

Statistical Data Analysis

Dr. Jana de Wiljes

14. November 2021

Universität Potsdam

Bayesian statistics

Setting: parameter θ is a random variable and distributed according a known distribution

Bayes Theorem:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A) \cdot \mathbb{P}(A)}{\mathbb{P}(B)} \quad (1)$$

Goal: determine $\mathbb{P}(A|B)$

Example

Example

Example

A-posteriori-distribution

Def: The a-posteriori-distribution of θ is the conditional distribution given the information $X_1 = x_1, \dots, X_n = x_n$, i.e.,

$$q(\theta_i | x_1, \dots, x_n) := \mathbb{P}[\theta = \theta_i | X_1 = x_1, \dots, X_n = x_n], \quad i = 1, 2, \dots$$

A-posteriori-distribution

Bayes method

Def: The Bayes estimator is defined as the expectation of the a-posteriori-distribution

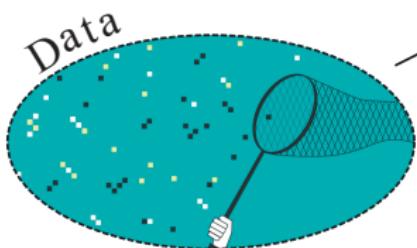
$$\hat{\theta}_{\text{Bayes}} = \sum_i \theta_i q(\theta_i | x_1, \dots, x_n)$$

What is Data Assimilation?

Data Assimilation

Model

$$\rho \dot{\vec{v}} = \rho \left(\frac{\partial \vec{v}}{\partial t} + (\vec{v} \cdot \nabla) \vec{v} \right)$$



Uncertainty?

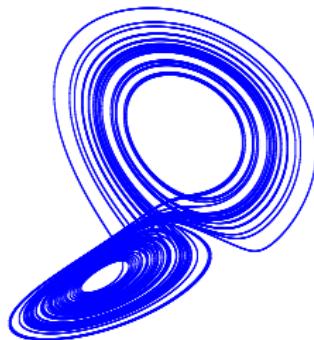
A toy atmospheric model

Lorenz equations:

$$\dot{x} = \sigma(y - x)$$

$$\dot{y} = x(\rho - z) - y$$

$$\dot{z} = xy - \beta z$$



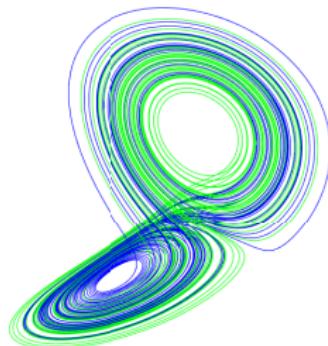
Uncertainty in initial conditions

Lorenz equations:

$$\dot{x} = \sigma(y - x)$$

$$\dot{y} = x(\rho - z) - y$$

$$\dot{z} = xy - \beta z$$



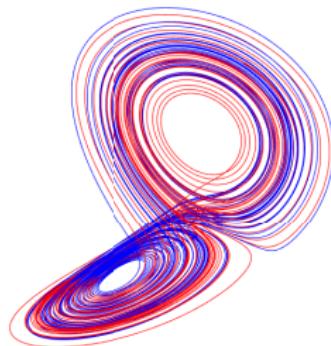
Uncertainty in parameters

Lorenz equations:

$$\dot{x} = \sigma(y - x)$$

$$\dot{y} = x(\rho - z) - y$$

$$\dot{z} = xy - \beta z$$



Numerical discretization and differentiation

Lorenz equations:

$$x_n = x_{n-1} + [\sigma(y_{n-1} - x_{n-1})]dt$$

$$y_n = y_{n-1} + [x_{n-1}(\rho - z_{n-1}) - x_{n-1}]dt$$

$$z_n = z_{n-1} + [x_{n-1}y_{n-1} - \beta z_{n-1}]dt$$

The Model

Model (deterministic)

Evolution equation

$$\textcolor{blue}{z}_n = \Psi(\textcolor{blue}{z}_{n-1}, \textcolor{red}{\lambda})$$

where

$$z_0 \sim \mathcal{N}(\textcolor{blue}{m}_0, \textcolor{blue}{C}_0)$$

Model

Evolution equation

$$\textcolor{blue}{z}_n = \Psi(\textcolor{blue}{z}_{n-1}, \lambda) + \xi_{n-1}$$

where

$$\textcolor{blue}{z}_0 \sim \mathcal{N}(\textcolor{blue}{m}_0, \textcolor{blue}{C}_0)$$

$$\xi_n \sim \mathcal{N}(0, B) \quad \text{i.i.d.} \quad \forall n$$

Parameter estimation

Augmented state space

$$\textcolor{blue}{z}_n = \Psi(\textcolor{blue}{z}_{n-1}, \lambda_{n-1}) + \xi_{n-1}$$

$$\lambda_n = \lambda_{n-1}$$

where

$$[\textcolor{blue}{z}_0, \lambda_0]^\top \sim \mathcal{N}(\textcolor{blue}{m}_0, \textcolor{blue}{C}_0)$$

