Bayesian Inference and Data Assimilation

Prof. Dr.-Ing. Sebastian Reich

Universität Potsdam

12 April 2021

6 Basic data assimilation algorithms

In this section, we return to the state **estimation problem for dynamical systems** as initially raised in the Prolog.

However, in contrast to the Prolog, data assimilation algorithms will be based on a probabilistic interpretation of model-based forecasts in addition to measurement processes.

Hence **state estimation from partial and noisy observations** can be put into the framework of Bayesian inference with model-based forecast uncertainties taking the role of prior distributions.

Most of the mathematical ingredients have already been provided in the previous chapters.

In this chapter, we define **sequential (online) Bayesian inference**, explain what it is used for and introduce some mathematical tools for applying it.

Definition (Model)

Stochastic difference equation

$$Z^{n+1} = Z^n + \delta t f(Z^n) + \sqrt{2\delta t}\,\Xi^n, \qquad t_{n+1} = t_n + \delta t, \tag{1}$$

where $\Xi^n : \Omega \to \mathbb{R}^{N_z}$ are i.i.d. Gaussian RVs with mean zero and covariance matrix Q, $\delta t > 0$ is a given step-size, and $f : \mathbb{R}^{N_z} \to \mathbb{R}^{N_z}$ is a given vector field.

We assume that the initial distribution $\pi_Z(z, 0)$ for Z^0 is given.

The time evolution of the marginal PDFs $\pi_Z(z, t_n)$ $(n \geq 1)$ under (1) is given recursively by the **Chapman-Kolmogorov equation**

$$\pi_Z(z', t_{n+1}) = \int_{\mathbb{R}^{N_z}} \pi(z'|z)\, \pi_Z(z, t_n)\, \mathrm{d}z,$$

with transition kernel (conditional PDF)

$$\pi(z'|z) = \frac{1}{(4\pi\delta t)^{N_z/2}|Q|^{1/2}} \times \exp\left(-\frac{1}{4\delta t}(z' - z - \delta t f(z))^{\mathrm{t}} Q^{-1} (z' - z - \delta t f(z)) \right). \tag{2}$$

Definition (Data)

Noisy partial observations

$$y_{\rm obs}(t_k) \in \mathbb{R}^{N_y}, \quad t_k = k\Delta t_{\rm out}, \quad k = 1, 2, \ldots,$$

become available in regular time intervals of width $\Delta t_{\rm out} > 0$.

We assume that these observations are related to the unknown **reference trajectory** $z_{\rm ref}(t)$ of the dynamical model (1) by

$$y_{\rm obs}(t_k) = h(z_{\rm ref}(t_k)) + \Sigma^k, \tag{3}$$

where Σ^k denotes the realisation of a Gaussian random variable with mean zero and covariance matrix R and $h : \mathbb{R}^{N_z} \to \mathbb{R}^{N_y}$ is the **forward operator**.

The **conditional PDF (likelihood)** for $y_{\rm obs}(t_k)$ given a state z is therefore given by

$$\pi_Y(y|z) = \frac{1}{(2\pi)^{N_y/2}|R|^{1/2}} \exp\left(-\frac{1}{2}(y - h(z))^T R^{-1}(y - h(z))\right). \tag{4}$$

Definition (Prediction, filtering, and smoothing)

In the case where there are no observations available, our knowledge about a process $z_{\rm ref}(t)$ is encoded in the marginal PDFs $\pi_Z(z,t_n)$ for Z^n.

When we are given observations $y_{\rm obs}(\mathrm{t}_k)$, $k=1,\ldots,N_{\rm obs}$, collected in an $N_y \times N_{\rm obs}$ matrix

$$y^{\rm obs}_{\mathrm{t}_{1:N_{\rm obs}}} = (y_{\rm obs}(\mathrm{t}_1), y_{\rm obs}(\mathrm{t}_2), \ldots, y_{\rm obs}(\mathrm{t}_{N_{\rm obs}})) \in \mathbb{R}^{N_y \times N_{\rm obs}},$$

our knowledge is encoded in the conditional, marginal PDFs $\pi_Z(z,t_n|y^{\rm obs}_{\mathrm{t}_{1:N_{\rm obs}}})$ for $Z^n|y^{\rm obs}_{\mathrm{t}_{1:N_{\rm obs}}}$.

The task of calculating these PDFs takes one of the following three names:

1. **prediction** if $t_n > \mathrm{t}_{N_{\rm obs}}$,
2. **filtering** if $t_n = \mathrm{t}_{N_{\rm obs}}$, and
3. **smoothing** if $t_n < \mathrm{t}_{N_{\rm obs}}$.

We continue with a closer inspection of the **filtering problem** and derive recursive representations of the conditional, marginal PDFs $\pi_Z(z, \mathrm{t}_k | y^{\mathrm{obs}}_{\mathrm{t}_{1:k}})$ for $k = 1, \ldots, N_{\mathrm{obs}}$.

Let us assume, for simplicity, that

$$\Delta t_{\mathrm{out}} = N_{\mathrm{out}} \delta t,$$

in which case the N_{out}-fold application of the **Chapman-Kolmogorov equation** leads from the initial PDF $\pi_Z(z, 0)$ at $t = 0$ to

$$\pi_Z(z', \mathrm{t}_1) = \int_{\mathbb{R}^{N_z}} \pi_{N_{\mathrm{out}}}(z'|z) \pi_Z(z, 0) \mathrm{d}z,$$

with the transition kernel $\pi_{N_{\mathrm{out}}}(z'|z)$ appropriately defined.

For example, $N_{\mathrm{out}} = 2$:

$$\pi_2(z'|z) = \int_{\mathbb{R}^{N_z}} \pi(z'|z'') \pi(z''|z) \, \mathrm{d}z'' \, .$$

Filtering involves the following **recursion**.

(i) Given the PDF $\pi_Z(z, \mathrm{t}_{k-1}|y^{\mathrm{obs}}_{\mathrm{t}_{1:k-1}})$ for the filtering problem at $t = \mathrm{t}_{k-1}$, the **Chapman-Kolmogorov equation**

$$\pi_Z(z, \mathrm{t}_k|y^{\mathrm{obs}}_{\mathrm{t}_{1:k-1}}) = \int_{\mathbb{R}^{N_z}} \pi_{N_{\mathrm{out}}}(z|z')\, \pi_Z(z', \mathrm{t}_{k-1}|y^{\mathrm{obs}}_{\mathrm{t}_{1:k-1}})\, \mathrm{d}z', \tag{5}$$

yields the PDF for the prediction problem at $t = \mathrm{t}_k$ conditioned on having observed $y^{\mathrm{obs}}_{\mathrm{t}_{1:k-1}}$.

This marginal PDF is then used as a prior for assimilating an observation $y_{\mathrm{obs}}(\mathrm{t}_k)$ via **Bayes' theorem**

$$\pi_Z(z, \mathrm{t}_k|y^{\mathrm{obs}}_{\mathrm{t}_{1:k}}) = \frac{\pi_Y(y_{\mathrm{obs}}(\mathrm{t}_k)|z)\, \pi_Z(z, \mathrm{t}_k|y^{\mathrm{obs}}_{\mathrm{t}_{1:k-1}})}{\int_{\mathbb{R}^{N_z}} \pi_Y(y_{\mathrm{obs}}(\mathrm{t}_k)|z)\, \pi_Z(z, \mathrm{t}_k|y^{\mathrm{obs}}_{\mathrm{t}_{1:k-1}})\, \mathrm{d}z}, \tag{6}$$

which solves the filtering problem at $t = \mathrm{t}_k$.

Definition (Sequential Data Assimilation)

Given a stochastic dynamical system (1) with initial PDF $\pi(z,0)$ and observations $y_{\rm obs}(t)$ for $t = {\rm t}_1, \ldots, {\rm t}_{N_{\rm obs}}$ in observation intervals of $\Delta t_{\rm out} = N_{\rm out}\delta t$, $N_{\rm out} \geq 1$, and forward model (3), the marginal filtering distributions $\pi_Z(z, {\rm t}_k | y^{\rm obs}_{{\rm t}_{1:k}})$, $k = 1, \ldots, N_{\rm obs}$, are recursively defined by first solving the **prediction problem** (5) followed by the **Bayesian assimilation step** (6).

