The ensemble Kalman filter (EnKF; Evensen 1994)

Approximate version of the KF for large or nonlinear SSMs, in which the state distribution is represented by a sample or "ensemble"

Assume ensemble $\mathbf{x}_{t-1|t-1}^{(1)}, \ldots, \mathbf{x}_{t-1|t-1}^{(N)}$ is sample from filtering distribution at time $t-1$.
For $i = 1, \ldots, N$:

1. Forecast Step: Apply the evolution model:

$$\mathbf{x}_{t|t-1}^{(i)} = \mathcal{M}_t(\mathbf{x}_{t-1|t-1}^{(i)}) + \mathbf{w}_t^{(i)}, \quad \mathbf{w}_t^{(i)} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{Q}_t)$$

2. Update Step: Update by applying a linear "shift":

$$\mathbf{x}_{t|t}^{(i)} = \mathbf{x}_{t|t-1}^{(i)} + \widehat{\mathbf{K}}_t(\widetilde{\mathbf{y}}_t^{(i)} - \mathbf{H}_t\mathbf{x}_{t|t-1}^{(i)})$$

where $\widehat{\mathbf{K}}_t$ is an approximation of the Kalman gain, and $\widetilde{\mathbf{y}}_t^{(i)} = \mathbf{y}_t + \mathbf{v}_t^{(i)}$ is a perturbed observation

The ensemble Kalman filter (EnKF; Evensen 1994)

Approximate version of the KF for large or nonlinear SSMs, in which the state distribution is represented by a sample or "ensemble"

Assume ensemble $\mathbf{x}_{t-1|t-1}^{(1)}, \ldots, \mathbf{x}_{t-1|t-1}^{(N)}$ is sample from filtering distribution at time $t-1$.
For $i = 1, \ldots, N$:

1. Forecast Step: Apply the evolution model:

$$\mathbf{x}_{t|t-1}^{(i)} = \mathcal{M}_t(\mathbf{x}_{t-1|t-1}^{(i)}) + \mathbf{w}_t^{(i)}, \quad \mathbf{w}_t^{(i)} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{Q}_t)$$

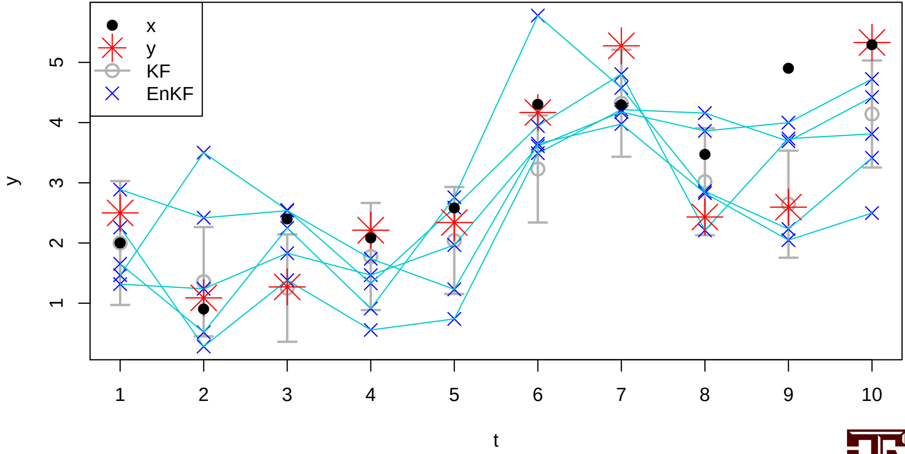
2. Update Step: Update by applying a linear "shift":

$$\mathbf{x}_{t|t}^{(i)} = \mathbf{x}_{t|t-1}^{(i)} + \widehat{\mathbf{K}}_t(\widetilde{\mathbf{y}}_t^{(i)} - \mathbf{H}_t \mathbf{x}_{t|t-1}^{(i)})$$

where $\widehat{\mathbf{K}}_t$ is an approximation of the Kalman gain, and $\widetilde{\mathbf{y}}_t^{(i)} = \mathbf{y}_t + \mathbf{v}_t^{(i)}$ is a perturbed observation

EnKF: Illustration

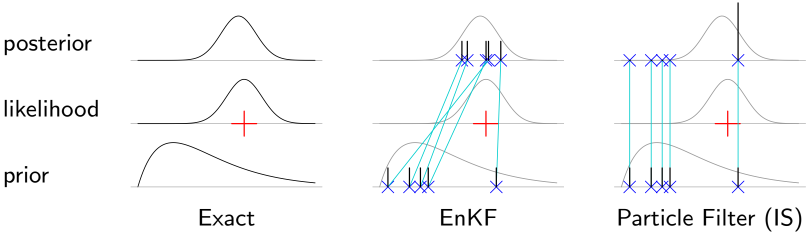
Simulation from a linear one-dimensional SSM for 10 time points:



State, observation, Kalman filter, and EnKF ($N = 5$)

Illustration of the update step

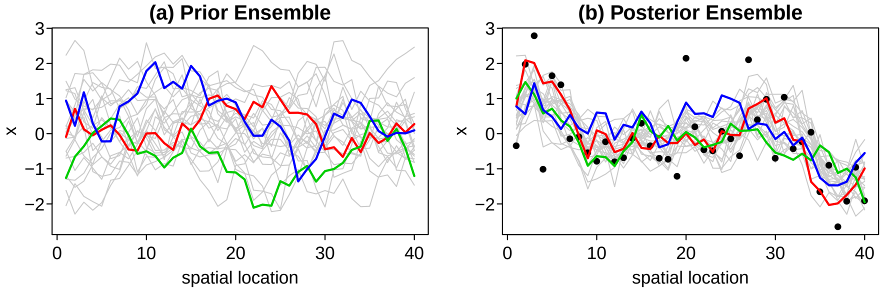
Illustration of different updating schemes at a single time point



Particle weights degenerate for large n (Snyder et al., 2008)

Spatial update step

Update step for a one-dimensional spatial field



Estimation of the Kalman gain

For EnKF update step, need to estimate $n \times m_t$ Kalman gain from ensemble of size N, where typically $N \ll n$, and so we need regularization.

If $\mathbf{x}_t$ consists of one or more spatial fields, we can set

$$\widehat{\mathbf{K}}_t := \widehat{\Sigma}_{t|t-1}\mathbf{H}_t'(\mathbf{H}_t\widehat{\Sigma}_{t|t-1}\mathbf{H}_t' + \mathbf{R}_t)^{-1},$$

where $\widehat{\Sigma}_{t|t-1} = \mathbf{S}_t \circ \mathcal{T}_t$ is the Hadamard (entrywise) product of

- $\mathbf{S}_t$, the sample covariance matrix of the forecast ensemble, and
- $\mathcal{T}_t$, a sparse positive definite correlation ("tapering") matrix (Houtekamer and Mitchell, 2001; Furrer and Bengtsson, 2007)

Estimation of the Kalman gain

For EnKF update step, need to estimate $n \times m_t$ Kalman gain from ensemble of size N, where typically $N \ll n$, and so we need regularization.

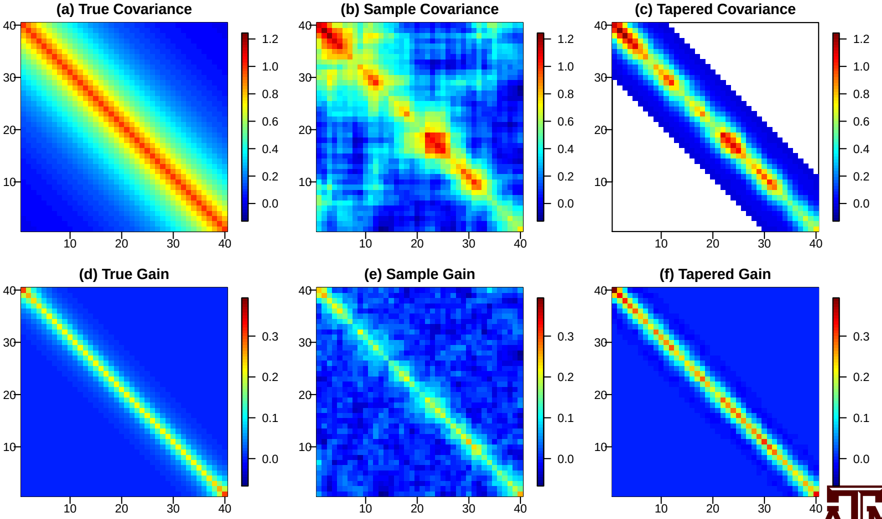
If $\mathbf{x}_t$ consists of one or more spatial fields, we can set

$$\widehat{\mathbf{K}}_t := \widehat{\boldsymbol{\Sigma}}_{t|t-1}\mathbf{H}_t'(\mathbf{H}_t\widehat{\boldsymbol{\Sigma}}_{t|t-1}\mathbf{H}_t' + \mathbf{R}_t)^{-1},$$

where $\widehat{\boldsymbol{\Sigma}}_{t|t-1} = \mathbf{S}_t \circ \mathcal{T}_t$ is the Hadamard (entrywise) product of

- $\mathbf{S}_t$, the sample covariance matrix of the forecast ensemble, and
- $\mathcal{T}_t$, a sparse positive definite correlation ("tapering") matrix (Houtekamer and Mitchell, 2001; Furrer and Bengtsson, 2007)