$$\xi_n \sim \mathcal{N}(0, B) \quad \text{i.i.d.} \quad \forall n$$

Observations

Observations

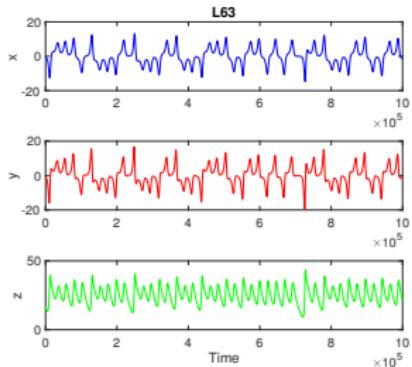
Partial and noisy data:

$$\textcolor{blue}{y}_n = h(\textcolor{blue}{z}_n) + \textcolor{red}{\eta}_n$$

where

$$\textcolor{red}{\eta}_n \sim \mathcal{N}(0, R) \quad \text{i.i.d.} \quad \forall n$$

Example



Lorenz equations:

$$x_n = x_{n-1} + [\sigma(y_{n-1} - x_{n-1})]dt$$

$$y_n = y_{n-1} + [x_{n-1}(\rho - z_{n-1}) - x_{n-1}]dt$$

$$z_n = z_{n-1} + [x_{n-1}y_{n-1} - \beta z_{n-1}]dt$$

The Math behind it...

Conditional probability

Definition (Conditional probability)

For a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and events $A, B \in \mathcal{F}$ the conditional probability of B given A is

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A, B)}{\mathbb{P}(B)}.$$

Bayes theorem

Theorem (Bayes)

For a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$ the following holds for two events A and B in \mathcal{F}

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

Bayesian data assimilation ansatz

$$\mathbb{P}(\text{Model}|\text{Obs}) = \frac{\mathbb{P}(\text{Obs}|\text{Model})\mathbb{P}(\text{Model})}{\mathbb{P}(\text{Obs})}$$

Bayesian data assimilation ansatz

$$\mathbb{P}(\text{Model}|\text{Obs}) \propto \mathbb{P}(\text{Obs}|\text{Model})\mathbb{P}(\text{Model})$$

Bayesian data assimilation for densities

$$\begin{aligned}\pi(\mathbf{z}_{n+1} | \mathbf{y}_{1:n+1}) &= \pi(\mathbf{z}_{n+1} | \mathbf{y}_{1:n}, \mathbf{y}_{n+1}) \\&= \frac{\pi(\mathbf{y}_{n+1} | \mathbf{y}_{1:n}, \mathbf{z}_{n+1})\pi(\mathbf{z}_{n+1} | \mathbf{y}_{1:n})}{\pi(\mathbf{y}_{n+1} | \mathbf{y}_{1:n})} \\&= \frac{\pi(\mathbf{y}_{n+1} | \mathbf{z}_{n+1})\pi(\mathbf{z}_{n+1} | \mathbf{y}_{1:n})}{\pi(\mathbf{y}_{n+1} | \mathbf{y}_{1:n})} \\ \implies \pi(\mathbf{z}_{n+1} | \mathbf{y}_{1:n+1}) &\propto \pi(\mathbf{y}_{n+1} | \mathbf{z}_{n+1})\pi(\mathbf{z}_{n+1} | \mathbf{y}_{1:n})\end{aligned}\tag{2}$$

Special case

Linear model: Ψ is linear, e.g.,

$$\textcolor{blue}{z}_n = A\textcolor{blue}{z}_{n-1} + \xi_{n-1} \quad (3)$$

with $A \in \mathbb{R}^{N_z} \times \mathbb{R}^{N_z}$

Linear observation operator

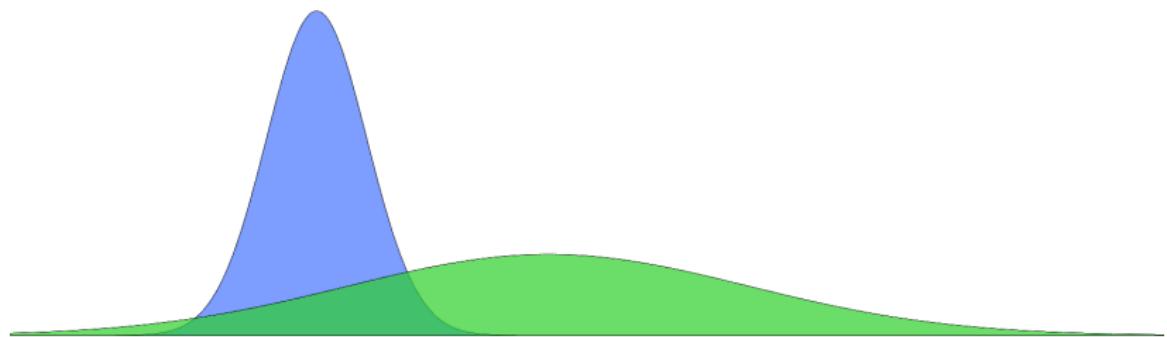
$$h = H \quad \text{with} \quad H \in \mathbb{R}^{N_y} \times \mathbb{R}^{N_y}$$

