

Statistical Data Analysis

Dr. Jana de Wiljes

16. November 2021

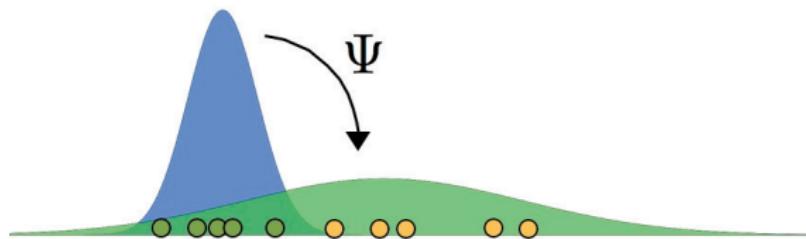
Universität Potsdam

Ensemble Kalman filter



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

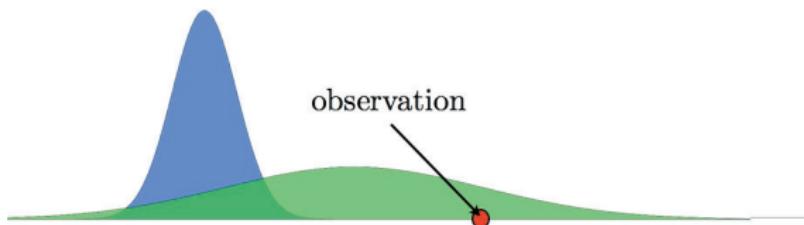
Ensemble Kalman filter



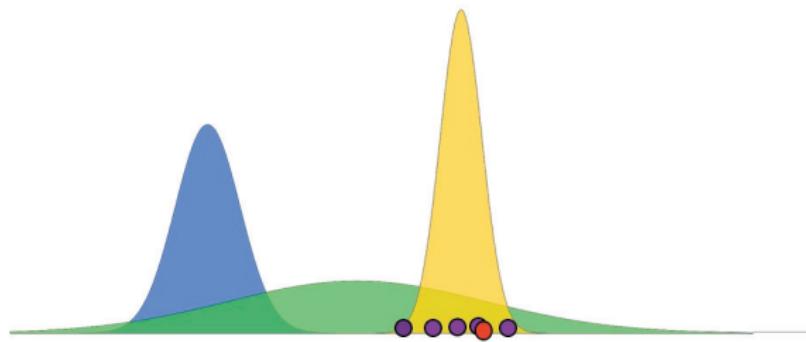
$$\mathcal{N}(\hat{\mathbf{m}}_1, \hat{\mathbf{C}}_1) \text{ with } \hat{\mathbf{m}}_1 \approx \frac{1}{M} \sum_{i=1}^M \hat{\mathbf{z}}_1^i = \frac{1}{M} \sum_{i=1}^M \Psi(\mathbf{z}_0^i)$$

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

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

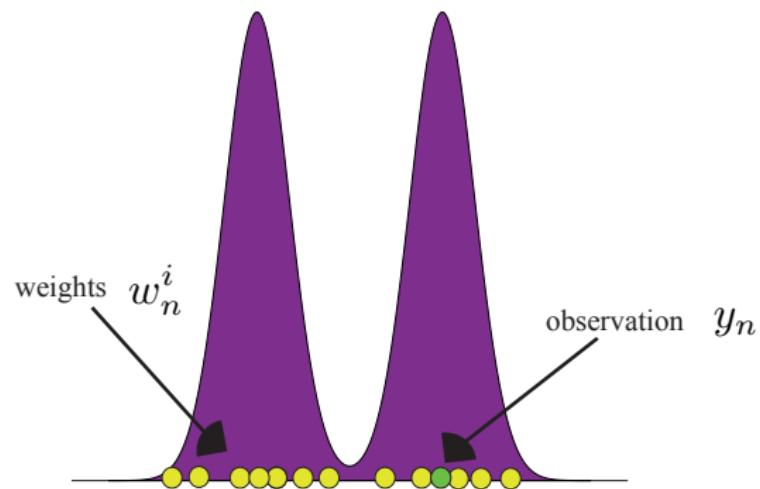
Works well in practice: e.g., EnKF is used for operational NWP
for \mathbf{z}_n^i with dimension 10^8 only using $M = 100$