Remark (Coupling of measures)

*Sequential data assimilation, as stated above, provides us with a sequence of marginal filtering distributions $\pi_Z(z, {\rm t}_k | y^{\rm obs}_{{\rm t}_{1:k}})$ for conditioned random variables $\tilde{Z}({\rm t}_k) := Z({\rm t}_k)|y^{\rm obs}_{{\rm t}_{1:k}}$, $k = 1, \ldots, N_{\rm obs}$. However, it does **not** specify joint distributions $\pi_{\tilde{Z}_{{\rm t}_{0:k}}}(\tilde{z}_{{\rm t}_{0:k}} | y^{\rm obs}_{{\rm t}_{1:k}})$ for the family of conditioned random variables*

$$\tilde{Z}_{{\rm t}_{0:k}} = (\tilde{Z}({\rm t}_0), \tilde{Z}({\rm t}_1), \ldots, \tilde{Z}({\rm t}_k)).$$

In particular, nothing is being said about mutual dependencies amongst $\tilde{Z}({\rm t}_{k_1})$ and $\tilde{Z}({\rm t}_{k_2})$ for $k_1 \neq k_2$. In order to define such families of random variables and their joint distributions we would need to employ the concept of coupling as introduced in Chapter 2.

Definition (Abstract formulation of sequential data assimilation)

We introduce the following two operators:

- **Prediction:** $\pi_{Z'} = \mathcal{P}\pi_Z$ defined by (Chapman-Kolmogorov)

$$\pi_{Z'}(z') = \int_{\mathbb{R}^{N_z}} \pi_{N_{\rm out}}(z'|z)\,\pi_Z(z)\,{\rm d}z$$

- **Bayesian inference:** $\pi_{Z'} = \mathcal{L}_k\pi_Z$ defined by (Bayes)

$$\pi_{Z'}(z) = \frac{\pi_Y(y_{\rm obs}(t_k)|z,t_k)\,\pi_Z(z)}{\int_{\mathbb{R}^{N_z}} \pi_Y(y_{\rm obs}(t_k)|z,t_k)\,\pi_Z(z)\,{\rm d}z}.$$

The **sequential data assimilation** (filtering/online learning) problem becomes

$$\pi_Z(z,t_k|y^{\rm obs}_{t_{1:k}}) = \mathcal{L}_k\mathcal{P}\mathcal{L}_{k-1}\mathcal{P}\cdots\mathcal{L}_1\mathcal{P}\pi_{Z^0}(z).$$

6.1 Kalman filter for linear model systems

There is a particular case for which the sequential filtering formulation from the previous section can be implemented algorithmically in closed form; namely when the initial PDF is Gaussian, the evolution model is linear,

$$Z^{n+1} = Z^n + \delta t(DZ^n + b) + \sqrt{2\delta t}\Xi^n, \tag{7}$$

and the forward model is also linear,

$$y_{\rm obs}(\mathrm{t}_k) = Hz_{\rm ref}(\mathrm{t}_k) + \Sigma^k.$$

Here $H \in \mathbb{R}^{N_z \times N_y}$, $D \in \mathbb{R}^{N_z \times N_z}$ and $b \in \mathbb{R}^{N_z}$ denote constant matrices and a constant vector, respectively.

The random variables Ξ^n and Σ^k are independent and Gaussian with mean zero and covariance matrices Q and R, respectively.

Remark (Notations)

We only need to keep track of the mean $\bar{z}$ and the covariance matrix P of the associated Gaussian distributions $\mathrm{N}(\bar{z}, P)$.

*Following standard notation from meteorology, we denote variables arising from a prediction step by superscript f (**forecast**) and those arising from the assimilation of data via Bayes' theorem by superscript a (**analysis**).*

In order to simplify notation we also introduce the shorthand

$$z_k^f = z^f(\mathrm{t}_k),\ z_k^a = z^a(\mathrm{t}_k),\ P_k^f = P^f(\mathrm{t}_k),\ P_k^a = P^a(\mathrm{t}_k).$$

Hence we are dealing with sequences of Gaussian distributions

$$\pi(z, \mathrm{t}_k | y^{\mathrm{obs}}_{\mathrm{t}_{1:k-1}}) = \mathrm{n}(z, \bar{z}_k^f, P_k^f), \qquad \pi(z, \mathrm{t}_k | y^{\mathrm{obs}}_{\mathrm{t}_{1:k}}) = \mathrm{n}(z, \bar{z}_k^a, P_k^a).$$

in terms of given mean vectors and covariance matrices.

Lemma (Linear SDEs)

A single **prediction step** *under the linear model (7) leads to the update*

$$\bar{z}^{n+1} = [I + \delta t D]\bar{z}^n + \delta t b, \tag{8}$$

for the mean, and the update

$$P^{n+1} = [I + \delta t D]P^n[I + \delta t D]^T + 2\delta t Q, \tag{9}$$

for the covariance matrix, where Q is the covariance matrix of the stochastic forcing terms Ξ^n.

Proof.

Recall Chapter 3 and see Section 6.1 from the textbook. □

Remark (Prediction step)

The prediction step is repeated $N_{\rm out}$ times to propagate the mean and the covariance matrix over a time interval $\Delta t_{\rm out} = \delta t N_{\rm out}$.

This results in an explicit transformation of the analysis pair $(\bar{z}^a(t_{k-1}), P^a(t_{k-1}))$ into the forecast pair $(\bar{z}^f(t_k), P^f(t_k))$, given by

$$\bar{z}^f(t_k) = [I + \delta t D]^{N_{\rm out}} \bar{z}^a(t_{k-1}) + \delta t \sum_{i=1}^{N_{\rm out}} [I + \delta t D]^{i-1} b,$$

and

$$P^f(t_k) = [I + \delta t D]^{N_{\rm out}} P^a(t_{k-1}) ([I + \delta t D]^{N_{\rm out}})^T + 2\delta t \sum_{i=1}^{N_{\rm out}} [I + \delta t D]^{i-1} Q ([I + \delta t D]^{i-1})^T.$$

Can be proved by induction.

Lemma (Kalman update step)

We analyse the **Bayesian inference step**:

$$(P_k^a)^{-1} = (P_k^f)^{-1} + H^T R^{-1} H,$$
$$\bar{z}_k^a = \bar{z}_k^f - P_k^a H^T R^{-1}(H\bar{z}_k^f - y_{\rm obs}(t_k)).$$

We employ the **Sherman-Morrison-Woodbury** *matrix inversion formula*

$$(M + U^T N U)^{-1} = M^{-1} - M^{-1}U^T(N^{-1} + UM^{-1}U^T)^{-1}UM^{-1}, \qquad (10)$$

with $M = (P_k^f)^{-1}$, $U = H$ *and* $N = R^{-1}$:

$$P_k^a = P_k^f - P_k^f H^T (R + HP_k^f H^T)^{-1} HP_k^f = P_k^f - KHP_k^f,$$

with **Kalman gain matrix**

$$K = P_k^f H^T (R + HP_k^f H^T)^{-1}.$$

Using the Kalman gain matrix, the update for the mean can also be reformulated as

$$\bar{z}_k^a = \bar{z}_k^f - K(H\bar{z}_k^f - y_{\rm obs}(t_k)).$$

Definition (Kalman Filter)

Given a mean $\bar{z}_0^a$ and a covariance matrix P_0^a at time $t = 0$ and a sequence of observations $y_{\rm obs}(\mathrm{t}_k)$ with error covariance matrix R, $\mathrm{t}_k = k\Delta t_{\rm out}$ and observation interval $\Delta t_{\rm out} = \delta t N_{\rm out}$, the following sequence of steps is performed for $k \geq 1$:

(i) Set $\bar{z}^0 := \bar{z}_{k-1}^a$, $P^0 := P_{k-1}^a$ and iteratively determine $\bar{z}^{n+1}$ and P^{n+1} for $n = 0, \ldots, N_{\rm out} - 1$ *via*

$$\bar{z}^{n+1} = [I + \delta t D]\bar{z}^n + \delta t b, \tag{11}$$

$$P^{n+1} = [I + \delta t D]P^n[I + \delta t D]^T + 2\delta t Q. \tag{12}$$

Set $\bar{z}_k^f := \bar{z}^{N_{\rm out}}$ and $P_k^f := P^{N_{\rm out}}$.