Linear model



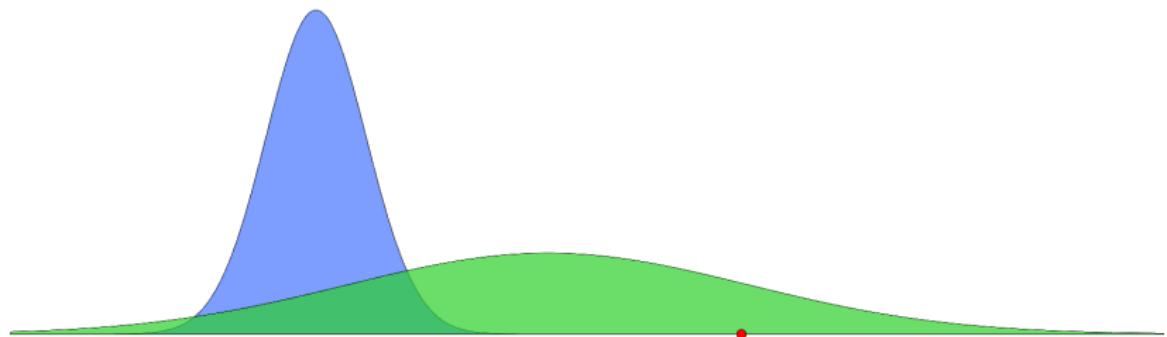
Initial distribution: $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{m}_0, \mathbf{C}_0)$

Linear model



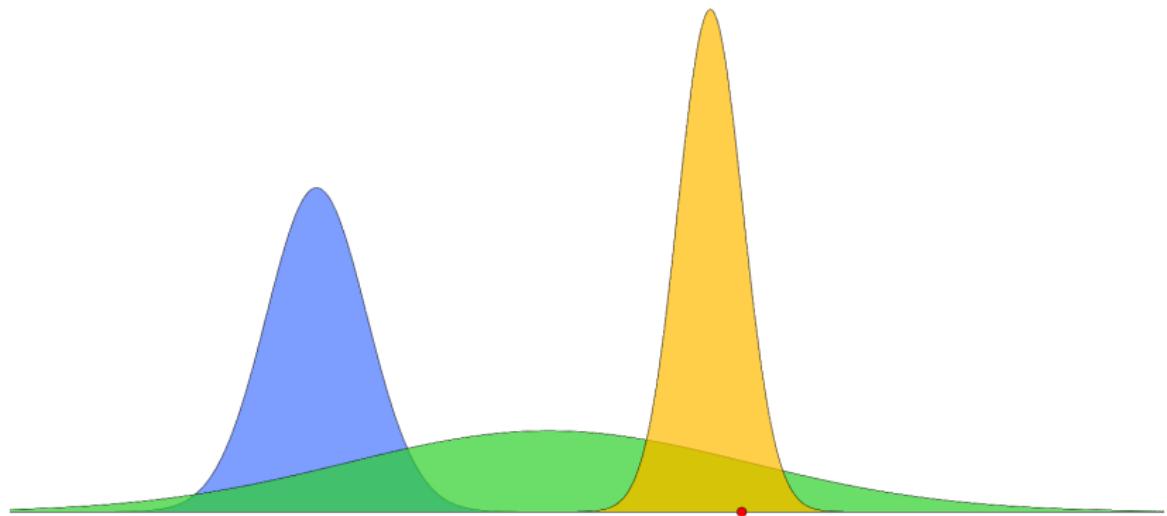
Prior distribution: $\mathcal{N}(\hat{m}_1, \hat{C}_1)$

Linear model



Likelihood: $\mathcal{N}(H\hat{\mathbf{m}}_1, R)$

Linear model



Posterior: $\mathcal{N}(\mathbf{m}_1, \mathbf{C}_1) \propto \mathcal{N}(H\hat{\mathbf{m}}_1, R)\mathcal{N}(\hat{\mathbf{m}}_1, \hat{\mathbf{C}}_1)$

Algorithms

Kalman filter

Two steps:

Kalman filter

Two steps:

- Forecast: $(m_n, C_n) \mapsto (\hat{m}_{n+1}, \hat{C}_{n+1})$

Kalman filter

Two steps:

- Forecast: $(m_n, C_n) \mapsto (\hat{m}_{n+1}, \hat{C}_{n+1})$
- Analysis: $(\hat{m}_{n+1}, \hat{C}_{n+1}) \mapsto (m_{n+1}, C_{n+1})$

Two steps:

- Forecast: $(\mathbf{m}_n, \mathbf{C}_n) \mapsto (\hat{\mathbf{m}}_{n+1}, \hat{\mathbf{C}}_{n+1})$
- Analysis: $(\hat{\mathbf{m}}_{n+1}, \hat{\mathbf{C}}_{n+1}) \mapsto (\mathbf{m}_{n+1}, \mathbf{C}_{n+1})$

Forecast formulas

$$\hat{\mathbf{m}}_{n+1} = A\mathbf{m}_n$$

$$\hat{\mathbf{C}}_{n+1} = A\mathbf{C}_n A^\top + B$$

Kalman filter

Two steps:

- Forecast: $(\mathbf{m}_n, \mathbf{C}_n) \mapsto (\hat{\mathbf{m}}_{n+1}, \hat{\mathbf{C}}_{n+1})$
- Analysis: $(\hat{\mathbf{m}}_{n+1}, \hat{\mathbf{C}}_{n+1}) \mapsto (\mathbf{m}_{n+1}, \mathbf{C}_{n+1})$

Forecast formulas

$$\hat{\mathbf{m}}_{n+1} = A\mathbf{m}_n$$

$$\hat{\mathbf{C}}_{n+1} = A\mathbf{C}_n A^\top + B$$

Analysis formulas

$$\mathbf{m}_{n+1} = \hat{\mathbf{m}}_{n+1} - K_{n+1}(H\hat{\mathbf{m}}_{n+1} - \mathbf{y}_{n+1})$$

$$\mathbf{C}_{n+1} = \hat{\mathbf{C}}_{n+1} - K_{n+1}H\hat{\mathbf{C}}_{n+1}$$

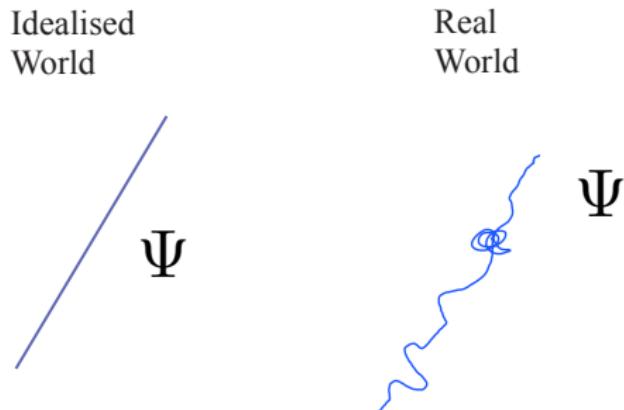
Kalman gain

$$K_{n+1} = \hat{\mathbf{C}}_{n+1} H^\top (R + H\hat{\mathbf{C}}_{n+1} H^\top)^{-1}$$

Kalman filter



Reality check



Problem: Kalman Filter is not applicable anymore

Approach: use approximative Algorithms, e.g.,

- Extended Kalman Filter: linearize model function

Monte Carlo approximation



Approach: approximative via empirical measure

$$\pi(\mathbf{z}_n | \mathbf{y}_{1:n}) = \frac{1}{M} \sum_{i=1}^M \delta(z - \mathbf{z}_n^i)$$

where

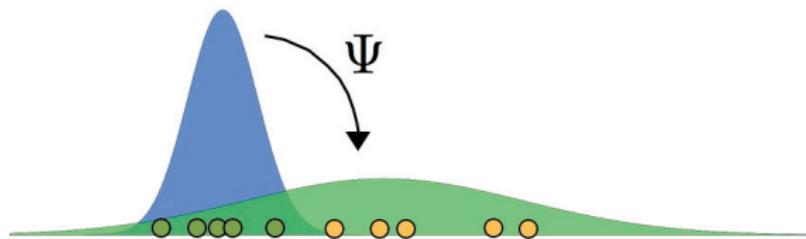
$$\mathbf{z}_n^i \sim \pi(\mathbf{z}_n | \mathbf{y}_n)$$

Ensemble Kalman filter



$$\mathcal{N}(\textcolor{blue}{m}_0, \textcolor{blue}{C}_0) \text{ with } \textcolor{blue}{m}_0 \approx \frac{1}{M} \sum_{i=1}^M \textcolor{blue}{z}_0^i$$
$$\textcolor{blue}{C}_0 \approx \frac{1}{M} \sum_{i=1}^M (\textcolor{blue}{z}_0^i - \textcolor{blue}{m}_0)(\textcolor{blue}{z}_0^i - \textcolor{blue}{m}_0)^\top$$

