MTMD02 - Operational Data Assimilation Techniques

Part II: Sequential Estimation

Stefano Migliorini
Department of Meteorology, University of Reading
s.migliorini@reading.ac.uk
Room 1U11

January 26, 2011

Contents

1 Introduction

Since the beginning of the 20th century it has been known that predicting the weather can be quantitatively thought of as an initial-value problem. This means that the knowledge of the state of the flow (here assumed to be the atmospheric flow) $\mathbf{x}_{t_1}$ at a given time t_1 depends on the knowledge of the state $\mathbf{x}_{t_0}$ at a previous time t_0 as well as on a set of laws (hereafter, the weather prediction model) describing the evolution of the flow. It is also necessary to know the values of the state variables over the boundaries of the weather prediction domain and the values of any parameters on which the model may depend.

When initial (or boundary) conditions are not known exactly, the solution of the equations is not unique: forecasting the weather is an ill-posed problem. In other words, no matter how small initial condition (or "analysis") errors are: as long as they are not zero, it is no longer possible to find a unique solution, e.g., a unique value of surface temperature at a given location and at a given future time. To minimize forecast errors, not only we need accurate models but also good knowledge of initial (and boundary) conditions. It is essential that good quality, widespread and frequent measurements (e.g., from satellites) are assimilated in NWP models in order to achieve good quality forecasts.

The main challenge to pursue this strategy is given by the very large number of components of the state (also called the state vector) for weather forecasting applications: the current configuration of the UK Met Office operational global model, for example, includes 1024 x 769 model levels over 70 vertical levels. This means that it is necessary to know the initial conditions of each model variable over more than 55 million grid points. However, the fundamental difficulty is given by the inevitable error that affects every observation, so that initial conditions are always known with finite precision. The insufficient amount of observational information is complemented with information from short-range model forecasts, also affected by errors. Combining observational and prior information to determine the best estimate of the state of the system (here the atmosphere) over a given time window is the aim of data assimilation.

1.1 Smoothers and filters

In this part of the course we will introduce concepts related to sequential data assimilation techniques, which deal with determining the best estimate of the state at a given time. Variational techniques, on the other hand, have the aim of estimating the optimal state over a fixed time interval (also known as the data assimilation window). We will see that this corresponds to considering the solution of either the *filtering* problem or the *smoothing* problem. This means that 4D-Var provides the solution to the fixed-interval smoothing problem, by using all observations available in the interval, while the Kalman filter provides the solution of the filtering problem at time t by using only observational information available up to time t.

However, the distinction between variational and sequential methods is only methodological (i.e., they differ in the way the estimate is determined), as it is possible to prove that 4D-Var is equivalent to a fixed-interval Kalman smoother initialized with the same background state and covariance matrix at the beginning of the assimilation window. Let us now prove this result in a simple case, when the system model is considered to be perfect and linear, so that we can write

$$\mathbf{x}(t_{i+1}) = \mathbf{M}\mathbf{x}(t_i) \tag{1}$$

and we are given a set of observations

$$\mathbf{y}(t_{i+1}) = \mathbf{H}\mathbf{x}(t_{i+1}) + \boldsymbol{\epsilon}^o(t_{i+1}) \tag{2}$$

with $E\{\boldsymbol{\epsilon}^o(t_{i+1})\boldsymbol{\epsilon}^o(t_{i+1})^T\} = \mathbf{R}(t_{i+1})$. First let us define the optimal smoothed estimate as the estimate generated by combining the optimal estimate $\hat{\mathbf{x}}(t_i^+)$ from the *forward filter* – which takes into account measurements up to and including time t_i, with initial conditions $\hat{\mathbf{x}}(t_0)$ and $\mathbf{P}(t_0)$ – and the optimal estimate $\hat{\mathbf{x}}_b(t_i^-)$ from the *backward filter*, which takes into account measurements after time t_i, with initial conditions $\hat{\mathbf{x}}(t_N) = \mathbf{0}$ and $\mathbf{P}^{-1}(t_N) = \mathbf{0}$.

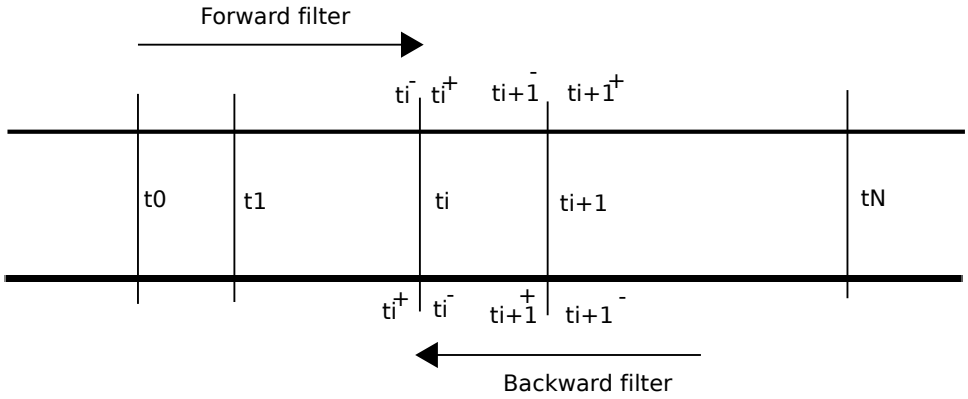

Figure 1: Schema of a smoother.