(ii) Compute the Kalman gain matrix

$$K = P_k^f H^T (R + H P_k^f H^T)^{-1},$$

and update the mean and covariance matrix according to

$$\bar{z}_k^a := \bar{z}_k^f - K(H\bar{z}_k^f - y_{\rm obs}(\mathrm{t}_k)), \tag{13}$$

$$P_k^a := P_k^f - KHP_k^f. \tag{14}$$

Example (one-dimensional example)

Consider the scalar stochastic difference equation

$$Z^{n+1} = Z^n + \delta t\, d\, Z^n + \delta t\, b + \sqrt{2\delta t}\, \Xi^n \tag{15}$$

with $d = -0.1$, $b = 1$, $\Xi^n \sim \mathrm{N}(0,1)$, and $\delta t = 0.01$.

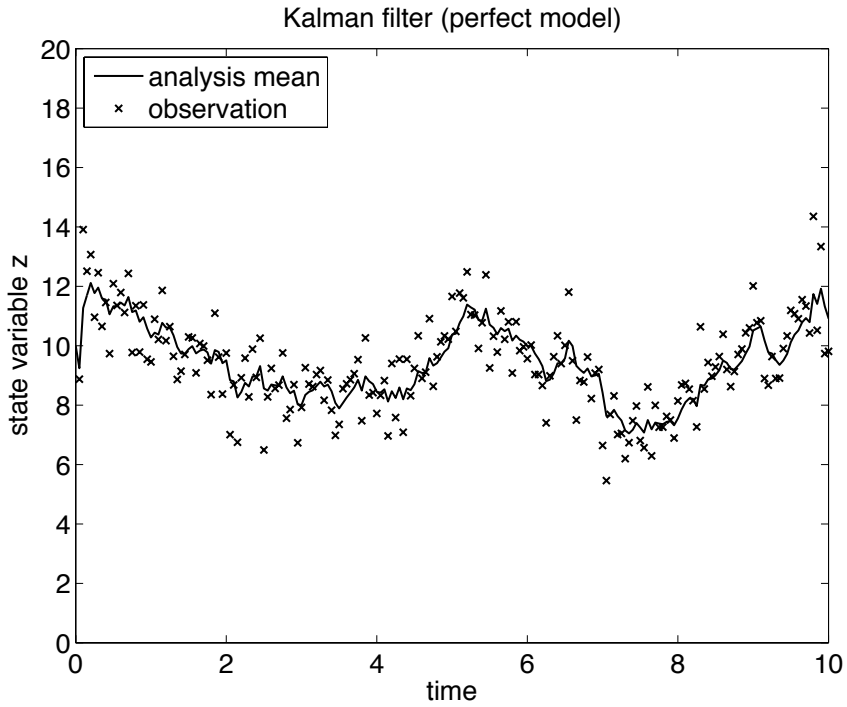
The initial condition has probability distribution $Z^0 \sim \mathrm{N}(10,2)$.

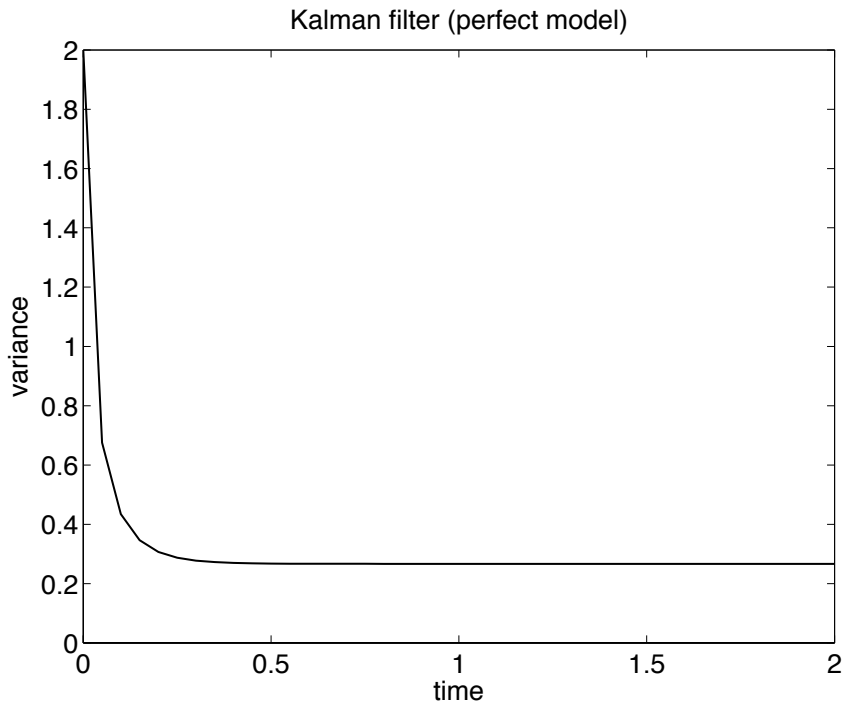
Observations are given in intervals of $\Delta t_{\mathrm{out}} = 0.05$ (*i.e.*, $N_{\mathrm{out}} = 5$), the forward operator is $H = 1$, and the measurement error has distribution $\mathrm{N}(0,1)$.

The reference trajectory $z_{\mathrm{ref}}(t)$ is generated from the same model (15) with initial condition $z_{\mathrm{ref}}(0) = 10$. This setting corresponds to the **perfect model** scenario.

Example (continued)

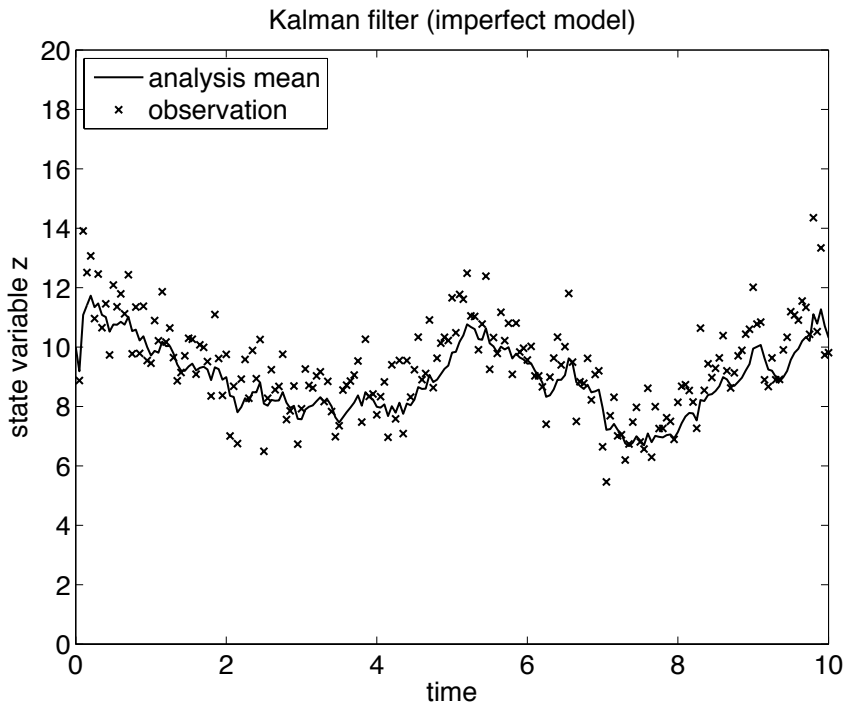
A simple inspection of the Kalman filter equations for the variances reveals that they are independent of the actual observations and, indeed, we find that the analysis and forecast variances quickly converge to stationary values $P_*^a \approx 0.2666$ and $P_f^* \approx 0.3636$, respectively. As an immediate consequence the Kalman gain matrix also becomes stationary with $K \approx 0.2666$.