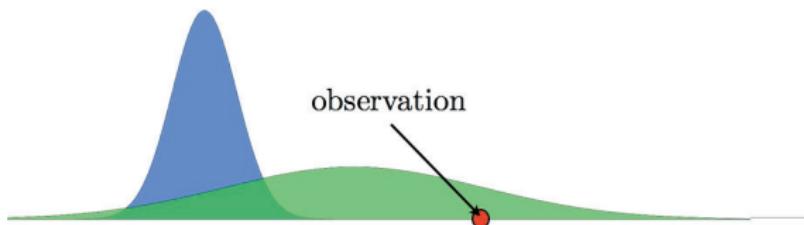
Ensemble Kalman filter



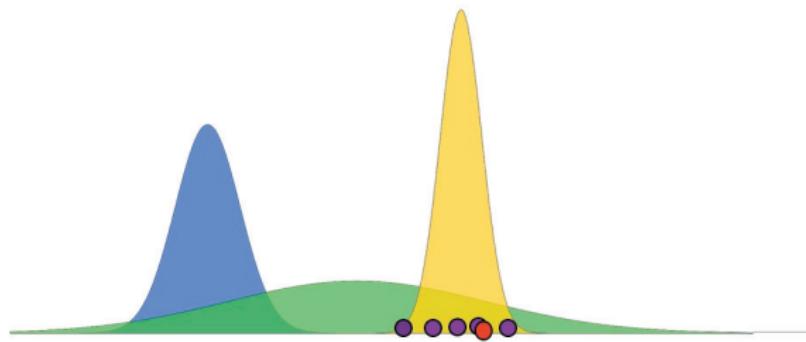
$$\mathcal{N}(\hat{\mathbf{m}}_1, \hat{\mathbf{C}}_1) \text{ with } \hat{\mathbf{m}}_1 \approx \frac{1}{M} \sum_{i=1}^M \hat{\mathbf{z}}_1^i = \frac{1}{M} \sum_{i=1}^M \Psi(\mathbf{z}_0^i)$$

$$\hat{\mathbf{C}}_1 \approx \frac{1}{M} \sum_{i=1}^M (\hat{\mathbf{z}}_1^i - \hat{\mathbf{m}}_1)(\hat{\mathbf{z}}_1^i - \hat{\mathbf{m}}_1)^\top$$

Ensemble Kalman filter



Ensemble Kalman filter



$$\mathcal{N}(\textcolor{blue}{m}_1, \textcolor{blue}{C}_1) \text{ with } \textcolor{blue}{m}_1 \approx \frac{1}{M} \sum_{i=1}^M \textcolor{blue}{z}_1^i$$

$$\textcolor{blue}{C}_1 \approx \frac{1}{M} \sum_{i=1}^M (\textcolor{blue}{z}_1^i - \textcolor{blue}{m}_1)(\textcolor{blue}{z}_1^i - \textcolor{blue}{m}_1)^\top$$

Ensemble Kalman filter

Goal: approximate $\pi(\textcolor{blue}{z}_n | \textcolor{blue}{y}_{1:n})$

Ensemble Kalman filter

Goal: approximate $\pi(\textcolor{blue}{z}_n | \textcolor{blue}{y}_{1:n})$

Approach: propagate samples $\hat{\textcolor{blue}{z}}_{n+1}^i$ with Kalman formula

$$\textcolor{blue}{z}_{n+1}^i = \hat{\textcolor{blue}{z}}_{n+1}^i - \textcolor{blue}{K}_{n+1}(H\hat{\textcolor{blue}{z}}_{n+1}^i - \tilde{\textcolor{red}{y}}_{n+1}^i)$$

Ensemble Kalman filter

Goal: approximate $\pi(\mathbf{z}_n | \mathbf{y}_{1:n})$

Approach: propagate samples $\hat{\mathbf{z}}_{n+1}^i$ with Kalman formula

$$\mathbf{z}_{n+1}^i = \hat{\mathbf{z}}_{n+1}^i - \mathcal{K}_{n+1}(H\hat{\mathbf{z}}_{n+1}^i - \tilde{\mathbf{y}}_{n+1}^i)$$

Need: perturbed observations

$$\tilde{\mathbf{y}}_{n+1}^i = \mathbf{y}_{n+1} + \epsilon_{n+1}^i$$

with $\epsilon_{n+1}^i \sim \mathcal{N}(0, R)$ i.i.d. to get the correct mean and covariance
in the linear case for $M \rightarrow \infty$

Ensemble Kalman filter

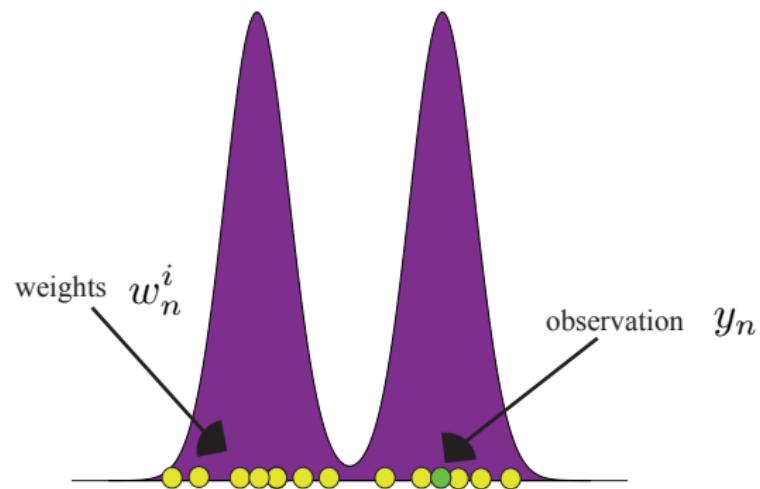
Works well in practice: e.g., EnKF is used for operational NWP
for \mathbf{z}_n^i with dimension 10^8 only using $M = 100$