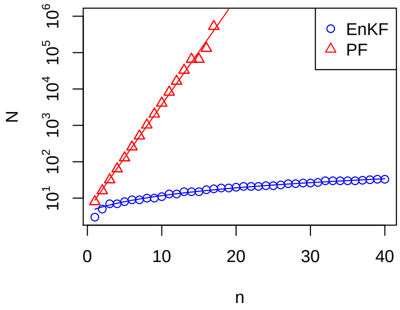
Tapering: Illustration



$n = m = 40, \ N = 25$

EnKF likelihood

- Can also approximate likelihood using EnKF
- Properties:
  - Biased estimator of the true likelihood
  - But much lower variance than the unbiased particle likelihood
    - For bounded variance of loglikelihood in independent case:
      EnKF: $N = \mathcal{O}(n)$; PF: $N = \mathcal{O}(e^n)$
    - Necessary N to keep the variance of the loglikelihood below 2 for independent case:



Outline

Hierarchical state-space model (HSSM)

Starting with $\mathbf{x}_0 \sim \mathcal{N}_n(\boldsymbol{\mu}_{0|0}, \boldsymbol{\Sigma}_{0|0})$, we assume for $t = 1, 2, \ldots$:

$$\begin{aligned}
\mathbf{z}_t | \mathbf{y}_t, \boldsymbol{\theta}_t &\sim f_t(\mathbf{y}_t; \boldsymbol{\theta}_t) \\
\mathbf{y}_t &= \mathbf{H}_t(\boldsymbol{\theta}_t)\,\mathbf{x}_t + \mathbf{v}_t, \quad \mathbf{v}_t \sim \mathcal{N}_{m_t}(\mathbf{0}, \mathbf{R}_t(\boldsymbol{\theta}_t)) \\
\mathbf{x}_t &= \mathcal{M}_t(\mathbf{x}_{t-1}; \boldsymbol{\theta}_t) + \mathbf{w}_t, \quad \mathbf{w}_t \sim \mathcal{N}_n(\mathbf{0}, \mathbf{Q}_t(\boldsymbol{\theta}_t)) \\
\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1} &\sim p_t(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1})
\end{aligned}$$

where $\mathbf{z}_t$ are the actual measurements, and we have added a transformation layer and a parameter layer to the additive Gaussian SSM (in black).

Existing methods for parameter inference

Most popular: state augmentation (Anderson, 2001), but does not work well if the correlation between states and parameters is low (DelSole and Yang, 2010)

Other possible approaches for parameter estimation are sequential maximum likelihood (e.g., Dee and da Silva, 1999; Mitchell and Houtekamer, 2000) and sequential Bayes (e.g., Stroud and Bengtsson, 2007; Frei and Kunsch, 2012) — for specific parameters

Also some work on choosing tuning parameters (e.g., Anderson, 2007a,b)

Basic idea of extended EnKFs

Conditional on $\mathbf{y}_t$ and $\boldsymbol{\theta}_t$, the HSSM reduces to the "standard" SSM, for which the EnKF is applicable

$\rightarrow$ Take existing techniques for Bayesian inference (e.g., Gibbs sampler, particle filter), but replace the part requiring integrating out or sampling from $\mathbf{x}_t$ by the EnKF

Examples:

- Gibbs-EnKF
- Particle-EnKF
- Gibbs-EnKS
- ...

Basic idea of extended EnKFs

Conditional on $\mathbf{y}_t$ and $\boldsymbol{\theta}_t$, the HSSM reduces to the "standard" SSM, for which the EnKF is applicable

$\rightarrow$ Take existing techniques for Bayesian inference (e.g., Gibbs sampler, particle filter), but replace the part requiring integrating out or sampling from $\mathbf{x}_t$ by the EnKF

Examples:

- Gibbs-EnKF
- Particle-EnKF
- Gibbs-EnKS
- ...

The Gibbs Ensemble Kalman Filter (GEnKF)

Assume forecasts of state and parameters are independent. Then:

Initialize $\mathbf{x}_{0|0}^{(j)} \overset{iid}{\sim} \mathcal{N}_n(\boldsymbol{\mu}_{0|0}, \boldsymbol{\Sigma}_{0|0})$, $j = 1, \ldots, N$. For $t = 1, 2, \ldots$:

1. Forecast step: $\mathbf{x}_{t|t-1}^{(j)} = \mathcal{M}_t(\mathbf{x}_{t-1|t-1}^{(j)}) + \mathbf{w}_t^{(j)}$, $j = 1, \ldots, N$.
2. Find starting values for $\mathbf{y}_t^{(j)}$ and $\boldsymbol{\theta}_t^{(j)}$, $j = 1, \ldots, N$.
3. For $j = 1, \ldots, N$, repeat G times (until convergence):
   (a) EnKF update of $\mathbf{x}_{t|t}^{(j)}$ from $\mathbf{x}_{t|t-1}^{(j)}$ based on $\mathbf{y}_t^{(j)}$, $\boldsymbol{\theta}_t^{(j)}$, $\mathbf{x}_{t|t-1}^{(1:N)}$.
   (b) Sample $\boldsymbol{\theta}_t^{(j)}$ from FCD $[\boldsymbol{\theta}_t | \mathbf{y}_t^{(j)}, \mathbf{x}_{t|t}^{(j)}, \mathbf{z}_t]$.
   (c) Sample $\mathbf{y}_t^{(j)}$ from FCD $[\mathbf{y}_t | \mathbf{x}_{t|t}^{(j)}, \boldsymbol{\theta}_t^{(j)}, \mathbf{z}_t]$.