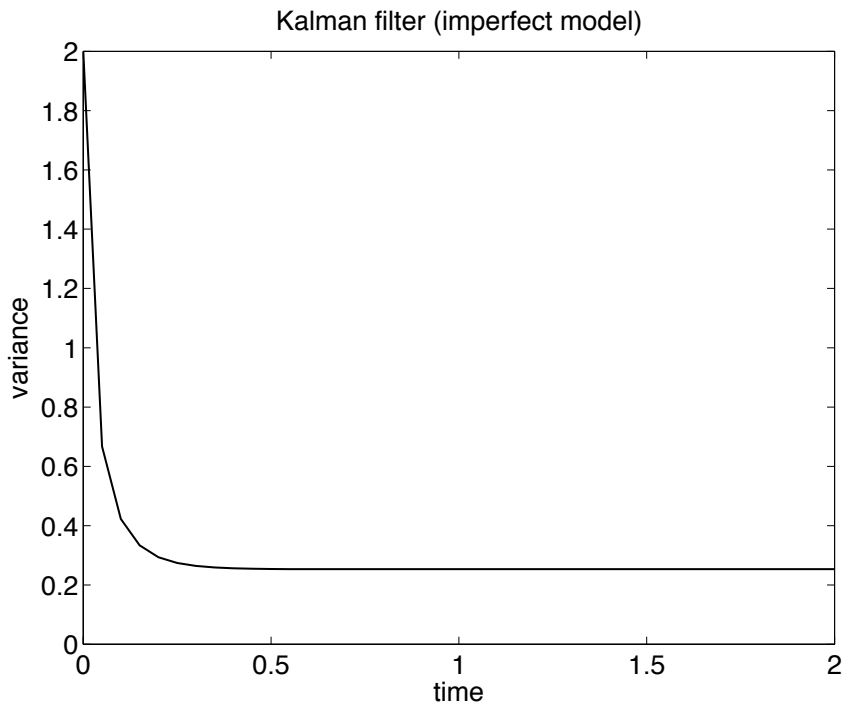




Example (continued)

We now investigate an **imperfect model scenario**, which is more relevant for practical applications. Here we make the model "imperfect" by choosing that the trajectory $z_{\rm ref}(t)$ is still obtained from (15) with $d = -0.1$ while the Kalman filter implementation is based on $d = -0.5$. All other parameters are as specified before (but could have been modified as well). The imperfect model scenario can be recognised from the fact that about 73% of the observed values $y_{\rm obs}(t_k)$ are larger than the analysis $\bar{z}_k^a$, confirming that the model (15) is no longer calibrated.





Example (continued)

The difference between the perfect and the imperfect model scenario can be assessed in terms of time averaged root mean square error (RMSE), and the averaged continuously ranked probability score (CRPS), as defined in Chapter 4.

A long simulation over $N_{\rm obs} = 10^8$ observations reveals that the perfect model scenario leads to slightly better scores (imperfect model: RMSE $= 0.7692$/ perfect model: RMSE $= 0.5162$; imperfect model: CRPS $= 0.6345$/ perfect model: CRPS $= 0.4118$).

The asymptotic value for the analysis variances is slightly reduced for the imperfect model scenario (imperfect model: $P_*^a = 0.2530$ / perfect model $P_*^a = 0.2666$).

It should be noted that the perfect/ imperfect model experiments are based on one and the same set of observations $y_{\rm obs}(t_k)$.

In summary, we find that the performance of the Kalman filter is quite robust with respect to the change in d in (15).

6.2 Variational data assimilation

In the Prolog, a nonlinear method of least squares was used to estimate model states from observations. This was a particular instance of a **variational data assimilation** techniques; such techniques are based on minimising appropriate cost functionals subject to model constraints.

Similar techniques are typically used in **machine learning** and the training of neural networks in particular (back propagation, stochstastic gradient descent, ...). The only difference is that we estimate here the states of a dynamic process while in machine learning you learn, for example, the parameters of a neural network.

In this section, a **connection will be established between the Kalman filter and a variational data assimilation technique**, called four dimensional variational data assimilation (*4DVar*).

Definition (Linear dynamics)

Assumption. It is assumed throughout this subsection that the stochastic model perturbations Ξ^n vanish, *i.e.*, the model error covariance matrix is $Q = 0$. We also set $b = 0$ for notational convenience.

Kalman filter. The $N_{\rm out}$-fold application of a linear model dynamics becomes:

$$z(t_k) = \Psi z(t_{k-1}), \qquad \Psi := (I + \delta t D)^{N_{\rm out}} \in \mathbb{R}^{N_z \times N_z},$$

and the Kalman filter equations are reformulated as follows: For $k = 1, \ldots, N_{\rm obs}$ define

$$\bar{z}_k^f = \Psi \bar{z}_{k-1}^a, \tag{16}$$

$$P_k^f = \Psi P_{k-1}^a \Psi^T, \tag{17}$$

$$(P_k^a)^{-1} = (P_k^f)^{-1} + H^T R^{-1} H, \tag{18}$$

$$\bar{z}_k^a = \bar{z}_k^f - P_k^a H^T R^{-1} (H \bar{z}_k^f - y_{\rm obs}(t_k)), \tag{19}$$

recursively from given initial mean $\bar{z}_0^a$, initial covariance matrix P_0^a, and a sequence of observations $y_{\rm obs}(t_k)$, $k = 1, \ldots, N_{\rm obs}$.

Remark (Smoothing and 4DVar)

We demonstrate that variational data assimilation is closely related to the **smoothing problem**: *determine the marginal distribution at* $\mathrm{t}_0 = 0$ *from observations at* $\mathrm{t}_k > 0$, $k = 1, \ldots, N_{\rm obs}$.

These marginal distributions are again Gaussian under the assumptions stated (linear model and initial Gaussian) and we introduce the notation

$$\pi_Z(z, \mathrm{t}_0 | y^{\rm obs}_{\mathrm{t}_{1:k}}) = \mathrm{n}(z; \bar{z}^s_{0:k}, P^s_{0:k}) \tag{20}$$

in terms of the involved mean vectors $\bar{z}^s_{0:k}$ *and covariance matrices* $P^s_{0:k}$.

Lemma (Variational formulation of the optimal state estimation)

The initial random variable Z_0^a has mean $\bar{z}_0^a$ and covariance matrix P_0^a at time $\mathrm{t}_0 = 0$. Given a sequence of observations $y_{\rm obs}(\mathrm{t}_k)$, with $k = 1, \ldots, N_{\rm obs}$.

The associated Kalman filtering problem at time $\mathrm{t}_{N_{\rm obs}}$ is solved by:

Step (i). *Minimise*

$$L(z) = \frac{1}{2}(z - \bar{z}_0^a)^T (P_0^a)^{-1}(z - \bar{z}_0^a) + \frac{1}{2}\sum_{k=1}^{N_{\rm obs}} (H\Psi^k z - y_{\rm obs}(\mathrm{t}_k))^T R^{-1}(H\Psi^k z - y_{\rm obs}(\mathrm{t}_k)),$$

with respect to z. Denote the **minimiser** *by $\bar{z}^s_{0:N_{\rm obs}}$.*

Step (ii). *Set*

$$\bar{z}^a_{N_{\rm obs}} = \Psi^{N_{\rm obs}} \bar{z}^s_{0:N_{\rm obs}} \qquad P^a_{N_{\rm obs}} = \Psi^{N_{\rm obs}} \mathcal{H}^{-1} (\Psi^{N_{\rm obs}})^T, \qquad P^s_{0:N_{\rm obs}} = \mathcal{H}^{-1},$$

where $\mathcal{H} \in \mathbb{R}^{N_z \times N_z}$ is the **Hessian matrix of second derivatives** *of L with respect to z given by*

$$\mathcal{H} = (P_0^a)^{-1} + \sum_{k=1}^{N_{\rm obs}} (\Psi^k)^T H^T R^{-1} H \Psi^k.$$

Proof.