The combined optimal estimate $\mathbf{x}^s(t_i)$ is determined by interpreting $\hat{\mathbf{x}}(t_i^+)$ and $\hat{\mathbf{x}}_b(t_i^-)$ as two distinct measurements of $\mathbf{x}(t_i)$, with covariance $\mathbf{P}(t_i^+)$ and $\mathbf{P}_b(t_i^-)$, respectively, so that the weighted average $\mathbf{x}^s(t_i)$ is given by

$$\hat{\mathbf{x}}^s(t_i) = [\mathbf{P}^{-1}(t_i^+) + \mathbf{P}_b^{-1}(t_i^-)]^{-1}[\mathbf{P}^{-1}(t_i^+)\hat{\mathbf{x}}(t_i^+) + \mathbf{P}_b^{-1}(t_i^-)\hat{\mathbf{x}}_b(t_i^-)]. \tag{3}$$

Now let us assume that for simplicity we want to determine the combined estimate at t_0 when the only observations are available at time t_1, assumed to be the end of the time interval. In this case we can write

$$\begin{aligned}
\hat{\mathbf{x}}(t_0^+) &= \hat{\mathbf{x}}(t_0) && (4)\\
\mathbf{P}(t_0^+) &= \mathbf{P}(t_0) && (5)\\
\hat{\mathbf{x}}_b(t_0^-) &= \mathbf{M}^{-1}\hat{\mathbf{x}}_b(t_1^+) && (6)\\
\mathbf{P}_b(t_0^-) &= \mathbf{M}^{-1}\mathbf{P}_b(t_1^+)\mathbf{M}^{-T} && (7)
\end{aligned}$$

where we have assumed that there is no model error. We can also write

$$\begin{aligned}
\hat{\mathbf{x}}_b(t_1^+) &= \mathbf{P}_b(t_1^+)[\mathbf{P}_b^{-1}(t_1^-)\hat{\mathbf{x}}_b(t_1^-) + \mathbf{H}^T(t_1)\mathbf{R}^{-1}(t_1)\mathbf{y}_1] = \mathbf{P}_b(t_1^+)\mathbf{H}^T(t_1)\mathbf{R}^{-1}(t_1)\mathbf{y}_1 && (8)\\
\mathbf{P}_b(t_1^+) &= (\mathbf{P}_b^{-1}(t_1^-) + \mathbf{H}^T(t_1)\mathbf{R}^{-1}(t_1)\mathbf{H}(t_1))^{-1} = (\mathbf{H}^T(t_1)\mathbf{R}^{-1}(t_1)\mathbf{H}(t_1))^{-1} && (9)
\end{aligned}$$

so that

$$\hat{\mathbf{x}}_b(t_1^+) = (\mathbf{H}^T(t_1)\mathbf{R}^{-1}(t_1)\mathbf{H}(t_1))^{-1}\mathbf{H}^T(t_1)\mathbf{R}^{-1}(t_1)\mathbf{y}_1. \tag{10}$$

Given that $\mathbf{P}_b(t_0^-) = \mathbf{M}^{-1}(\mathbf{H}^T(t_1)\mathbf{R}^{-1}(t_1)\mathbf{H}(t_1))^{-1}\mathbf{M}^{-T}$ It follows that

$$\begin{aligned}
\hat{\mathbf{x}}^s(t_0) &= (\mathbf{P}^{-1}(t_0) + \mathbf{M}^T\mathbf{H}^T(t_1)\mathbf{R}^{-1}(t_1)\mathbf{H}(t_1)\mathbf{M})^{-1}(\mathbf{P}^{-1}(t_0)\hat{\mathbf{x}}(t_0) + && (11)\\
&\quad \mathbf{M}^T\mathbf{H}^T(t_1)\mathbf{R}^{-1}(t_1)\mathbf{H}(t_1)\mathbf{M}\mathbf{M}^{-1}(\mathbf{H}^T(t_1)\mathbf{R}^{-1}(t_1)\mathbf{H}(t_1))^{-1}\mathbf{H}^T(t_1)\mathbf{R}^{-1}(t_1)\mathbf{y}_1)\\
&= (\mathbf{P}^{-1}(t_0) + \mathbf{M}^T\mathbf{H}^T(t_1)\mathbf{R}^{-1}(t_1)\mathbf{H}(t_1)\mathbf{M})^{-1}(\mathbf{P}^{-1}(t_0)\hat{\mathbf{x}}(t_0) + \mathbf{M}^T\mathbf{H}^T(t_1)\mathbf{R}^{-1}(t_1)\mathbf{y}_1).
\end{aligned}$$

The 4D-Var method aims to determine the initial conditions $\hat{\mathbf{x}}^{4DVAR}(t_0)$ by finding the minimum of

$$J(\mathbf{x}(t_0)) = \frac{1}{2}(\mathbf{x}(t_0)-\hat{\mathbf{x}}(t_0))^T\mathbf{P}^{-1}(t_0)(\mathbf{x}(t_0)-\hat{\mathbf{x}}(t_0))+\frac{1}{2}\sum_{i=1}^{N}(\mathbf{H}(t_i)\mathbf{x}(t_i)-\mathbf{y}(t_i))^T\mathbf{R}^{-1}(\mathbf{H}(t_i)\mathbf{x}(t_i)-\mathbf{y}(t_i)) \tag{12}$$

where we can write

$$\mathbf{x}(t_i) = \mathbf{M}(t_{i-1})\mathbf{M}(t_{i-2})\cdots\mathbf{M}(t_0)\mathbf{x}(t_0) \equiv \mathbf{M}(t_{i-1}:t_0)\mathbf{x}(t_0). \tag{13}$$

The minimum of $J(\mathbf{x}(t_0))$ is found by finding the state that makes $\nabla J(\mathbf{x}(t_0)) = \mathbf{0}$. We have

$$\nabla J(\mathbf{x}(t_0)) = \mathbf{P}^{-1}(t_0)(\mathbf{x}(t_0)-\hat{\mathbf{x}}(t_0))+\sum_{i=1}^{N}\mathbf{M}(t_{i-1}:t_0)^T\mathbf{H}(t_i)^T\mathbf{R}^{-1}(\mathbf{H}(t_i)\mathbf{M}(t_{i-1}:t_0)\mathbf{x}(t_0)-\mathbf{y}(t_i)). \tag{14}$$

When $N = 1$ we get $\nabla J(\mathbf{x}(t_0)) = \mathbf{0}$ for

$$(\mathbf{P}^{-1}(t_0)+\mathbf{M}(t_0)^T\mathbf{H}(t_i)^T\mathbf{R}^{-1}\mathbf{H}(t_i)\mathbf{M}(t_0))\hat{\mathbf{x}}^{4DVAR}(t_0) = \mathbf{P}^{-1}(t_0)\hat{\mathbf{x}}(t_0)+\mathbf{M}(t_0)^T\mathbf{H}(t_i)^T\mathbf{R}^{-1}\mathbf{y}(t_i). \tag{15}$$

This proves that $\hat{\mathbf{x}}^s(t_0) = \hat{\mathbf{x}}^{4DVAR}(t_0)$.