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:

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

Weighting: unnormalized weights

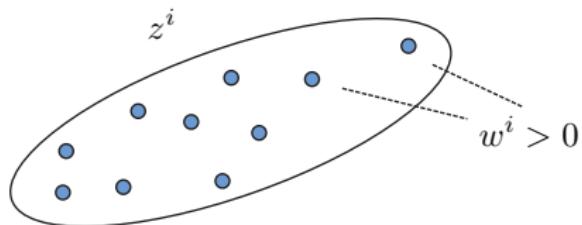
$$\tilde{w}_n^i = \pi(\mathbf{y}_n | \mathbf{z}_n^i) w_{n-1}^i \text{ with } w_0^i = \frac{1}{M}$$

and normalized weights

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

Particle collapse

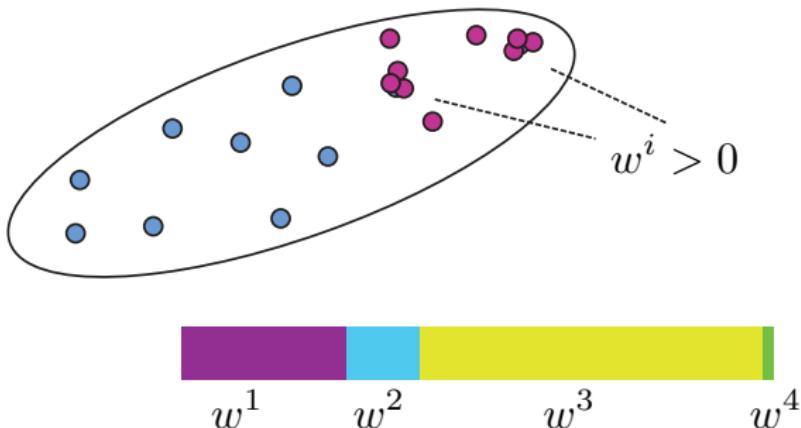
y
Degeneracy
of Particle Filter



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

$$i^* = \arg \min_{i \geq 1} \sum_{j=1}^i w_j \geq u \quad (2)$$

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

Return ξ_i

Still a lot of challenges....