Yet: mathematical foundation largely missing

Recent study: accuracy results for EnKF for idealized setting:
 $H = Id$ and observational error small

Particle filter

Particle Filter



Particle filter

Problem: sampling from $\pi(\mathbf{z}_n | \mathbf{y}_{1:n})$ to approximate posterior via

$$\pi(\mathbf{z}_n | \mathbf{y}_{1:n}) = \frac{1}{M} \sum_{i=1}^M \delta(z - \mathbf{z}_n^i)$$

Idea: sampling from $\pi(\mathbf{z}_n | \mathbf{y}_{1:n-1})$ instead i.e.,

$$\pi(\mathbf{z}_n | \mathbf{y}_{1:n}) = \sum_{i=1}^M w_n^i \delta(z - \hat{\mathbf{z}}_n^i)$$

Bayes:

$$\pi(\mathbf{z}_{n+1} | \mathbf{y}_{1:n}) \propto \pi(\mathbf{y}_n | \mathbf{z}_n) \pi(\mathbf{z}_n | \mathbf{y}_{1:n-1}) \quad (4)$$

Weighting: unnormalized weights

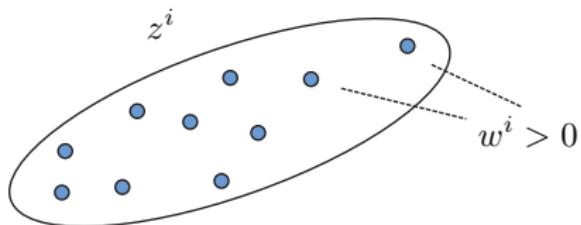
$$\tilde{w}_n^i = \pi(\mathbf{y}_n | \mathbf{z}_n^i) w_{n-1}^i \text{ with } w_0^i = \frac{1}{M}$$

and normalized weights

$$w_n^i = \frac{\tilde{w}_n^i}{\sum_{j=1}^M \tilde{w}_n^j}$$

Particle collapse

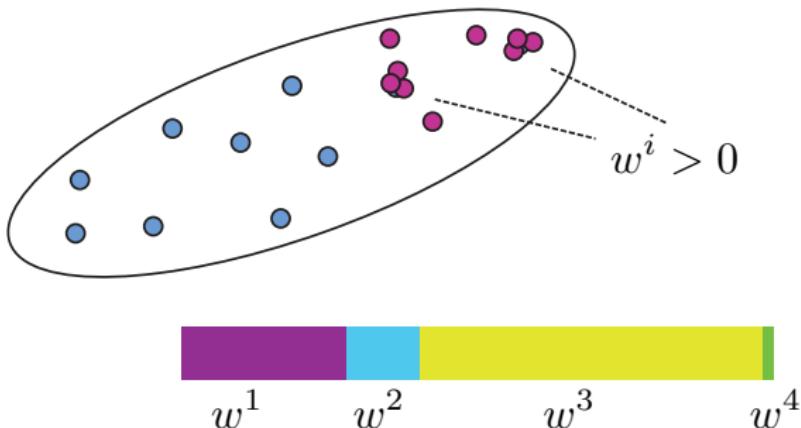
 y
Degeneracy
of Particle Filter



Resampling

• y

Resampling



Resampling

Problem: weights w_n^i become very small

Ansatz: resampling

Input: w_n^i

For($k = 1 : M$)

1. Draw a number $u \in [0, 1]$ from the uniform distribution $U[0, 1]$
2. Compute $i^* \in \{1, \dots, M\}$ which satisfies

$$i^* = \arg \min_{i \geq 1} \sum_{j=1}^i w_j \geq u \quad (5)$$

3. Set $\xi_{i^*} = \xi_{i^*} + 1$

Return ξ_i

Still a lot of challenges....

Ansatz: approximative via empirical measure

$$\pi(\mathbf{z}_n | \mathbf{y}_{1:n}) = \frac{1}{M} \sum_{i=1}^M \delta(z - \mathbf{z}_n^i)$$