   Then, each $(\mathbf{x}_{t|t}^{(j)}, \boldsymbol{\theta}_t^{(j)})$ is a joint sample from $[\mathbf{x}_t, \boldsymbol{\theta}_t | \mathbf{z}_{1:t}]$.

Particle EnKF (for low-dimensional parameters)

Initialize the algorithm with an ensemble of ensembles: $(\boldsymbol{\theta}_0^{(i)}, \mathbf{x}_{0|0}^{(i,j)})$ with $w_0^{(i)} = 1/M$, $i = 1, \ldots, M$; $j = 1, \ldots, N$. Then, for $t = 1, 2, \ldots$:

1. For $i = 1, \ldots, M$:
   (a) Propagate $\boldsymbol{\theta}_t^{(i)}$ and compute corresponding forecast ensemble $\mathbf{x}_{t|t-1}^{(i,1:N)}$
   (b) Calculate the particle weight: $w_t^{(i)} \propto w_{t-1}^{(i)} \mathcal{L}_t^Z(\mathbf{z}_t | \boldsymbol{\theta}_t^{(i)}, \mathbf{x}_{t|t-1}^{(i,1:N)})$
   (c) Compute filtering states $\mathbf{x}_{t|t}^{(i,1:N)}$ via EnKF update
2. Filtering distribution:
   $[\boldsymbol{\theta}_t, \mathbf{x}_t | \mathbf{z}_{1:t}] \approx \sum_{i=1}^M w_t^{(i)} \frac{1}{N} \sum_{j=1}^N \delta_{(\boldsymbol{\theta}_t^{(i)}, \mathbf{x}_{t|t}^{(i,j)})}(\boldsymbol{\theta}_t, \mathbf{x}_t)$
3. If desired, resample the particles $(\boldsymbol{\theta}_t^{(i)}, \mathbf{x}_{t|t}^{(i,1:N)})$

Likelihood $\mathcal{L}_t^Z(\mathbf{z}_t | \theta_t, \mathbf{x}_{t|t-1}^{(i,1:N)})$ is approximated using EnKF

In the case of forecast independence of states and parameters, only need a single ensemble (cf. Frei & Kunsch, 2012)

Particle EnKF (for low-dimensional parameters)

Initialize the algorithm with an ensemble of ensembles: $(\boldsymbol{\theta}_0^{(i)}, \mathbf{x}_{0|0}^{(i,j)})$ with $w_0^{(i)} = 1/M$, $i = 1, \ldots, M$; $j = 1, \ldots, N$. Then, for $t = 1, 2, \ldots$:

1. For $i = 1, \ldots, M$:
   (a) Propagate $\boldsymbol{\theta}_t^{(i)}$ and compute corresponding forecast ensemble $\mathbf{x}_{t|t-1}^{(i,1:N)}$
   (b) Calculate the particle weight: $w_t^{(i)} \propto w_{t-1}^{(i)} \mathcal{L}_t^Z(\mathbf{z}_t | \boldsymbol{\theta}_t^{(i)}, \mathbf{x}_{t|t-1}^{(i,1:N)})$
   (c) Compute filtering states $\mathbf{x}_{t|t}^{(i,1:N)}$ via EnKF update
2. Filtering distribution:
   $[\boldsymbol{\theta}_t, \mathbf{x}_t | \mathbf{z}_{1:t}] \approx \sum_{i=1}^M w_t^{(i)} \frac{1}{N} \sum_{j=1}^N \delta_{(\boldsymbol{\theta}_t^{(i)}, \mathbf{x}_{t|t}^{(i,j)})}(\boldsymbol{\theta}_t, \mathbf{x}_t)$
3. If desired, resample the particles $(\boldsymbol{\theta}_t^{(i)}, \mathbf{x}_{t|t}^{(i,1:N)})$

Likelihood $\mathcal{L}_t^Z(\mathbf{z}_t | \boldsymbol{\theta}_t, \mathbf{x}_{t|t-1}^{(i,1:N)})$ is approximated using EnKF

In the case of forecast independence of states and parameters, only need a single ensemble (cf. Frei & Kunsch, 2012)

Gibbs-EnKS

Want smoothing distribution $[\mathbf{x}_{1:T}, \boldsymbol{\theta}_{1:T} | \mathbf{z}_{1:T}]$ for fixed T.