Ansatz: approximative via empirical measure

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

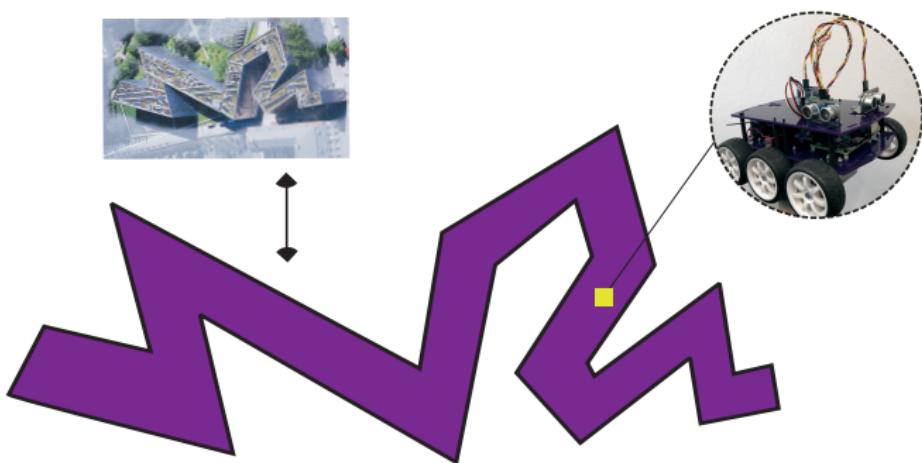
where

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

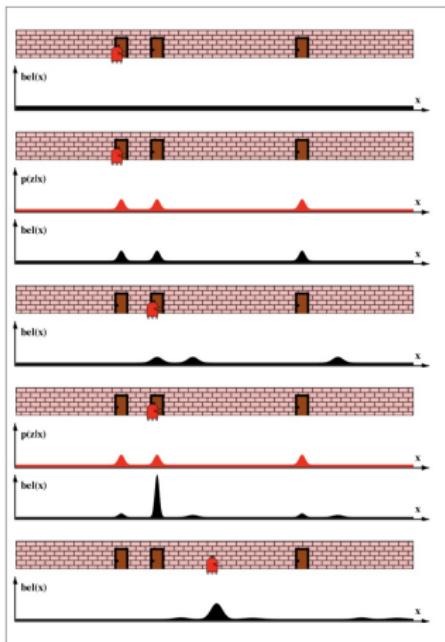
Monte Carlo approximation leads to a variety of filters e.g.,