Proof by induction. See textbook. □

Remark

The **minimisation problem** *can be solved by (stochastic)* **gradient descent** *using* **backpropagation** *in order to compute the gradient.*

Remark

In this linear setting **without stochastic model contributions**, *our discussion suggests that 4DVar is equivalent to first running a Kalman filter up to* $\mathrm{t}_{N_{\rm obs}}$ *resulting in* $\bar{z}^a_{N_{\rm obs}}$, *then applying the inverse dynamic operator* Ψ^{-1} $N_{\rm obs}$*-times to obtain*

$$\bar{z}^s_{0:N_{\rm obs}} = \Psi^{-N_{\rm obs}} \bar{z}^a_{N_{\rm obs}}.$$

Similarly, we obtain the smoothing covariance matrix

$$P^s_{0:N_{\rm obs}} = \Psi^{-N_{\rm obs}} P^a_{N_{\rm obs}} (\Psi^{-N_{\rm obs}})^T.$$

This reformulation of the smoothing problem allows for a generalisation to **nonlinear dynamical systems** of type

$$z^{n+1} = z^n + \delta t\, f(z^n), \tag{21}$$

as considered in the Prolog.

If $\Delta t_{\rm out} = \delta t N_{\rm out}$ (as assumed before), then we define the map ψ as the result of an $N_{\rm out}$-fold application of (21) and formally introduce the iteration

$$Z^{k+1} = \psi(Z^k), \tag{22}$$

in order to describe the model dynamics in between observations.

Again, we have assumed that stochastic contributions to the model dynamics can be ignored, *e.g.*, ψ could correspond to any **deterministic dynamical system**.

Extensions to stochastic models are possible but beyond the scope of this course.

Definition (Four-Dimensional Variational Data Assimilation (4DVar))

Let a mean $\bar{z}_0^a$ and a covariance matrix P_0^a at time $t = 0$ be given as well as a sequence of observations $y_{\rm obs}(t_k)$ with $k = 1, \ldots, N_{\rm obs}$.

The **smoothing distribution** at $t_0 = 0$ is given by

$$\pi_Z(z, t_0 | y^{\rm obs}_{t_{1:N_{\rm obs}}}) \propto e^{-\frac{1}{2}(z-\bar{z}_0^a)^T (P_0^a)^{-1}(z-\bar{z}_0^a)} \, \pi_{Y_{t_{1:N_{\rm obs}}}}(y^{\rm obs}_{t_{1:N_{\rm obs}}} | z) \, .$$

The associated **MAP estimator** (see Chapter 4), denoted by $\hat{z}^s_{0:N_{\rm obs}}$, is the minimiser of the functional $L(z) = -\log \pi_Z(z, t_0 | y^{\rm obs}_{t_{1:N_{\rm obs}}})$, that is,

$$L(z) = \frac{1}{2}(z - \bar{z}_0^a)^T (P_0^a)^{-1}(z - \bar{z}_0^a) + \frac{1}{2} \sum_{k=1}^{N_{\rm obs}} (H\psi^k(z) - y_{\rm obs}(t_k))^T R^{-1} (H\psi^k(z) - y_{\rm obs}(t_k)). \qquad (23)$$

The associated $\hat{z}^a_{N_{\rm obs}} = \psi^{N_{\rm obs}}(\hat{z}^s_{0:N_{\rm obs}})$ provides the MAP estimator at the end of the observation interval $t_{N_{\rm obs}}$.

The process of obtaining the MAP estimator is called **4DVar**.

Remark (Cycled 4DVar)

Instead of applying 4DVar to a complete set of observations, we can also apply 4DVar recursively to **patches of observations** *from intervals*

$$[0, t_{N_a}], \quad [t_{N_a}, t_{2N_a}], \quad \cdots$$

with the integer N_a chosen appropriately.

One often works with a **fixed covariance matrix** *B instead of an observation adjusted (P_0^a), $(P_{N_a}^a)$, etc. The* **MAP estimator** *at the end of the last patch is used as the mean in the cost functional for the next patch.*

Consider the second patch over the interval $[t_{N_a}, t_{2N_a}]$ with cost functional

$$L(z) = \frac{1}{2}(z - \hat{z}_{N_a}^a)^T B^{-1}(z - \hat{z}_{N_a}^a) + \frac{1}{2}\sum_{k=1}^{N_a}(H\psi^k(z) - y_{\rm obs}(t_{k+N_a}))^T R^{-1}(H\psi^k(z) - y_{\rm obs}(t_{k+N_a})), \qquad (24)$$

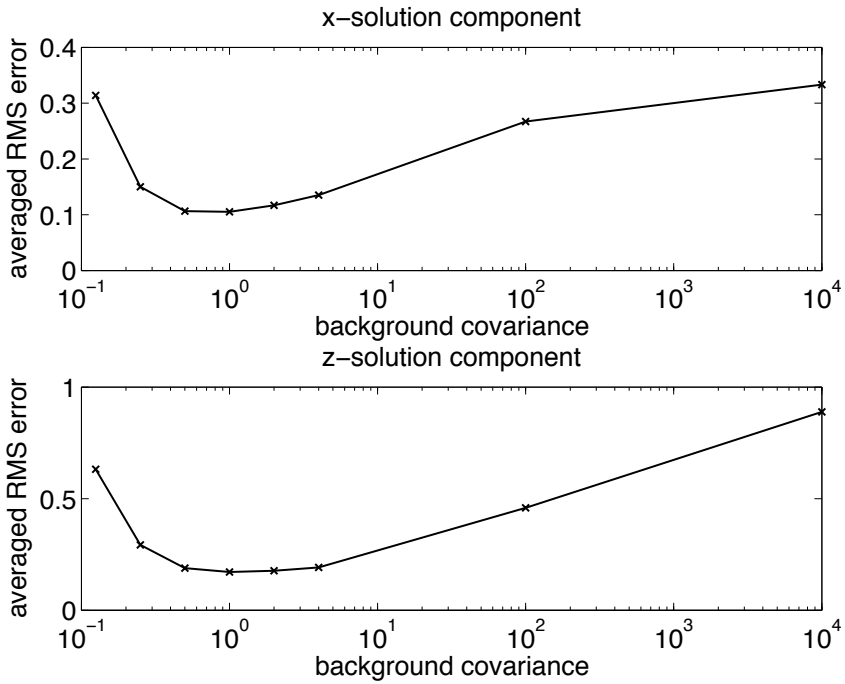
with the analysis $\hat{z}_{N_a}^a$ from the assimilation of the first patch of observations.

Example (Lorenz-63)

We return now to the Lorenz-63 example from the Prolog. There we minimised

$$L(z) = \frac{1}{2}\sum_{k=1}^{N_a}(H\psi^k(z) - y_{\rm obs}(t_k))^T R^{-1}(H\psi^k(z) - y_{\rm obs}(t_k)), \qquad (25)$$

i.e., the **prior distribution was ignored**. Here we use instead a covariance matrix $B = \delta I$ for all assimilation windows. Results for varying δ:



Remark (4DVar, smoothing, Monte Carlo, numerical implementation)

We conclude this discussion by emphasising that 4DVar provides a mathematically elegant approach to state estimation in the context of smoothing.

In combination with Monte Carlo sampling, variational data assimilation can also be used to compute expectation values with respect to $\pi_Z(z, t_0 | y^{\rm obs}_{t_{1:N_{\rm obs}}})$.

The impact of stochastic model contributions on state estimation can also be accounted for within a variational framework.

Efficient implementations of 4DVar use Gauss-Newton-type methods and adjoint formulations to compute gradients.

Closely related to many training algorithms in machine learning which are based on minimising appropriate cost functions.

6.3 Particle filters

Sequential Monte Carlo (SMC) methods (also called **particle filters**) combine **ensemble prediction** (or Monte Carlo) for the underlying model dynamics with **importance sampling** in the Bayesian inference step.

We first recall the basic idea of ensemble prediction.

Definition (ensemble prediction)

The ensemble prediction approach to the approximation of marginal PDFs $\pi_Z(z,t)$ of an SDE model

$$\mathrm{d}Z = f(Z)\mathrm{d}t + \sqrt{2}Q^{1/2}\mathrm{d}W(t)$$

is build upon an ensemble of time-discrete approximations

$$z_i^{n+1} = z_i^n + \delta t f(z_i^n) + \sqrt{2\delta t}Q^{1/2}\xi_i^n, \qquad i = 1,\ldots,M, \tag{26}$$

$n \geq 0$, with the values $\{\xi_i^n\}$ denoting realisations of independent and identically distributed Gaussian random variables with mean zero and covariance matrix equal to the identity matrix.