1. Initialize $\mathbf{y}_{1:T|T}$ and $\boldsymbol{\theta}_{1:T|T}$.
2. Iterate between following steps until convergence:
   (a) Obtain samples $\mathbf{x}_{1:T|T}^{(1:N)}$ from $[\mathbf{x}_{1:T} | \boldsymbol{\theta}_{1:T|T}, \mathbf{y}_{1:T|T}]$ using the EnKS (Evensen & van Leeuwen, 2000), and set $\mathbf{x}_{1:T|T} = \mathbf{x}_{1:T|T}^{(j)}$ for j sampled uniformly at random from $\{1, \ldots, N\}$.
   (b) Draw a sample $\mathbf{y}_{1:T|T}$ from $[\mathbf{y}_{1:T} | \mathbf{z}_{1:T}, \mathbf{x}_{1:T|T}, \boldsymbol{\theta}_{1:T|T}]$.
   (c) Draw a sample $\boldsymbol{\theta}_{1:T|T}$ from $[\boldsymbol{\theta}_{1:T} | \mathbf{x}_{1:T|T}, \mathbf{y}_{1:T|T}, \mathbf{z}_{1:T|T}]$.

Works even for high-dimensional parameters, if full conditional distribution is available in closed form

Properties of the extended EnKFs

- For linear Gaussian SSMs, convergence to the true posteriors as $N \to \infty$ and M or the number of MCMC iterations increase
- For small N, algorithms will tend to perform well for HSSMs for which the embedded SSM is well suited for inference using the EnKF and EnKS
- Computational complexity:
  - EnKF: Have to apply $\mathcal{M}_t$ to N ensemble members; update is $\mathcal{O}(nN^2)$ for most EnKF variants (e.g., Tippett et al 2003)
  - Extended EnKFs: In general, have to carry out EnKF several times. But: Often only a small number of iterations or particles is necessary, and only the update has to be repeated
  - Thus, increased computational cost of extended EnKFs is minor in some applications

Outline

Data with outliers

Heavy-tailed noise distribution: $v_{t,i} \sim t_\nu(0, \sigma_t^2)$, ν small

Special case of our HSSM:

$\mathbf{z}_t = \mathbf{y}_t$ and $\mathbf{R}_t(\theta_t) = \sigma_t^2 \, diag(\theta_{t,1}, \ldots, \theta_{t,m_t})$, where $\theta_{t,l} \stackrel{ind.}{\sim} IG(\nu/2, \nu/2)$

GEnKF (Robust EnKF):

Update step: For $j = 1, \ldots, N$, repeat G times (until convergence):

(a) EnKF update of $\mathbf{x}_{t|t}^{(j)}$ from $\mathbf{x}_{t|t-1}^{(j)}$ based on $\mathbf{y}_t$, $\theta_t^{(j)}$, $\mathbf{x}_{t|t-1}^{(1:N)}$.

(b) Sample $\theta_{t,l}^{(j)} \stackrel{ind.}{\sim} IG(\nu/2 + \frac{1}{2}, \nu/2 + \big(\frac{y_{t,l} - (\mathbf{H}_t \mathbf{x}_{t|t}^{(j)})_l}{\sigma_t}\big)^2/2)$, $l = 1, \ldots, m_t$

Data with outliers

Heavy-tailed noise distribution: $v_{t,i} \sim t_\nu(0, \sigma_t^2)$, ν small

Special case of our HSSM:

$\mathbf{z}_t = \mathbf{y}_t$ and $\mathbf{R}_t(\boldsymbol{\theta}_t) = \sigma_t^2\, diag(\theta_{t,1}, \dots, \theta_{t,m_t})$, where $\theta_{t,l} \overset{ind.}{\sim} IG(\nu/2, \nu/2)$

GEnKF (Robust EnKF):

Update step: For $j = 1, \dots, N$, repeat G times (until convergence):

(a) EnKF update of $\mathbf{x}_{t|t}^{(j)}$ from $\mathbf{x}_{t|t-1}^{(j)}$ based on $\mathbf{y}_t$, $\boldsymbol{\theta}_t^{(j)}$, $\mathbf{x}_{t|t-1}^{(1:N)}$.

(b) Sample $\theta_{t,l}^{(j)} \overset{ind.}{\sim} IG\big(\nu/2 + \frac{1}{2}, \nu/2 + \big(\frac{y_{t,l} - (\mathbf{H}_t \mathbf{x}_{t|t}^{(j)})_l}{\sigma_t}\big)^2/2\big)$, $l = 1, \dots, m_t$

Data with outliers

Heavy-tailed noise distribution: $v_{t,i} \sim t_\nu(0, \sigma_t^2)$, ν small

Special case of our HSSM:

$\mathbf{z}_t = \mathbf{y}_t$ and $\mathbf{R}_t(\boldsymbol{\theta}_t) = \sigma_t^2 \, diag(\theta_{t,1}, \ldots, \theta_{t,m_t})$, where $\theta_{t,l} \overset{ind.}{\sim} IG(\nu/2, \nu/2)$