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

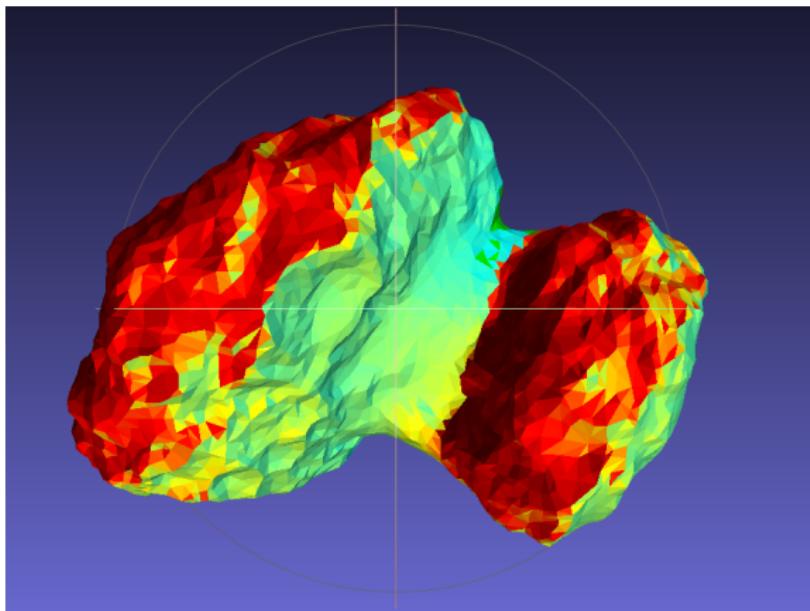
Same Math different Applications



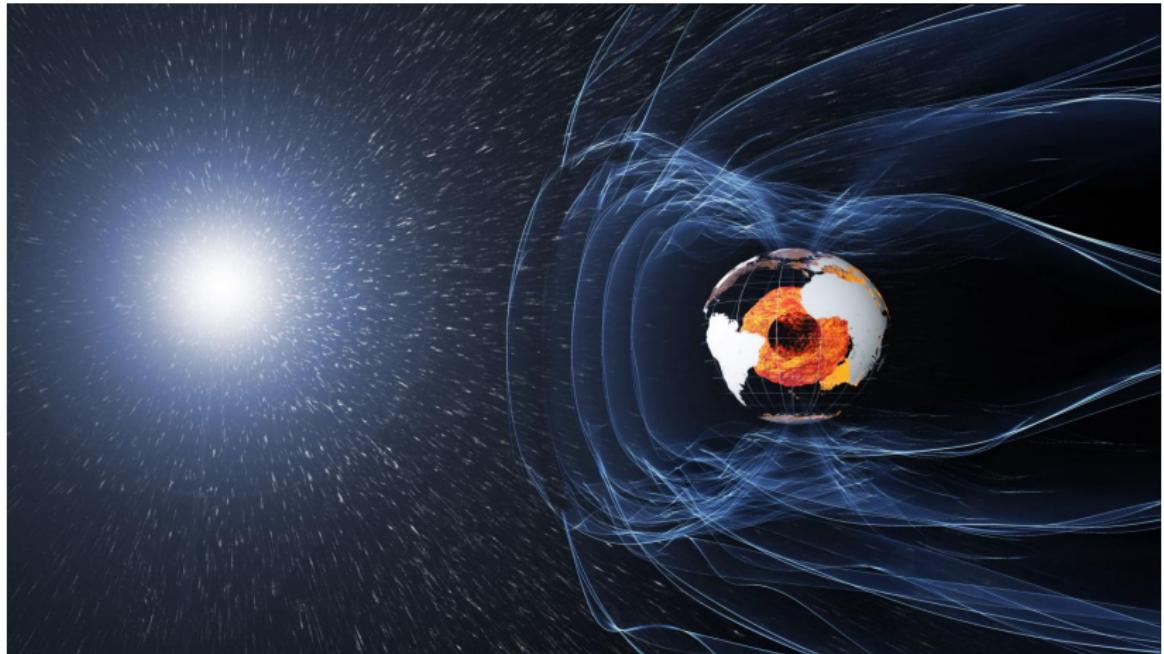
Simultaneous state and parameter estimation



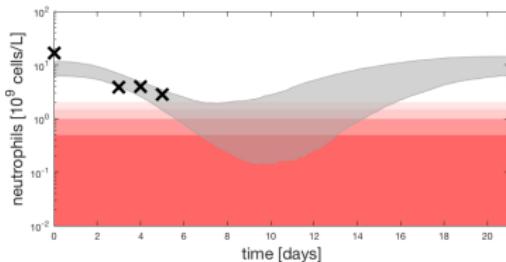
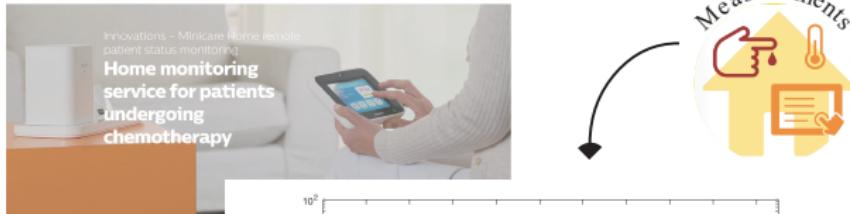
Thermophysical modeling



Space weather



Pharmacology

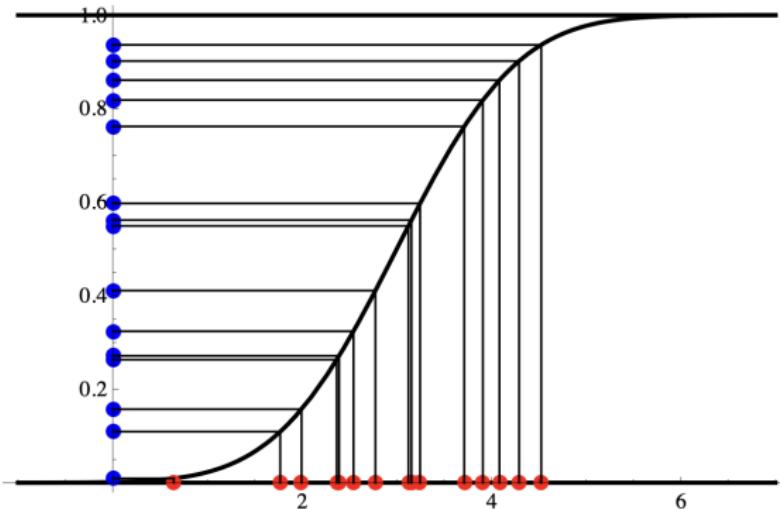


Personalised medicine

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

$$z_1 \cdot \dots \cdot z_k \leq \frac{1}{k^k}. \quad (3)$$

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

$$\hat{\theta}_{MS} = \arg \max_{\theta \in \Theta} \prod_{i=1}^{n+1} (F_\theta(x_{(i)}) - F_\theta(x_{(i-1)})) \quad (4)$$