The initial conditions z_i^0 at $t = 0$ are realisations of a random variable with PDF $\pi_{Z^0}(z) = \pi_Z(z,0)$.

Remark

Ensemble prediction method is a Monte Carlo implementation of the prediction step

$$\pi_{Z^{n+1}}(z^{n+1}) = \mathcal{P}_{\delta t}\, \pi_{Z^n}(z^{n+1}) := \int_{\mathbb{R}^{N_z}} \pi(z^{n+1}|z^n)\, \pi_{Z^n}(z^n)\, \mathrm{d}z^n$$

with transition kernel given by (2) and

$$\pi_{Z^{n+N_{\rm out}}} = \mathcal{P}\, \pi_{Z^n} := (\mathcal{P}_{\delta t})^{N_{\rm out}}\, \pi_{Z^n}$$

for the propagation of the model states from one observation to the next.

The idea of particle filters is to combine ensemble prediction with **importance sampling**.

Each ensemble member is assigned a weight $w_i^0 = 1/M$ at $t = 0$. These weights are kept constant during each ensemble prediction step (26) provided that there is no observation to be assimilated.

We now assume that after $N_{\rm out}$ steps under (26), an observation $y_{\rm obs}(t_1)$ becomes available at $t_1 = \Delta t_{\rm out} := \delta t N_{\rm out}$ with likelihood function $\pi_Y(y_{\rm obs}(t_1)|z)$.

Then, an application of **Bayes' formula** to the forecast ensemble $z_i^f(t_1) = z_i^{N_{\rm out}}$ leads to rescaling the weights according to

$$w_i^1 = \frac{w_i^0 \pi_Y(y_{\rm obs}(t_1)|z_i^f(t_1))}{\sum_{j=1}^M w_j^0 \pi_Y(y_{\rm obs}(t_1)|z_j^f(t_1))} = \frac{\pi_Y(y_{\rm obs}(t_1)|z_i^f(t_1))}{\sum_{j=1}^M \pi_Y(y_{\rm obs}(t_1)|z_j^f(t_1))},$$

$i = 1, \ldots, M$, since $w_i^0 = 1/M$.

The new weights w_i^1 are again kept constant whilst the states are incremented according to (26), until a second set of observations $y_{\rm obs}(t_2)$ becomes available at $t_2 = 2\Delta t_{\rm out}$ and a new forecast ensemble is defined by $z_i^f(t_2) = z_i^{2N_{\rm out}}$.

At that point the weights are changed to

$$w_i^2 = \frac{w_i^1 \pi_Y(y_{\rm obs}(t_2)|z_i^f(t_2))}{\sum_{j=1}^M w_j^1 \pi_Y(y_{\rm obs}(t_2)|z_j^f(t_2))}, \quad i = 1, \ldots, M.$$

Definition (Sequential Importance Sampling (SIS))

We obtain the general update formula

$$w_i^k = \frac{w_i^{k-1} \pi_Y(y_{\rm obs}(t_k)|z_i^f(t_k))}{\sum_{j=1}^M w_j^{k-1} \pi_Y(y_{\rm obs}(t_k)|z_j^f(t_k))}, \qquad z_i^f(t_k) = z_i^{kN_{\rm out}}, \tag{27}$$

with the variables $\{z_i^n\}$, $n \geq 0$, obtained by simple propagation of the initial ensemble $\{z_i^0\}$ under the discrete model dynamics (26) throughout the whole assimilation process.

The resulting particle filter is often referred to as **sequential importance sampling** (SIS) filter.

It is to be expected that the weights will become non-uniform after a few assimilation steps. In order to quantify the non-uniformity in the weights we introduce the **effective sample size**.

Definition (Effective sample size)

Given a set of weights $w_i \geq 0$, $i = 1, \ldots, M$, the **effective sample size** is defined by

$$M_{\text{effective}} = \frac{1}{\sum_{i=1}^{M} w_i^2}. \tag{28}$$

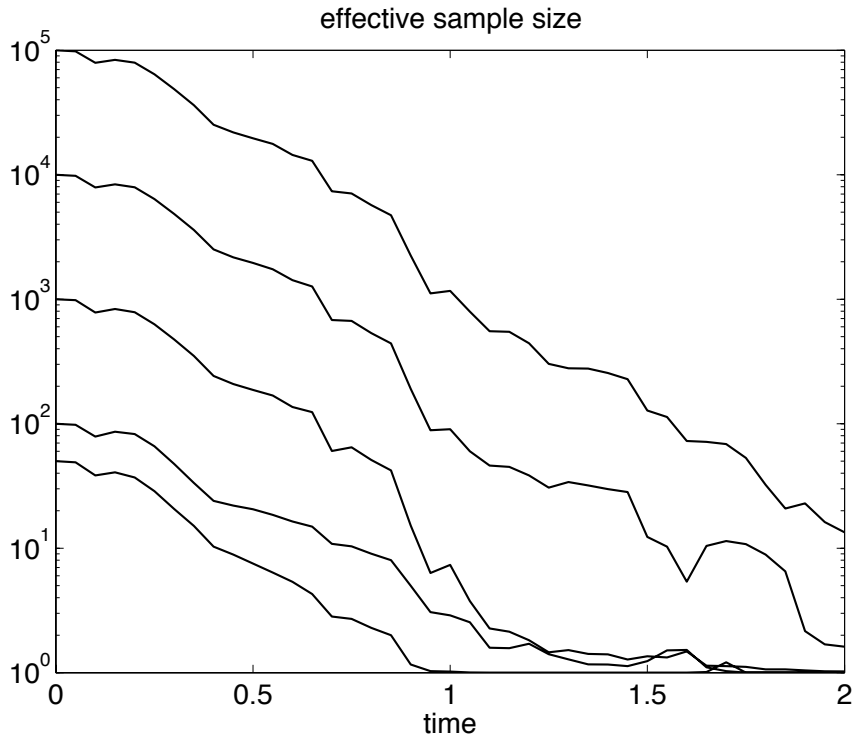
The effective sample size is equal to M in the case of uniform weights $w_i = 1/M$. In the extreme non-uniform case $w_i = 1$ for one single ensemble member i and $w_j = 0$ for all $j \neq i$, we obtain $M_{\text{effective}} = 1$.

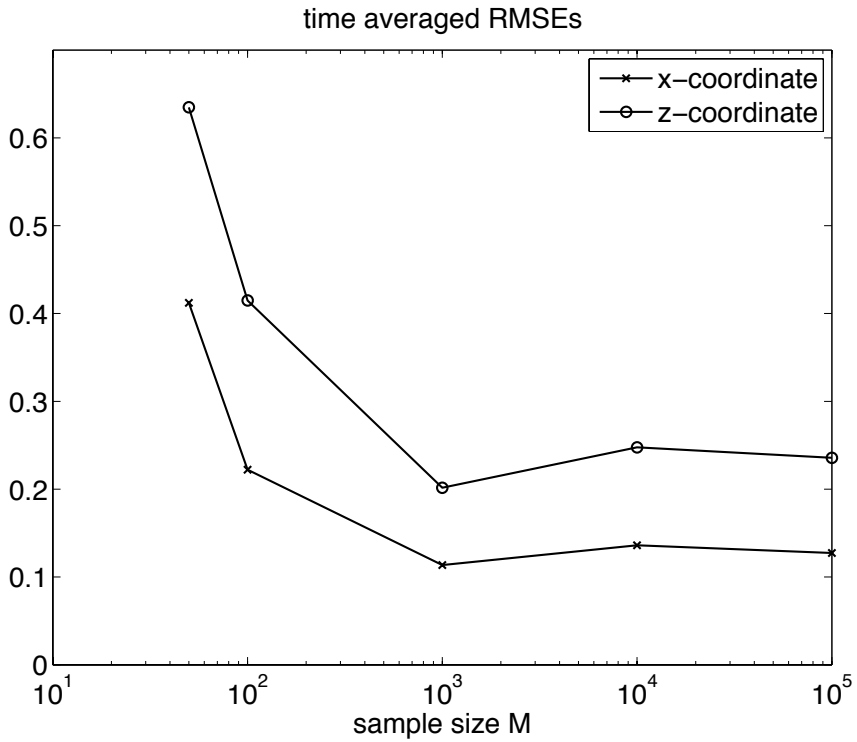
Example (SIS for Lorenz-63)

We consider a stochastically perturbed Lorenz-63 model and only observe its x-component. The required likelihood function is given by

$$\pi_Y(\mathrm{x}_{\mathrm{obs}}(\mathrm{t}_k)|z_i^f(\mathrm{t}_k)) = \frac{1}{\sqrt{2\pi R}} e^{-\frac{1}{2R}(\mathrm{x}_{\mathrm{obs}}(\mathrm{t}_k)-\mathrm{x}_i^f(\mathrm{t}_k))^2},$$

with error variance $R = 1/15$. Observations are collected every $\Delta t_{\mathrm{out}} = 0.05$.