1.2 A simple sequential estimation problem

When a system described by a number of quantities evolving in time is not known exactly, e.g., because our deterministic model of the system is not perfect or because the initial conditions are only known approximately, we would like to improve our knowledge of the system by complementing our information using a measuring device, also affected by errors. From a Bayesian point of view, the aim of our estimation strategy is to determine the evolution of the *conditional probability density* of the quantities describing the state of the system, that is the probability density of the quantities conditioned on a set of available measurements (up to a given time t).

A filter is a recursive algorithm that process the available data in such a way to reduce the noise of the data and propagates the conditional density of the state of the system. An optimal estimate may be the mean or the mode of the conditional density. The Kalman filter provides an estimate of the evolution of the conditional density for a *linear model* and a set of observations, both with *white* and *Gaussian* errors. The optimal estimate provided by the Kalman filter, which minimizes the variance of the estimation errors, coincides with the mean as well as the mode (and the median) of the conditional density.

Let us now consider a simple example. Suppose we want to determine the position of an object moving along the direction defined by the x axis. We also know that the speed of the object is approximately constant, so that we can write

$$\frac{dx}{dt} = u + w \tag{16}$$

where u is the nominal velocity and the noise w represents the fact that the nominal velocity is only approximately equal to the actual speed of the object. We will assume w to be a white Gaussian noise, with zero mean and variance σ_w^2. Also assume that at

time t_0 the best estimate of the position of the object was $\hat{x}(t_0)$ with uncertainty $\sigma_x^2(t_0)$. At time t_1 we have

$$x(t_1) = x(t_0) + (u + w)(t_1 - t_0) \tag{17}$$

so that the mean and variance of $x(t_1)$ are

$$\begin{aligned} \hat{x}(t_1) &= \hat{x}(t_0) + u(t_1 - t_0) && (18)\\ \sigma_x^2(t_1) &= E\{(x(t_1) - \hat{x}(t_1))^2\} = \sigma_x^2(t_0) + (t - t_0)^2\sigma_w^2. && (19) \end{aligned}$$

Assume that at t_1 we have a measurement of the position with value y and variance σ_y^2. We want to combine y and $\hat{x}(t_1)$ to get the best estimate of position at time t_1. As seen before for a multi-dimensional case, the combined estimate $\hat{x}^a(t_1)$ can be determined as

$$\hat{x}^a(t_1) = (\sigma_x^{-2}(t_1) + \sigma_y^{-2})^{-1}(\sigma_x^{-2}(t_1)\hat{x}(t_1) + \sigma_y^{-2}y) \tag{20}$$

which can be rewritten as

$$\hat{x}^a(t_1) = \hat{x}(t_1) + \frac{\sigma_x^2(t_1)}{\sigma_x^2(t_1) + \sigma_y^2}(y - \hat{x}(t_1)) = \hat{x}(t_1) + K(t_1)(y - \hat{x}(t_1)). \tag{21}$$

Note that when $\sigma_x^2(t_1)$ is large $K \to 1$ so that $\hat{x}^a(t_1) \to y$, while when σ_y^2 is large $K \leftarrow 0$ and $\hat{x}^a(t_1) \to \hat{x}(t_1)$. This suggests that the Kalman filter provides a reasonable answer, although we have not yet proved it provides the optimal estimate.

Before attempting to describe mathematically systems affected by random noise, we will first discuss deterministic system models in the time domain.

1.3 Dynamic system models

First consider a linear, time-invariant, single input single output system model, that can be described by the linear, constant coefficient ordinary, n-th order differential equation

$$\frac{d^n y(t)}{dt^n} + a_{n-1}\frac{d^{n-1}y(t)}{dt^{n-1}} + \cdots + a_0 y(t) = u(t) \tag{22}$$

where t is the independent variable, $y(t)$ is the dependent or output variable and $u(t)$ is the input. The system is time-invariant (or autonomous) because the coefficients $a_i(t)$ are constant. Let us now define

$$\begin{aligned} x_1(t) &\equiv y(t) && (23)\\ x_2(t) &\equiv \frac{dy(t)}{dt} = \frac{dx_1(t)}{dt} && (24)\\ &\vdots && (25)\\ x_n(t) &\equiv \frac{d^{n-1}y(t)}{dt^{n-1}} = \frac{dx_{n-1}(t)}{dt} && (26) \end{aligned}$$

The *state space representation* of Eq. 22 is a first order vector differential equation with associated output relation:

$$\begin{aligned} \frac{d\mathbf{x}(t)}{dt} &= \mathbf{A}\mathbf{x}(t) + \mathbf{b}u(t) && (27)\\ y(t) &= \mathbf{h}^T\mathbf{x} && (28)\\ & && (29) \end{aligned}$$

where $\mathbf{x}(t) = (x_1(t), x_2(t), \cdots, x_n(t))^T$,

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{pmatrix}, \tag{30}$$

$\mathbf{b} = (0, 0, \cdots, 0, 1)^T$ and $\mathbf{h} = (1, 0, \cdots, 0, 0)^T$. The state vector $\mathbf{x}$ completely describes the system at a given time t so that the knowledge of the input at time $\tau \geq t$ as well as of the initial conditions $\mathbf{x}(t_0) = \mathbf{x}_0$ allow us to know the state of the system at time $\tau \geq t$. It is important to note that the state space representation is not unique, as we can always define a new state vector $\mathbf{x} = \mathbf{T}\mathbf{x}'$ where $\mathbf{T}$ is a non-singular matrix so that we get a system of equations similar to Eq. 27, with $\mathbf{A}' = \mathbf{T}^{-1}\mathbf{A}\mathbf{T}$, $\mathbf{b}' = \mathbf{T}^{-1}\mathbf{b}$ and $\mathbf{h}'^T = \mathbf{h}^T\mathbf{T}$. The transformation of $\mathbf{A}$ to the matrix $\mathbf{A}'$ is a *similarity transformation*, which keeps the same eigenvalues so that the dynamics of the system is unchanged.