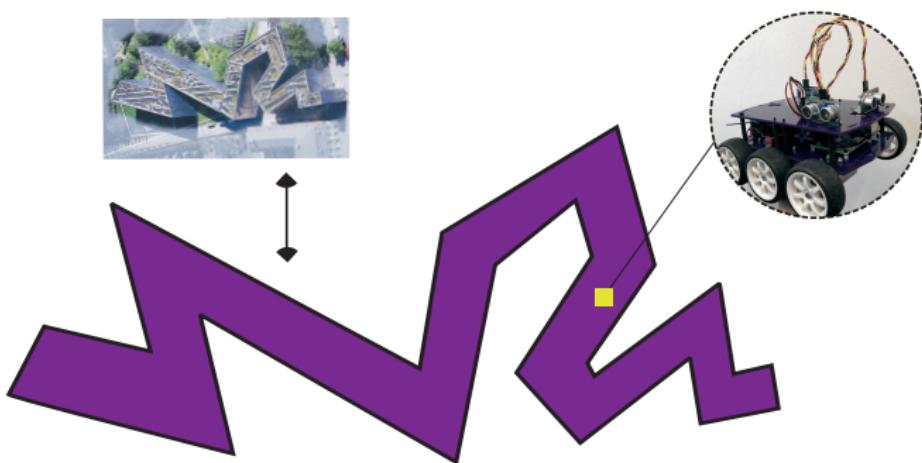
where

$$\mathbf{z}_n^i \sim \pi(\mathbf{z}_n | \mathbf{y}_n)$$

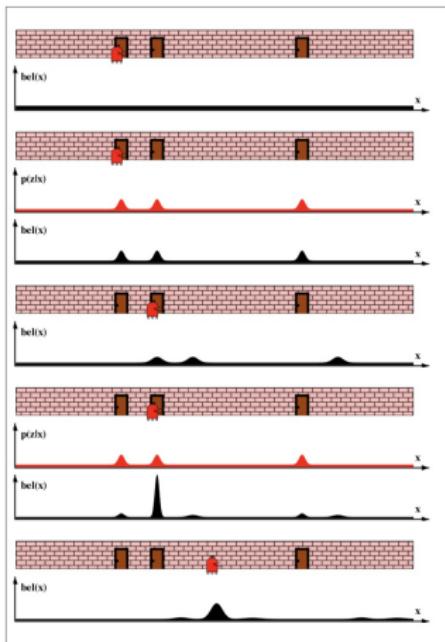
Monte Carlo approximation leads to a variety of filters e.g.,

- Particle filters (**curse of dimensionality**)
- Ensemble Kalman filter (**underlying Gaussian assumption**)