GEnKF (Robust EnKF):

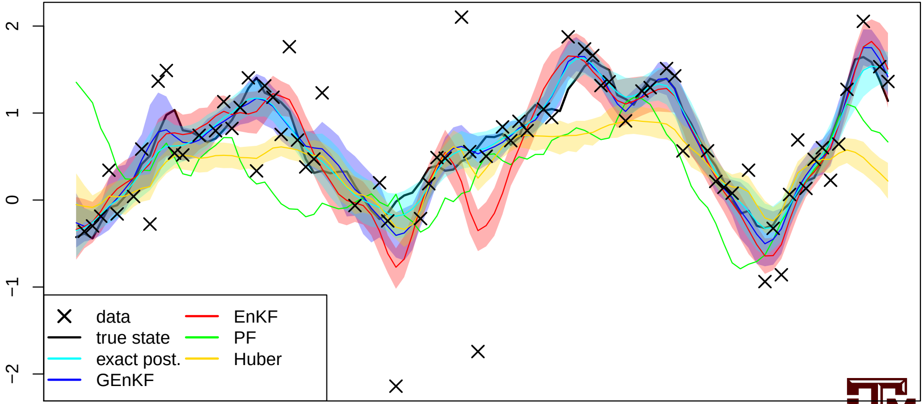
Update step: For $j = 1, \ldots, N$, repeat G times (until convergence):

(a) EnKF update of $\mathbf{x}_{t|t}^{(j)}$ from $\mathbf{x}_{t|t-1}^{(j)}$ based on $\mathbf{y}_t$, $\boldsymbol{\theta}_t^{(j)}$, $\mathbf{x}_{t|t-1}^{(1:N)}$.

(b) Sample $\theta_{t,l}^{(j)} \overset{ind.}{\sim} IG\big(\nu/2 + \frac{1}{2}, \nu/2 + \big(\frac{y_{t,l} - (\mathbf{H}_t \mathbf{x}_{t|t}^{(j)})_l}{\sigma_t}\big)^2/2\big)$, $l = 1, \ldots, m_t$

Example: Heavy-tailed data

Simulated heavy-tailed data (with $v_{t,l}/\sigma_t \sim t_2$)



Example: Threshold Models

- Challenge: Observation distributions with point masses (e.g., binary)
- Use transformation equations involving indicator functions
- Example: Rainfall amounts:

$$z_{t,l} = g(y_{t,l};\boldsymbol{\theta}_t) = \begin{cases} y_{t,l}^{\kappa_t}, & y_{t,l} > 0 \\ 0, & y_{t,l} \leq 0 \end{cases}$$

for some $\kappa_t > 1$. Assume $\mathbf{R}_t = diag(\sigma_{t,1}^2, \dots, \sigma_{t,m_t}^2)$.

Gibbs-EnKF for rainfall data:

- Step 3(c): Independently:

$$y_{t,l} | z_{t,l}, \mathbf{x}_t, \kappa_t \begin{cases} = z_{t,l}^{1/\kappa_t}, & z_{t,l} > 0 \\ \sim \mathcal{N}^-\big((\mathbf{H}_t\mathbf{x}_t)_l, \sigma_{t,l}^2\big), & z_{t,l} = 0 \end{cases}$$

- Step 3(b): If κ_t is unknown, sample from $[\kappa_t | \mathbf{x}_{t|t}^{(j)}, \mathbf{z}_t]$

Example: Threshold Models

- Challenge: Observation distributions with point masses (e.g., binary)
- Use transformation equations involving indicator functions
- Example: Rainfall amounts:

$$z_{t,l} = g(y_{t,l}; \boldsymbol{\theta}_t) = \begin{cases} y_{t,l}^{\kappa_t}, & y_{t,l} > 0 \\ 0, & y_{t,l} \leq 0 \end{cases}$$

for some $\kappa_t > 1$. Assume $\mathbf{R}_t = diag(\sigma_{t,1}^2, \ldots, \sigma_{t,m_t}^2)$.

Gibbs-EnKF for rainfall data:

- Step 3(c): Independently:

$$y_{t,l} | z_{t,l}, \mathbf{x}_t, \kappa_t \begin{cases} = z_{t,l}^{1/\kappa_t}, & z_{t,l} > 0 \\ \sim \mathcal{N}^-((\mathbf{H}_t \mathbf{x}_t)_l, \sigma_{t,l}^2), & z_{t,l} = 0 \end{cases}$$

- Step 3(b): If κ_t is unknown, sample from $[\kappa_t | \mathbf{x}_{t|t}^{(j)}, \mathbf{z}_t]$

Example: Threshold Models

- Challenge: Observation distributions with point masses (e.g., binary)
- Use transformation equations involving indicator functions
- Example: Rainfall amounts:

$$z_{t,l} = g(y_{t,l}; \boldsymbol{\theta}_t) = \begin{cases} y_{t,l}^{\kappa_t}, & y_{t,l} > 0 \\ 0, & y_{t,l} \leq 0 \end{cases}$$

for some $\kappa_t > 1$. Assume $\mathbf{R}_t = diag(\sigma^2_{t,1}, \ldots, \sigma^2_{t,m_t})$.