As an example of dynamical system, consider a container that moves frictionless with acceleration $-u(t)$ along the x-axis. Assume that a mass m is connected to the left wall of the container by a spring with spring constant k and rest position in $x = 0$ and by a damper with damping coefficient c. The forces on the mass m are given by $-kx - c(dx/dt)$, while the acceleration is $d^2x/dt^2 - u(t)$. We can then write

$$m\frac{d^2x}{dt^2} + c\frac{dx}{dt} + kx = mu(t). \tag{31}$$

The state space representation can be given by defining $\mathbf{x} = (x, dx/dt)^T$ so that

$$\frac{d\mathbf{x}}{dt} = \begin{pmatrix} 0 & 1 \\ -k/m & -c/m \end{pmatrix}\mathbf{x} + \begin{pmatrix} 0 \\ u(t) \end{pmatrix}. \tag{32}$$

The output is given by $y(t) = x(t)$. This is an example of second-order system dynamics as two state variable are sufficient to describe the system completely.

In general, a linear state space system can be expressed as

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \tag{33}$$

$$\mathbf{y}(t) = \mathbf{H}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) \tag{34}$$

$$\tag{35}$$

with $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times p}$, $\mathbf{H} \in \mathbb{R}^{m\times n}$ and $\mathbf{D} \in \mathbb{R}^{m\times p}$. This linear first-order system of differential equations represents a multiple-input multiple output system. From the linearity it follows that if $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ are solutions, the sum of the two is also a solution of the system of equations. We define $\mathbf{x}_h(t)$ as the solution with zero input (the homogeneous solution) and $\mathbf{x}_p(t)$ the solution with zero initial condition (the particular solution). Let us consider now the homogeneous system

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{A}\mathbf{x}(t). \tag{36}$$

If $\mathbf{A}$ and $\mathbf{x}$ are scalar (denoted as $a \in \mathbb{R}$ and $x \in \mathbb{R}$) the solution is given by $x(t) = e^{at}x(0)$. Let us now define the matrix exponential as

$$e^{\mathbf{X}} = \mathbf{I} + \mathbf{X} + \frac{\mathbf{X}^2}{2} + \frac{\mathbf{X}^3}{3!} + \cdots = \sum_{k=0}^{\infty}\frac{\mathbf{X}^k}{k!} \tag{37}$$

where $\mathbf{X} \in \mathbb{R}^{n\times n}$, where $\mathbf{X}^0 \equiv \mathbf{I}$. For $\mathbf{X} = \mathbf{A}t$ we have that

$$\frac{de^{\mathbf{A}t}}{dt} = \mathbf{A}e^{\mathbf{A}t} \tag{38}$$

so that the homogenous solution of the system of equation is given by

$$\mathbf{x}_h(t) = e^{\mathbf{A}t}\mathbf{x}(0). \tag{39}$$

The complete solution is given by

$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}(0) + \int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}(\tau)d\tau \tag{40}$$

as it can be proven by calculating the derivative of the solution:

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{A}e^{\mathbf{A}t}\mathbf{x}(0) + \int_0^t \mathbf{A}e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}(\tau)d\tau + \mathbf{B}\mathbf{u}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t). \tag{41}$$

It follows that

$$\mathbf{y}(t) = \mathbf{H}e^{\mathbf{A}t}\mathbf{x}(0) + \int_0^t \mathbf{H}e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}(\tau)d\tau + \mathbf{D}\mathbf{u}(t) \tag{42}$$

which shows that $\mathbf{y}(t)$ is linear in both inputs and initial conditions and it represents the solution of a system of linear differential equations with constant coefficients.

Let us consider now the case of a time-varying or non-autonomous system, described as

$$\begin{aligned} \frac{d\mathbf{x}(t)}{dt} &= \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) && (43)\\ \mathbf{y}(t) &= \mathbf{H}(t)\mathbf{x}(t) + \mathbf{D}(t)\mathbf{u}(t) && (44)\\ & && (45) \end{aligned}$$

with $\mathbf{x}(t_0) = \mathbf{x}_0$. If $\mathbf{A}(t)$ and $\mathbf{B}(t)\mathbf{u}(t)$ are piecewise continuous we can write the solution as

$$\mathbf{x}(t) = \mathbf{\Phi}(t, t_0)\mathbf{x}(0) + \int_0^t \mathbf{\Phi}(t,\tau)\mathbf{B}(\tau)\mathbf{u}(\tau)d\tau \tag{46}$$

where $\mathbf{\Phi}(t, t_0) \in \mathbb{R}^{n\times n}$ is the *state transition matrix*, solution of

$$\frac{d\mathbf{\Phi}(t, t_0)}{dt} = \mathbf{A}(t)\mathbf{\Phi}(t, t_0) \tag{47}$$

and initial condition $\mathbf{\Phi}(t_0, t_0) = \mathbf{I}$. The transition matrix determines $\mathbf{x}(t)$ provided we know $\mathbf{x}(0)$ and no inputs are fed into the system. We have

$$\begin{aligned} \mathbf{x}(t_1) &= \mathbf{\Phi}(t_1, t_0)\mathbf{x}(0) && (48)\\ \mathbf{x}(t_2) &= \mathbf{\Phi}(t_2, t_1)\mathbf{x}(1) = \mathbf{\Phi}(t_2, t_1)\mathbf{\Phi}(t_1, t_0)\mathbf{x}(0) && (49)\\ & && (50) \end{aligned}$$

so that

$$\mathbf{\Phi}(t_2, t_0) = \mathbf{\Phi}(t_2, t_1)\mathbf{\Phi}(t_1, t_0) \tag{51}$$

independent of the order of t_0, t_1 and t_2. It follows that

$$\mathbf{\Phi}(t, t) = \mathbf{\Phi}(t, t_0)\mathbf{\Phi}(t_0, t) = \mathbf{I} \tag{52}$$

which means that $\mathbf{\Phi}(t,t_0)$ is non-singular, with

$$\mathbf{\Phi}(t,t_0)^{-1}=\mathbf{\Phi}(t_0,t). \tag{53}$$

Note that the time-invariant case (constant $\mathbf{A}$) correspond to a stationary transition matrix, with $\mathbf{\Phi}(t,t_0)=\mathbf{\Phi}(t-t_0)=e^{\mathbf{A}(t-t_0)}$.