Example

Example

Unbiased estimators

Def:

- An estimator is an arbitrary (Borel-measurable) function

$$\hat{\theta} : \mathcal{X} \rightarrow \Theta, \quad x \mapsto \hat{\theta}(x) \quad (5)$$

- An estimator $\hat{\theta}$ is called unbiased, if

$$\mathbb{E}_{\theta}[\hat{\theta}(X)] = \theta \quad (6)$$

for all $\theta \in \Theta$.

- The bias of an estimator $\hat{\theta}$ is

$$\text{Bias}_{\theta}(\hat{\theta}) = \mathbb{E}_{\theta}[\hat{\theta}(X)] - \theta \quad (7)$$

Note: $\text{Bias}_{\theta}(\hat{\theta})$ is a function in $\hat{\theta}$

Example

Example

Example

Mean square error

Def: Let $\Theta = (a, b) \subset \mathbb{R}$ be an interval. The mean square error (MSE) of an estimator $\hat{\theta} : \mathcal{X} \rightarrow \Theta$

$$\text{MSE}_\theta(\hat{\theta}) = \mathbb{E}_\theta[(\hat{\theta}(X) - \theta)^2] \quad (8)$$

Mean square error

Lemma: The relationship between the mean square error (MSE) of an estimator $\hat{\theta} : \mathcal{X} \rightarrow \Theta$ and the BIAS is given by

$$\text{MSE}_\theta(\hat{\theta}) = \text{Var}_\theta \hat{\theta} + (\text{Bias}_\theta(\hat{\theta}))^2 \quad (9)$$

Proof

Proof

Consistently better

Let $\hat{\Theta}_1$ and $\hat{\Theta}_2$ be two estimators. The estimator θ_1 is consistently better than θ_2 if,

$$MSE_{\theta}(\hat{\theta}_1) \leq MSE_{\theta}(\hat{\theta}_2) \quad \forall \theta \in \Theta \quad (10)$$

Minimum-variance unbiased estimator

Def: An unbiased estimator $\hat{\theta}$ is called minimum-variance unbiased estimator if all unbiased estimators $\tilde{\theta}$ the following inequality holds

$$\text{Var}_\theta \hat{\theta} \leq \text{Var}_\theta \tilde{\theta} \quad (11)$$

for all $\theta \in \Theta$.

Minimum-variance unbiased estimator

Lemma: Let $\hat{\theta}_1, \hat{\theta}_2 : \mathcal{X} \rightarrow \Theta$ are two minimum-variance unbiased estimator the

$$\hat{\theta}_1 = \hat{\theta}_2 \quad \text{almost surely under } \mathbb{P} \text{ for all } \theta \in \Theta \quad (12)$$

for all $\theta \in \Theta$.

Proof

Proof

Proof

Sufficient statistic

Def: A function $T : \mathcal{X} \rightarrow \mathbb{R}^r$ is called a sufficient statistic if the function

$$\theta \mapsto \mathbb{P}_\theta[X = x | T(X) = t] \quad (13)$$

is constant for all $x \in \mathcal{X}$ and for all $t \in \mathbb{R}^r$, i.e.,

$$\mathbb{P}_{\theta_1}[X = x | T(X) = t] \mathbb{P}_{\theta_2}[X = x | T(X) = t] \quad (14)$$

for all $t \in \mathbb{R}^r$ and all $\theta_1, \theta_2 \in \Theta$ with $\mathbb{P}_{\theta_1}[T(X) = t] \neq 0$ and
 $\mathbb{P}_{\theta_2}[T(X) = t] \neq 0$