Gibbs-EnKF for rainfall data:

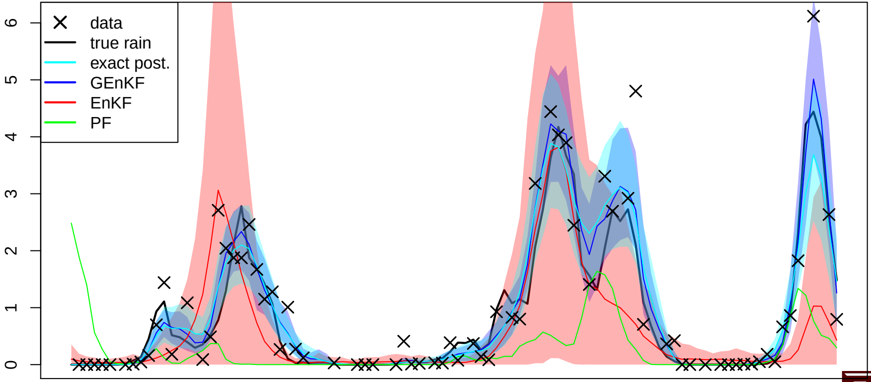
- Step 3(c): Independently:

$$y_{t,l} | z_{t,l}, \mathbf{x}_t, \kappa_t \begin{cases} = z_{t,l}^{1/\kappa_t}, & z_{t,l} > 0 \\ \sim \mathcal{N}^-\big((\mathbf{H}_t \mathbf{x}_t)_l, \sigma^2_{t,l}\big), & z_{t,l} = 0 \end{cases}$$

- Step 3(b): If κ_t is unknown, sample from $[\kappa_t | \mathbf{x}^{(j)}_{t|t}, \mathbf{z}_t]$

Example: Threshold models

Simulated rainfall data (with $\kappa_t = 3$)



Interlude: Probabilistic predictions

- Predictions should be probabilistic in nature
- Goal: maximize sharpness, subject to calibration
- Proper scoring rules: Joint assessment of calibration and sharpness
  - Proper: Minimized in expectation by true distribution (optimal pred.)
    - Examples: squared error, interval score (width + coverage)
  - Strictly proper: Uniquely minimized in expectation by true distribution
    - Examples: Log score (negative log-likelihood), continuous ranked probability score (CRPS – generalizes absolute error)
- Not proper: likelihood, interval width or empirical coverage, ...

For more details, see:
Gneiting, T., and Katzfuss, M. 2014. Probabilistic forecasting. *Annual Review of Statistics and its Application*, 1, 125–151.

Interlude: Probabilistic predictions

- Predictions should be probabilistic in nature
- Goal: maximize sharpness, subject to calibration
- Proper scoring rules: Joint assessment of calibration and sharpness
  - Proper: Minimized in expectation by true distribution (optimal pred.)
    - Examples: squared error, interval score (width + coverage)
  - Strictly proper: Uniquely minimized in expectation by true distribution
    - Examples: Log score (negative log-likelihood), continuous ranked probability score (CRPS – generalizes absolute error)
- Not proper: likelihood, interval width or empirical coverage, ...

For more details, see:

Gneiting, T., and Katzfuss, M. 2014. Probabilistic forecasting. *Annual Review of Statistics and its Application*, 1, 125–151.

Interlude: Probabilistic predictions

- Predictions should be probabilistic in nature
- Goal: maximize sharpness, subject to calibration
- Proper scoring rules: Joint assessment of calibration and sharpness
    - Proper: Minimized in expectation by true distribution (optimal pred.)
        - Examples: squared error, interval score (width + coverage)
    - Strictly proper: Uniquely minimized in expectation by true distribution
        - Examples: Log score (negative log-likelihood), continuous ranked probability score (CRPS – generalizes absolute error)
- Not proper: likelihood, interval width or empirical coverage, ...

For more details, see:
Gneiting, T., and Katzfuss, M. 2014. Probabilistic forecasting. *Annual Review of Statistics and its Application*, 1, 125–151.