To prove that the solution for a non-autonomous system is given by Eq. 46 we can see whether it satisfies Eq. 43. To this end, recall the Lebnitz's rule for differentiation under the integral sign: Let $\boldsymbol{\phi}(t)$ be defined as

$$\boldsymbol{\phi}(t)=\int_{u(t)}^{v(t)}\mathbf{f}(t,\tau)d\tau \tag{54}$$

so that we have

$$\frac{d\boldsymbol{\phi}(t)}{dt}=\int_{u(t)}^{v(t)}\frac{\partial\mathbf{f}(t,\tau)}{\partial t}d\tau+\mathbf{f}(t,v(t))\frac{dv(t)}{dt}-\mathbf{f}(t,u(t))\frac{du(t)}{dt} \tag{55}$$

A nonlinear model of the state of a system can be expressed as

$$\begin{aligned}\frac{d\mathbf{x}(t)}{dt} &= \mathbf{f}[\mathbf{x}(t),\mathbf{u}(t),t] &(56)\\ \mathbf{y}(t) &= H[\mathbf{x}(t),\mathbf{u}(t),t] &(57)\\ . & &(58)\end{aligned}$$

Now assume that we have found $\mathbf{x}_0(t)$, solution of Eq. 56 with initial condition $\mathbf{x}_0(t_0)$ and given input $\mathbf{u}_0(\tau)$ for $\tau\in[t_0,t]$. Now consider the solution that we find when the initial condition is perturbed by a "small" amount $\Delta\mathbf{x}(t_0)$ and the input function by a "small" $\Delta\mathbf{u}_0(\tau)$. For the perturbed solution $\mathbf{x}(t)=\mathbf{x}_0(t)+\Delta\mathbf{x}(t)$ we can write

$$\frac{d\mathbf{x}(t)}{dt}=\mathbf{f}[\mathbf{x}(t),\mathbf{u}(t),t]=\mathbf{f}[\mathbf{x}_0(t),\mathbf{u}_0(t),t]+\mathbf{A}(t)[\mathbf{x}(t)-\mathbf{x}_0(t)]+\mathbf{B}(t)[\mathbf{u}(t)-\mathbf{u}_0(t)]+\cdots \tag{59}$$

where $\mathbf{u}(t)=\mathbf{u}_0(t)+\Delta\mathbf{u}(t)$, $\mathbf{A}(t)=\frac{d\mathbf{f}}{d\mathbf{x}}(\mathbf{x}_0(t),\mathbf{u}_0(t),t)$ and $\mathbf{B}(t)=\frac{d\mathbf{f}}{d\mathbf{u}}(\mathbf{x}_0(t),\mathbf{u}_0(t),t)$. By definition we have that $\frac{d\mathbf{x}_0(t)}{dt}=\mathbf{f}[\mathbf{x}_0(t),\mathbf{u}_0(t),t]$ so that we can write

$$\frac{d[\mathbf{x}(t)-\mathbf{x}_0(t)]}{dt}=\mathbf{A}(t)[\mathbf{x}(t)-\mathbf{x}_0(t)]+\mathbf{B}(t)[\mathbf{u}(t)-\mathbf{u}_0(t)]+\cdots \tag{60}$$

If we only retain first order terms and define $\boldsymbol{\epsilon}(t)=\mathbf{x}(t)-\mathbf{x}_0(t)$ and $\delta\mathbf{u}(t)=\mathbf{u}(t)-\mathbf{u}_0(t)$ we can write

$$\frac{d\boldsymbol{\epsilon}(t)}{dt}=\mathbf{A}(t)\boldsymbol{\epsilon}(t)+\mathbf{B}(t)\delta\mathbf{u}(t) \tag{61}$$

which is a linear non-autonomous system of equations for the perturbations. Its solution can formally be written as

$$\boldsymbol{\epsilon}(t)=\mathbf{\Phi}(t,t_0)\boldsymbol{\epsilon}(0)+\int_0^t\mathbf{\Phi}(t,\tau)\mathbf{B}(\tau)\delta\mathbf{u}(\tau)d\tau. \tag{62}$$

Let us now consider the homogenous solution $\boldsymbol{\epsilon}_h(t)$ we get when $\delta\mathbf{u}(\tau)=\mathbf{0}$, that is, when the input function is not perturbed so that $\mathbf{u}(t)=\mathbf{u}_0(t)$. If we express $\mathbf{\Phi}(t,t_0)$ in terms

of its singular vector decomposition $\mathbf{\Phi}(t,t_0)=\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}^T$ (note that we omit to indicate that the singular values and vectors all depend on (t,t_0)) we can write

$$\boldsymbol{\epsilon}_h(t)=\sum_i(\mathbf{v}_i^T\boldsymbol{\epsilon}_h(t_0))\lambda_i^{1/2}\mathbf{u}_i. \tag{63}$$

This means that the perturbation at time t_0 in the direction of the i-th right singular vector evolves at time t into a perturbation in the direction of the i-th left singular vector, scaled by the i-th singular value. The square magnitude of the perturbation at time t is given by

$$||\boldsymbol{\epsilon}_h(t)||^2=\sum_i(\mathbf{v}_i^T\boldsymbol{\epsilon}_h(t_0))^2\lambda_i. \tag{64}$$

Note that if $\boldsymbol{\epsilon}_h(t_0)=\mathbf{v}_i$ then $||\boldsymbol{\epsilon}_h(t)||^2=\lambda_i$. It follows that the component of the perturbation at time t_0 that grows fastest is along $\mathbf{v}_0$, the right singular vector corresponding to the largest singular value. In this case, at time t the perturbation is in the direction of $\mathbf{\Phi}(t,t_0)\mathbf{v}_0=\lambda_0^{1/2}\mathbf{u}_0$, so that the original perturbation is dilated or contracted by a factor $\lambda_0^{1/2}$. In other words, if the perturbation at time t_0 is contained in a ball of unit radius, the perturbation at time t will be contained in a ellipsoid with axes in the direction of $\mathbf{u}_i$ (or, equivalently, of $\mathbf{\Phi}(t,t_0)\mathbf{v}_i$) and length $\lambda_i^{1/2}$. In analogous manner, we can write

$$\boldsymbol{\epsilon}_h(0)=\mathbf{\Phi}_h^{-1}(t,t_0)\boldsymbol{\epsilon}(t)=\mathbf{\Phi}_h(t_0,t)\boldsymbol{\epsilon}(t)=\mathbf{V}\mathbf{\Lambda}^{-1/2}\mathbf{U}^T\boldsymbol{\epsilon}(t)=\sum_i(\mathbf{u}_i^T\boldsymbol{\epsilon}_h(t))\lambda_i^{-1/2}\mathbf{v}_i. \tag{65}$$