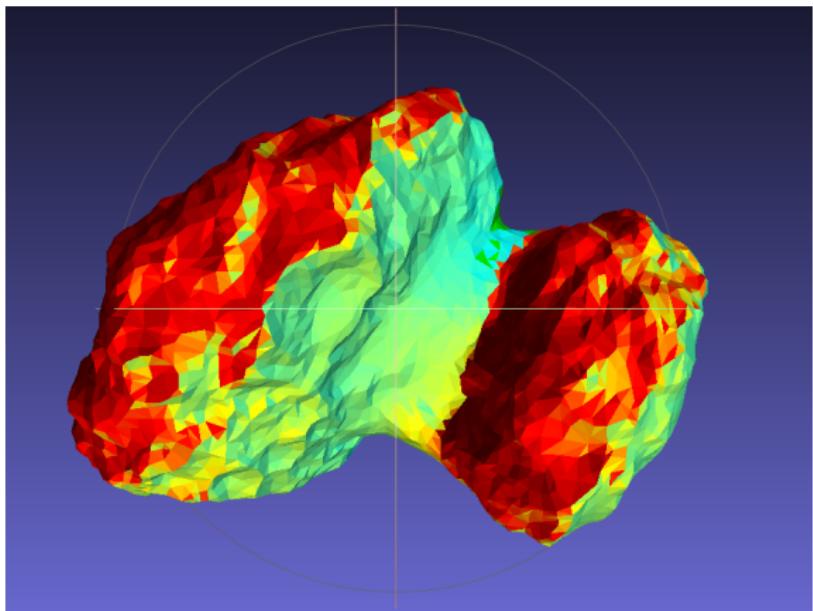
Same Math different Applications



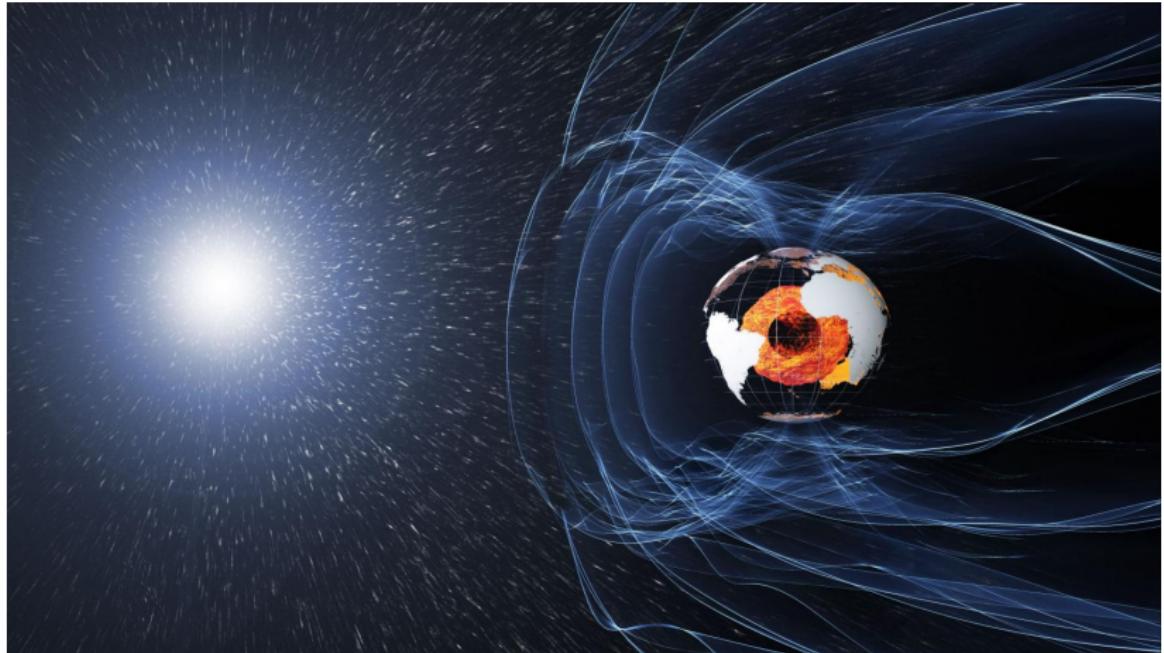
Simultaneous state and parameter estimation



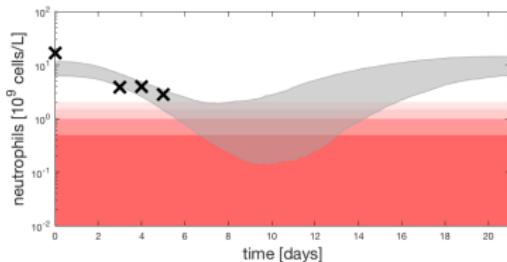
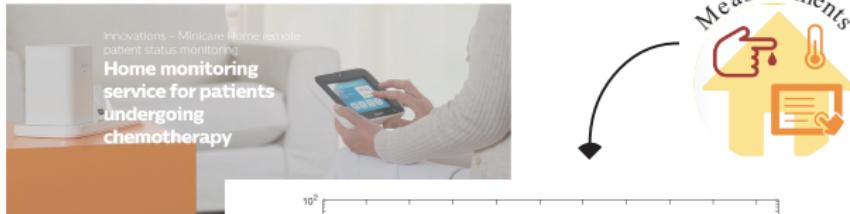
Thermophysical modeling



Space weather



Pharmacology

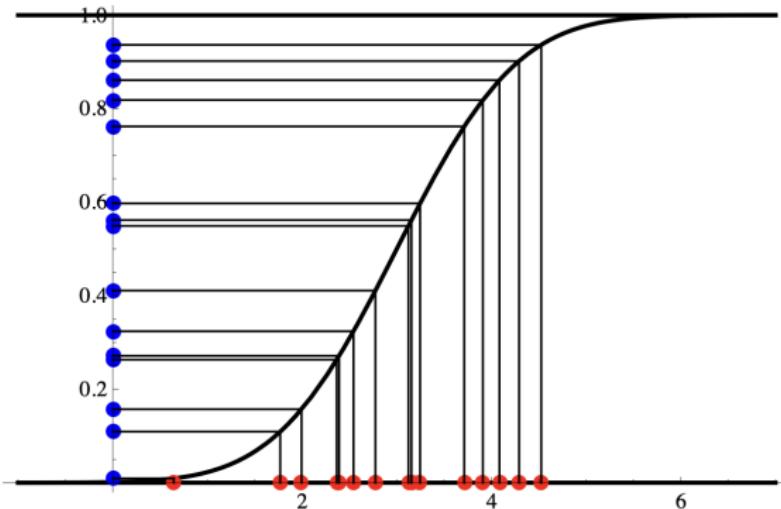


Personalised medicine

Maximum-spacing method

Maximum-spacing method

Lemma: Let the cdf F_θ be continuous and strictly monoton increasing. Under \mathbb{P}_θ the random variables $F_\theta(X_1), \dots, F_\theta(X_n)$ are independent and uniformly distributed on the $(0, 1)$ interval.



Proof

Proof

Maximum-spacing method

Lemma: Let $z_1, \dots, z_k \in [0, 1]$ be numbers that are subject to the condition $z_1 + \dots + z_k = 1$. Then

$$z_1 \cdot \dots \cdot z_k \leq \frac{1}{k^k}. \quad (6)$$

Equality is attained only if all the numbers are equal to $\frac{1}{k}$.

Example

Maximum-spacing method

Lemma: The maximum-spacing method is defined via

$$\widehat{\theta}_{MS} = \arg \max_{\theta \in \Theta} \prod_{i=1}^{n+1} (F_\theta(x_{(i)}) - F_\theta(x_{(i-1)})) \quad (7)$$

Example

Example