Definition (Sequential Importance resampling (SIR) filter)

Given an initial distribution $\pi_Z(z,0)$, we draw M independent realisations $z_i(t_0)$ from this distribution with equal weights $w_i^0 = 1/M$.

The following steps are performed recursively for $k = 1, \ldots, N_{\rm obs}$:

(i) In between observations, the ensemble is propagated under an evolution model, such as (26), with constant weights w_i^{k-1} producing a forecast ensemble $\{z_i^f(t_k)\}_{i=1}^M$.

(ii) An observation $y_{\rm obs}(t_k)$ at $t_k = k\Delta t_{\rm out} = \delta t k N_{\rm out}$ results in a change of weights from w_i^{k-1} to w_i^k, $i = 1, \ldots, M$, according to (27).

(iii) After each change of weights, the effective sample size is evaluated.

(a) If $M_{\rm effective}$ drops below $M/2$, a new set of M equally weighted ensemble members $z_i^a(t_k)$ is generated from the forecast ensemble $z_i^f(t_k)$ by the method of monomial/residual resampling. Finally we return to (i) with the resampled ensemble, *i.e.*, with ensemble members given by

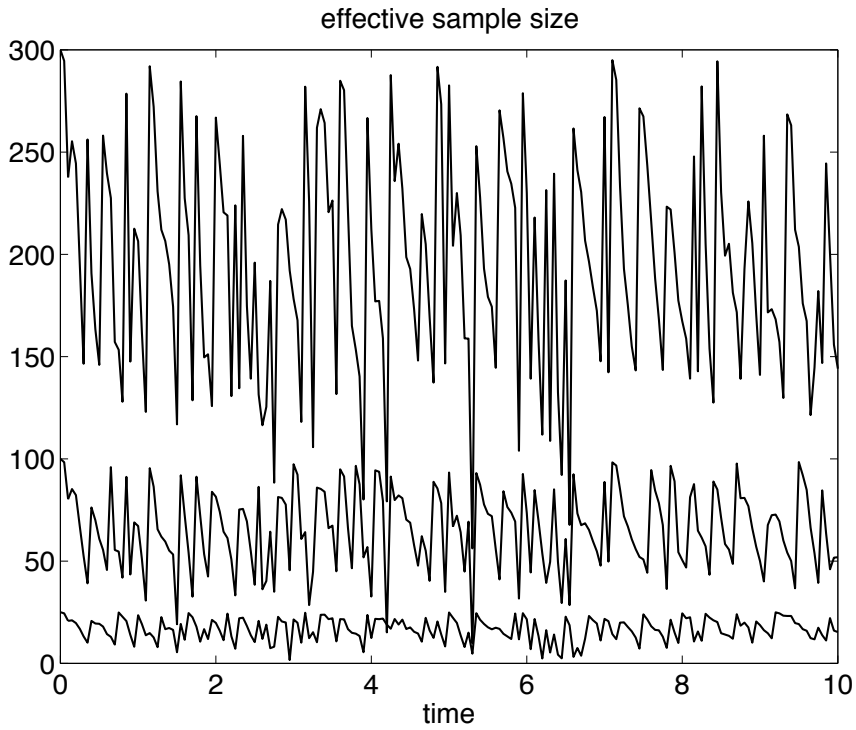
$$z_i(t_k) := z_i^a(t_k),$$

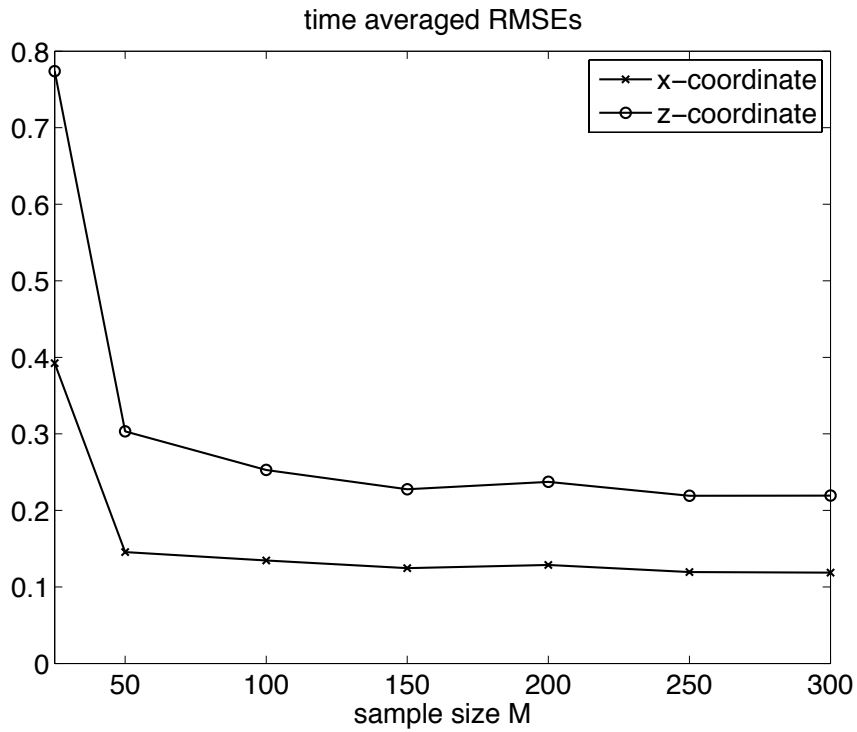
weights given by $w_i^k := 1/M$, and k incremented by one.

(b) Otherwise we continue in (i) with $z_i(t_k) = z_i^f(t_k)$, weights w_i^k as computed under (ii), and increment k by one.

Example (SIR for Lorenz-63)