Note also that

$$\mathbf{\Phi}_h^{-1}(t,t_0)\mathbf{u}_i = \mathbf{V}\mathbf{\Lambda}^{-1/2}\mathbf{U}^T\mathbf{u}_i=\lambda_i^{-1/2}\mathbf{v}_i \tag{66}$$

$$\mathbf{\Phi}_h^{T}(t,t_0)\mathbf{u}_i = \mathbf{V}\mathbf{\Lambda}^{1/2}\mathbf{U}^T\mathbf{u}_i=\lambda_i^{1/2}\mathbf{v}_i \tag{67}$$

$$. \tag{68}$$

This means that a perturbation that at time t is contained in a ball of unit radius originates from a perturbation at time t_0 contained in a ellipsoid with axes in the direction of $\mathbf{v}_i$ (or, equivalently, of $\mathbf{\Phi}(t,t_0)^{-1}\mathbf{u}_i$, or of $\mathbf{\Phi}_h^T(t,t_0)\mathbf{u}_i$) and length $\lambda_i^{-1/2}$. For this reason, the right singular vectors of $\mathbf{\Phi}(t,t_0)$ (eigenvectors of $\mathbf{\Phi}(t,t_0)^T\mathbf{\Phi}(t,t_0)$) are also called forward singular vectors, while the left singular vectors of $\mathbf{\Phi}(t,t_0)$ (eigenvectors of $\mathbf{\Phi}(t,t_0)\mathbf{\Phi}(t,t_0)^T$) are also called backward singular vectors.

The quantity σ_i defined as

$$\sigma_i=\frac{1}{t-t_0}\ln(||\mathbf{\Phi}_h(t,t_0)\mathbf{v}_i||)=\frac{1}{t-t_0}\ln(\lambda_i^{1/2}) \tag{69}$$

is the *finite-time Lyapunov exponent*. The reason of its name can be understood when we assume that $\mathbf{\Phi}_h(t,t_0)$ is a symmetric positive definite matrix that characterizes a stationary system. In this case $\mathbf{A}(t)=\mathbf{A}$ where the eigenvalues γ_i of $\mathbf{A}$ are time-independent. Also, in this case the singular vector and the eigenvector decomposition coincide, so that $||\mathbf{\Phi}_h(t,t_0)\mathbf{v}_i||=\exp(\gamma_i(t-t_0))$, and $\sigma_i=\gamma_i$. An important theorem (Oseledec's Theorem) states that $\lim_{t\to\infty}\sigma_i$ evolves to the i-th Lyapunov exponent, which is finite and unique for almost all linearization trajectory $\mathbf{x}_0(t)$ (which is uniquely determined by $\mathbf{x}_0(t_0)$). When the largest Lyapunov exponent (denoted as σ_∞) is positive, the perturbation $\boldsymbol{\epsilon}_h(t)$ grows exponentially: the system exhibits *sensitive dependence on initial conditions*. In this case, two trajectories that are very close at time t_0 will eventually diverge from each

other and will evolve towards states that are anywhere on the attractor (loosely defined as the set where all neighbouring trajectories converge): the exponential growth then stops when the perturbation is comparable to the "diameter" of the attractor. For such a system it is impossible to provide long-range predictions: even if we estimate the initial state with a very small error $\boldsymbol{\epsilon}_h(t_0)$, the error will double in a time $\Delta t = t_1 - t_0 = \ln(2)/\sigma_\infty$. We can't predict for much longer beyond Δt.

We conclude with a working definition of *chaos*: an aperiodic long-term behaviour (i.e., initial conditions will evolve to an attractor that is not a fixed point or a periodic or quasi-periodic orbit) in a deterministic system (i.e., without random inputs) that exhibits sensitive dependence on initial conditions.

2 Probability spaces and random variables

2.1 Fundamental concepts

Let us consider a system (e.g., the atmosphere) described by a set of equations, which provide us with a *model* describing the evolution of the system (say, the atmosphere). The model is inevitably inaccurate and the initial (or boundary) conditions also are uncertain. Any observation of the state of the system is made with instruments that are affected by errors. For these reasons, a deterministic model of the evolution of the state cannot provide an adequate description of the system, which must be instead given in probabilistic terms. Here we introduce some fundamental concepts related to probability theory.

Let Ω be the *sample space* containing all possible *elementary outcomes* ω of a given experiment. An *event* A is a subset of the sample space containing a collection of possible outcomes. If $\omega \in A \subseteq \Omega$, the event is said to occur. A σ-algebra $\mathcal{F}$ on Ω is a family of subsets of Ω such that

- $0 \in \Omega$
- If $A_i \in \mathcal{F}$ then the complement of A_i, $A^C = A \setminus C \in \Omega$
- If $A_1, A_2, \cdots \in \mathcal{F}$ then $\cup_{i=0}^{\infty} A_i \in \mathcal{F}$

The pair $(\Omega, \mathcal{F})$ is a measurable space. A probability measure P on $(\Omega, \mathcal{F})$ is a function $P : \mathcal{F} \to [0, 1]$ such that

- $P(\emptyset) = 0$, $P(\Omega) = 1$
- if $A_1, A_2, \cdots \in \mathcal{F}$ and $A_1, A_2, \cdots$ are disjoint (i.e., $A_i \cap A_j = \emptyset$ for $i \neq j$) then

$$P\left(\bigcup_{i=1}^{\infty}\right) = \sum_{i=1}^{\infty} P(A_i). \tag{70}$$

The triple $(\Omega, \mathcal{F}, P)$ is called a probability space. A *random variable* is a function $\mathrm{x} : \Omega \to \mathbb{R}$ which assigns a value x to a given outcome ω of an experiment, $\mathrm{x}(\omega) = x$ (where x is called a *realization*), such that for every $A = \{\omega : \mathrm{x}(\omega) \leq x\}$ we have that $A \in \mathcal{F}$ for any $x \in \mathbb{R}$. From the definition of random variable follows that $\mathrm{x}^{-1}((-\infty, x]) = A \in \mathcal{F}$ so that it is always possible to define the probability $P(A)$ for a random variable to have values in $(-\infty, x]$. Random vectors $\mathbf{x}(\omega) = \boldsymbol{x} \in \mathbb{R}^n$ are defined in a similar way, with

$A = \{\omega : \mathrm{x}_i(\omega) \leq x_i; i = 1, 2, \cdots, n\}$. From the definition it follows that random vectors and variables are measurable functions.