Simulation study: Non-Gaussian obs. at a single time point

Simulated heavy-tailed and rainfall data. Compared GEnKF update to (matrix-free) EnKF and PF (importance sampler):

	Heavy-tailed		Rainfall		Rain (κ unkn.)	
	MSPE	CRPS	MSPE	CRPS	MSPE	CRPS
exact	0.175	0.096	0.452	0.146	0.450	0.146
GEnKF	0.195	0.107	0.470	0.154	0.895	0.232
EnKF	0.289	0.154	13.907	3.848	$>$100	$>$100
PF	0.764	0.582	2.033	1.037	4.289	1.861
Huber (Roh et al 2013)	0.629	0.405				

Details:

- Simulated 100 true states of size $n = 100$; $m = 75$ randomly chosen observations; $\mathbf{H}_t$ is subset of identity matrix
- True state distribution: mean 0.2, powered exponential covariance with power 1.8 and scale 10. $\sigma_{t,l} \equiv 0.2$
- EnKF and PF: $N = 100$. GEnKF: $N = 30$; 1 or 3 iterations
- Wendland taper with range 20

Smoothing for Lorenz-96

- Mimics advection at n = 40 equally-spaced locations along a latitude circle on the globe
- $T = 10$
- Gaussian observations ($\mathbf{z}_t = \mathbf{y}_t$)
- $\mathcal{M}_t(\mathbf{x}_{t-1}) = \theta \, \text{Lorenz}_{8,0.2}(\mathbf{x}_{t-1})$
- $\mathbf{H}_t = \mathbf{R}_t = \mathbf{I}_n, \quad \mathbf{Q}_t = 0.2\boldsymbol{\Sigma}_L$
- Prior: $\theta \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2)$, with $\mu_\theta = 0.8$ and $\sigma_\theta = 0.2$

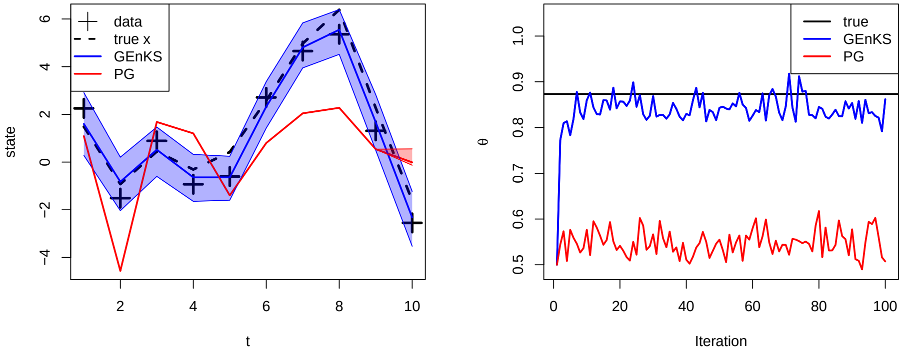
Goal: Find smoothing distribution $[\theta, \mathbf{x}_{1:T} | \mathbf{y}_{1:T}]$

Compared three methods on 100 simulated datasets:

- EnKS with $N = 1,000$ and state augmentation: $\theta_t | \theta_{t-1} \sim \mathcal{N}(\theta_{t-1}, 0.1^2)$
- Gibbs-EnKS
- Particle Gibbs sampler (Andrieu et al 2010)

For Gibbs-EnKS and particle-Gibbs: $N = 50$; 100 Gibbs iterations; $[\theta | \mathbf{x}_{1:T}, \mathbf{y}_{1:T}]$ is in closed form

Lorenz-96 results for one simulated dataset



(a) State at loc. 1 over time ($\mathbf{x}_{1:T,1}$)

(b) Trace plots for θ

Figure: True values and posterior distributions (posterior means and pointwise 80% prediction intervals) of the state at a single location (Panel (a)), and trace plots for θ (Panel (b))

Lorenz-96 results averaged over 100 datasets

	Parameter θ		State $\mathbf{x}_{1:T}$	
	MSPE	CRPS	MSPE	CRPS
Gibbs-EnKS	0.002	0.024	0.710	0.478
EnKS+SA	>100	>100	0.914	0.540
Particle Gibbs	0.111	0.262	12.380	2.495
Prior	0.042	0.118		

- EnKS with state augmentation diverges
- Particle Gibbs sampler produces worse inference on θ than simply using the prior distribution (i.e., completely ignoring any information in the data)

Outline

Summary

- EnKF can handle very large, nonlinear SSMs, but is less appropriate under non-Gaussianity or for unknown parameters
- Our extended EnKFs can handle more general, hierarchical SSMs, including unknown parameters and non-Gaussian observations
- In some cases, computational effort is similar to EnKF
- Results are approximate, but asymptotically correct for linear evolution

In general, extended EnKFs can only work well if the embedded EnKF for known parameters works well for the problem at hand

References

This talk is largely based on 2 papers:

- Katzfuss, M., Stroud, J.R., and Wikle, C.K. 2016. Understanding the ensemble Kalman filter. *The American Statistician*, 70(4), 350–357.
- Katzfuss, M., Stroud, J.R., and Wikle, C.K. 2017+. Extended ensemble Kalman filters for high-dimensional hierarchical state-space models. *arXiv:1704.06988*.

Funding

- Katzfuss: NSF DMS–1521676 and DMS–1654083
- Wikle: NSF SES-1132031