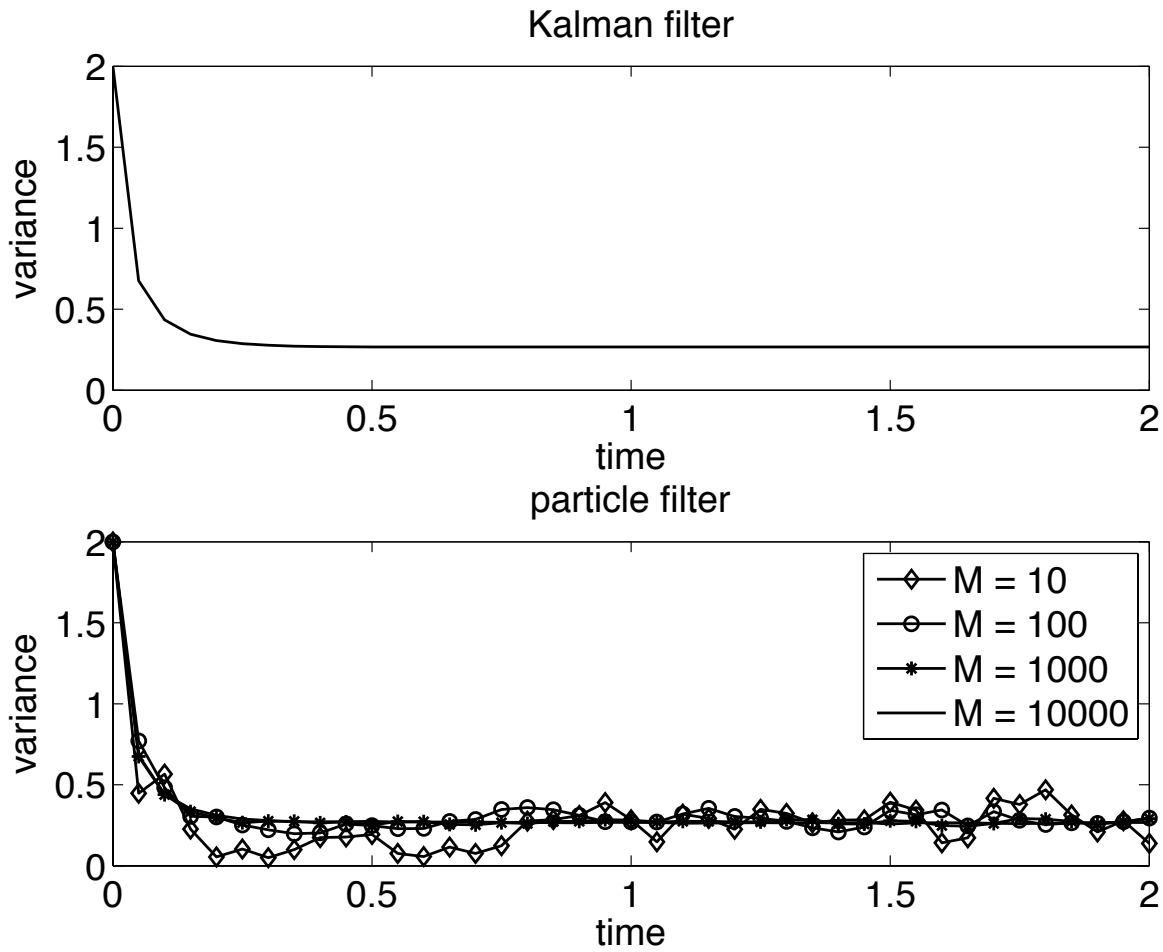
We repeat the Lorenz-63 experiment with the SIS filter being replaced by the SIR filter. The SIR filter is able to successfully track the reference solution as given by $z_{\rm ref}(t)$.





Example (one-dimensional linear problem)

We apply the SIR filter to a one-dimensional linear dynamical system. We find that the SIR filter converges as M is increased and, furthermore, a comparison with the Kalman filter reveals that the SIR filter converges to the same asymptotic value of the posterior variance.



Theorem (Convergence of SIR filter)

The SIR filter with resampling after each propagation step can be written as

$$\hat{\pi}_Z(z, \mathrm{t}_k|y^{\mathrm{obs}}_{\mathrm{t}_{1:k}}) = \mathcal{L}_k \left[\mathcal{S}^M \mathcal{P}_{\delta t}\right]^{N_{\mathrm{out}}} \mathcal{S}^M \mathcal{L}_{k-1} \left[\mathcal{S}^M \mathcal{P}_{\delta t}\right]^{N_{\mathrm{out}}} \mathcal{S}^M \cdots \mathcal{L}_1 \left[\mathcal{S}^M \mathcal{P}_{\delta t}\right]^{N_{\mathrm{out}}} \mathcal{S}^M \pi_{Z^0}.$$

Recall that the exact PDF satisfies

$$\pi_Z(z, \mathrm{t}_k|y^{\mathrm{obs}}_{\mathrm{t}_{1:k}}) = \mathcal{L}_k \mathcal{P} \mathcal{L}_{k-1} \mathcal{P} \cdots \mathcal{L}_1 \mathcal{P} \pi_{Z^0}(z).$$

If the likelihood function $\pi_Y(y|z)$ satisfies $\kappa \leq \pi_Y(y|z) \leq \kappa^{-1}$ for $\kappa \in (0, 1]$, then

$$d(\mathcal{L}_j \pi_Z, \mathcal{L}_j \hat{\pi}_Z) \leq 2\kappa^{-2} d(\pi_z, \hat{\pi}_Z)$$

for all $j = 1, \ldots, k$, and

$$d(\pi_Z(\cdot, \mathrm{t}_k|y^{\mathrm{obs}}_{\mathrm{t}_{1:k}}), \hat{\pi}_Z(\cdot, \mathrm{t}_k|y^{\mathrm{obs}}_{\mathrm{t}_{1:k}})) \leq \sum_{j=1}^{k} (2\kappa^{-2})^j \frac{N_{\mathrm{out}} + 1}{\sqrt{M}}.$$

Remark (Importance sampling)

We investigate the approximation properties of the importance sampling

$$\hat{\pi}_Z^{\rm a} = \mathcal{L}\mathcal{S}^M \pi_Z^{\rm f}$$

approximation to the posterior PDF $\pi_Z^{\rm a} = \mathcal{L}\pi_Z^{\rm f}$ *given* M *samples from the prior* $\pi_Z^{\rm f}$ *and an observation* $y_{\rm obs}$.

Agapio et al (2015) have shown that

$$d(\hat{\pi}_Z^{\rm a}, \pi_Z^{\rm a}) \leq 2\frac{\rho^{1/2}}{M^{1/2}}$$

with

$$\rho = \frac{\pi_Z^{\rm f}[\pi_Y(y_{\rm obs}|\cdot)^2]}{\pi_Z^{\rm f}[\pi_Y(y_{\rm obs}|\cdot)]^2}$$

Note that

$$\frac{M}{\rho} \approx M_{\rm effective}\,.$$

Remark (Curse of dimensionality)

Note that

$$\rho \geq \exp\left(D_{\mathrm{KL}}(\pi_Z^{\mathrm{a}} \| \pi_Z^{\mathrm{f}})\right).$$

(Proof by Jensen's inequality.)

The **Kullback-Leibler divergence** *is given by*

$$D_{\mathrm{KL}}(\pi_Z^{\mathrm{a}} \| \pi_Z^{\mathrm{f}}) = \int_{\mathbb{R}^{N_z}} \frac{\pi_Y(y|z)}{\pi_Y(y)} \log \frac{\pi_Y(y|z)}{\pi_Y(y)} \pi_Z^{\mathrm{f}}(z)\,\mathrm{d}z \geq 0$$

and $y = y_{\mathrm{obs}}$.

If we are in an **independent and identically distributed situation**, *that is,*

$$\pi_Z^{\mathrm{f}}(z) = \prod_{l=1}^{N_z} \pi_{Z_l}^{\mathrm{f}}(z_l), \qquad \pi_Y(y|z) = \prod_{l=1}^{N_z} \pi_{Y_l}(y_l|z_l), \qquad y_l, z_l \in \mathbb{R}$$

and $N_z \to \infty$, *then*

$$D_{\mathrm{KL}}(\pi_Z^{\mathrm{post}} \| \pi_Z^{\mathrm{prior}}) \to \infty\,.$$

Any resampling scheme from Chapter 4 can be used in a sequential Monte Carlo methods in particular those based on coupling arguments.

If the underlying dynamical model is deterministic, then resampling can lead to the creation of identical particles which remain identical under the dynamics.

This leads to the concept of particle rejuvenation. Particle rejuvenation is also often used to make a particle filter more robust against filter divergence.

Definition (Particle rejuvenation)

Let $\{z_i^a\}$ denote an equally weighted analysis ensemble with some ensemble members being identical, *i.e.*, there exist indices i_1 and i_2 such that $z_{i_1}^a = z_{i_2}^a$.

Let $B \in \mathbb{R}^{N_z \times N_z}$ denote a given covariance matrix and $\tau > 0$ a given *bandwidth parameter*. Then particle rejuvenation replaces each analysis ensemble member z_i^a, $i = 1, \ldots, M$, by a single draw from the Gaussian distribution $\mathrm{N}(z_i^a, \tau B)$.

Note that the analysis ensemble remains unaltered for $\tau = 0$ and that B can be chosen to be the empirical covariance matrix of the forecast ensemble.