We can now define the *probability distribution function* $F_{\mathbf{x}}(\boldsymbol{x}) \in \mathbb{R}$ as $F_{\mathbf{x}}(\boldsymbol{x}) = P(\{\omega : \mathbf{x}(\omega) \leq \boldsymbol{x}\})$, representing the probability to find $\mathbf{x} \leq \boldsymbol{x}$. For random vectors, $F_{\mathbf{x}}(\boldsymbol{x})$ is also called the joint probability distribution function of $\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_n$. If $F_{\mathbf{x}}(\boldsymbol{x})$ is differentiable (except perhaps in a countable number of points) it is possible to define a *probability density function* $p_{\mathbf{x}}(\boldsymbol{\xi})$ as

$$F_{\mathbf{x}}(x_1, x_2, \cdots, x_n) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_n} p_{\mathbf{x}}(\xi_1, \xi_2, \cdots, \xi_n)d\xi_1 d\xi_2 \cdots d\xi_n \tag{71}$$

$$= \int_{-\infty}^{\boldsymbol{x}} p_{\mathbf{x}}(\boldsymbol{\xi})d\boldsymbol{\xi}. \tag{72}$$

It follows that $p_{\mathbf{x}}(\boldsymbol{x})$ can be written as

$$p_{\mathbf{x}}(\boldsymbol{x}) = \frac{\partial^n F_{\mathbf{x}}(\boldsymbol{x})}{\partial x_1 \partial x_2 \cdots \partial x_n} \tag{73}$$

with $p_{\mathbf{x}}(\boldsymbol{x}) \geq 0$ and $\int_{-\infty}^{\boldsymbol{x}} p_{\mathbf{x}}(\boldsymbol{\xi})d\boldsymbol{\xi} = 1$.

The *expectation* or *mean* of a random vector $\mathbf{x}$ is defined as

$$E\{\mathbf{x}\} = \int_{-\infty}^{+\infty} \boldsymbol{x} p_{\mathbf{x}}(\boldsymbol{x})d\boldsymbol{x} \tag{74}$$

Let $\boldsymbol{\theta}(\mathbf{x}) \in \mathbb{R}^m$ be a function of the random vector $\mathbf{x} \in \mathbb{R}^n$. The expectation of $\boldsymbol{\theta}(\mathbf{x})$ is given by

$$E\{\boldsymbol{\theta}(\mathbf{x})\} = \int_{-\infty}^{+\infty} \boldsymbol{\theta}(\boldsymbol{x}) p_{\mathbf{x}}(\boldsymbol{x})d\boldsymbol{x}. \tag{75}$$

Note that the expectation $E\{\mathbf{x}\} \equiv \mathbf{m}$ of a random vector is not a random vector. If we now define $\boldsymbol{\theta}(\mathbf{x}) = (\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^T$, its expectation $\mathbf{P}$ is a symmetric positive-semidefinite matrix that can be written as

$$\mathbf{P} = E\{(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^T\} = \int_{-\infty}^{+\infty} (\boldsymbol{x}-\mathbf{m})(\boldsymbol{x}-\mathbf{m})^T p_{\mathbf{x}}(\boldsymbol{x})d\boldsymbol{x} \equiv \mathrm{cov}(\mathbf{x}) \tag{76}$$

and it is called the *covariance matrix* of $\mathbf{x}$. It can be proved that $\mathbf{P} = E\{\mathbf{x}\mathbf{x}^T\} - \mathbf{m}\mathbf{m}^T$.

2.2 Gaussian random variables

A random vector is said to be Gaussian or normally distributed when its probability density function is given by

$$p_{\mathbf{x}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \mathbf{P}}} \exp -[\frac{1}{2}(\boldsymbol{x}-\mathbf{m})^T \mathbf{P}^{-1}(\boldsymbol{x}-\mathbf{m})] \equiv N(\mathbf{m}, \mathbf{P}) \tag{77}$$

where $\mathbf{P}$ is assumed nonsingular, even though the definition of a Gaussian random vector can be extended to include a positive semidefinite $\mathbf{P}$. It can be showed that $\mathbf{m}$ and $\mathbf{P}$ are indeed the mean and covariance of $\mathbf{x}$, so that a Gaussian random vector is uniquely defined when the mean and covariance of $\mathbf{x}$ are known. Another important property of Gaussian vectors is the following: if $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian, with density $p_{\mathbf{x},\mathbf{y}}(\boldsymbol{x},\boldsymbol{y})$, and uncorrelated (i.e., $\mathrm{cov}(\mathbf{x}, \mathbf{y}) = 0$), they are also *independent*, so that we can write

$$p_{\mathbf{x},\mathbf{y}}(\boldsymbol{x},\boldsymbol{y}) = p_{\mathbf{x}}(\boldsymbol{x}) p_{\mathbf{y}}(\boldsymbol{y}). \tag{78}$$

2.3 Conditional probability

The conditional probability density of $\mathbf{x}$ given $\mathbf{y}$ or $\mathbf{x}|\mathbf{y}$ is the probability density that the random vector $\mathbf{x}$ assumes the realization $\boldsymbol{x}$ conditioned on knowledge that the random vector $\mathbf{y}$ assumes the realization $\boldsymbol{y}$ is defined as

$$p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y}) = \frac{p_{\mathbf{x},\mathbf{y}}(\boldsymbol{x},\boldsymbol{y})}{p_{\mathbf{y}}(\boldsymbol{y})}. \tag{79}$$

This definition is inspired by that for conditional probabilities, stating that the probability of an event A for $\mathbf{x}$ to occur, given an event B for $\mathbf{y}$ has occurred can be written as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{80}$$

where $P(A \cap B)$ is the probability for the event A and B for $\mathbf{x}$ and $\mathbf{y}$, respectively, to occur jointly. Note that if $\mathbf{x}$ and $\mathbf{y}$ are independent, from Eq. 79 it follows that $p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y}) = p_{\mathbf{x}}(\boldsymbol{x})$.

The *conditional expectation* or *conditional mean* of $\mathbf{x}$ given $\mathbf{y}$ has taken the value $\boldsymbol{x}$ can be written as

$$E_{\mathbf{x}}(\mathbf{x}|\mathbf{y} = \boldsymbol{y}) = \int_{-\infty}^{+\infty} \boldsymbol{x} p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x}, \tag{81}$$

where the subscript $\mathbf{x}$ is a reminder that the expectation is sought over the possible realizations of $\mathbf{x}$. Note that the values of the components of the conditional expectation vector depend on the value taken by the random vector $\mathbf{y}$, so that the conditional expectation is a random vector. It is possible to show that

$$E_{\mathbf{y}}(E_{\mathbf{x}}(\mathbf{x}|\mathbf{y} = \boldsymbol{y})) = E_{\mathbf{x}}(\mathbf{x}). \tag{82}$$

From Eq. 79 we can write

$$p_{\mathbf{x},\mathbf{y}}(\boldsymbol{x},\boldsymbol{y}) = p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y}) p_{\mathbf{y}}(\boldsymbol{y}) = p_{\mathbf{y}|\mathbf{x}}(\boldsymbol{y}|\boldsymbol{x}) p_{\mathbf{x}}(\boldsymbol{x}) \tag{83}$$

from which the *Bayes' rule* follows:

$$p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y}) = \frac{p_{\mathbf{y}|\mathbf{x}}(\boldsymbol{y}|\boldsymbol{x}) p_{\mathbf{x}}(\boldsymbol{x})}{p_{\mathbf{y}}(\boldsymbol{y})}. \tag{84}$$

Eq. 84 can also be written as

$$p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y}) = \frac{p_{\mathbf{y}|\mathbf{x}}(\boldsymbol{y}|\boldsymbol{x}) p_{\mathbf{x}}(\boldsymbol{x})}{\int_{-\infty}^{+\infty} p_{\mathbf{x},\mathbf{y}}(\boldsymbol{x},\boldsymbol{y}) d\boldsymbol{x}} = \frac{p_{\mathbf{y}|\mathbf{x}}(\boldsymbol{y}|\boldsymbol{x}) p_{\mathbf{x}}(\boldsymbol{x})}{\int_{-\infty}^{+\infty} p_{\mathbf{y}|\mathbf{x}}(\boldsymbol{y}|\boldsymbol{x}) p_{\mathbf{x}}(\boldsymbol{x}) d\boldsymbol{x}}. \tag{85}$$

3 Stochastic processes

A stochastic process is a family of random vectors $\{\mathbf{x}_t\}_{t \in T} \in \mathbb{R}^n$ defined on a probability space, where T is the parameter space. At a given $t \in T$ the process is represented by the random vector $\mathbf{x}_t$, while for a given $\omega \in \Omega$ the stochastic process represents a *path* or *sample* in parameter space. We will assume that $T \subseteq \mathbb{R}$ and that t represents time. A stochastic process can be defined over a discrete number of times (discrete-time stochastic process) or over a interval of $\mathbb{R}$ (continuous-time stochastic process). The probability density function of a stochastic process defined at times $t_1, t_2, \cdots, t_N$ is defined as $p_{\mathbf{x}_{t_1},\mathbf{x}_{t_2},\cdots,\mathbf{x}_{t_N}}(\boldsymbol{x}_{t_1}, \boldsymbol{x}_{t_2}, \cdots, \boldsymbol{x}_{t_N})$.

3.1 Markov processes and Brownian motion

A stochastic process $\mathbf{x}_t(\omega)$ is a *Markov process* if for any $t_1, t_2, \cdots, t_n \in T$ we can write

$$p_{\mathbf{x}_{t_n}|\mathbf{x}_{t_1},\mathbf{x}_{t_2},\cdots,\mathbf{x}_{t_{n-1}}}(\boldsymbol{x}_{t_n}|\boldsymbol{x}_{t_1},\boldsymbol{x}_{t_2},\cdots,\boldsymbol{x}_{t_{n-1}}) = p_{\mathbf{x}_{t_n}|\mathbf{x}_{t_{n-1}}}(\boldsymbol{x}_{t_n}|\boldsymbol{x}_{t_{n-1}}). \tag{86}$$

This means that the value assumed by the random variable at time t_{n-1} completely determines the probability density of the random variable at time t_n: generalized causality principle. Markov processes are of fundamental importance for the study of stochastic dynamical systems. A stochastic process $\mathbf{x}_t(\omega)$ is said to be *white Gaussian* if $\mathbf{x}_{t_1}, \mathbf{x}_{t_1}, \cdots, \mathbf{x}_{t_n}$ are independent Gaussian random vectors, for any $t_1, t_2, \cdots, t_n \in T$. From the independence property follows that $\text{cov}(\mathbf{x}_{t_i}, \mathbf{x}_{t_j}) = 0$ for $i \neq j$. For a continuous-time independent process we can write

$$\text{cov}(\mathbf{x}_{t_i}, \mathbf{x}_{t_j}) = \mathbf{Q}(t_i)\delta(t_i - t_j) \tag{87}$$

where $\mathbf{Q}(t_i)$ is the covariance matrix of $\mathbf{x}_{t_i}$ and $\delta(t_i - t_j)$ is the Dirac delta function. A continuous-time independent process is also referred to as *delta correlated* process. It follows that the Fourier transform of the covariance of a delta-correlated process is constant given by $\mathbf{Q}(t_i)$. This means that the power spectral density is constant over all frequencies and that is why a delta-correlated process is also called white process. A process that is both Gaussian and Markovian is called Gauss-Markov process.

A process $\mathbf{x}_t(\omega)$ is said to have *independent increments* if, for any $t_1, t_2, \cdots, t_n \in T$, the increments $\mathbf{x}_1 - \mathbf{x}_0, \mathbf{x}_2 - \mathbf{x}_1, \cdots, \mathbf{x}_n - \mathbf{x}_{n-1}$ are mutually independent. A *Brownian motion* or *Wiener* process is a process with a) independent Gaussian increments; b) $E\{\mathbf{x}(t)\} = \mathbf{0}$; c) $E\{[\mathbf{x}(t_2) - \mathbf{x}(t_1)][\mathbf{x}(t_2) - \mathbf{x}(t_1)]^T\} = \int_{t_1}^{t_2} \mathbf{Q}(t)dt$ with $t_2 \geq t_1$ and where $\mathbf{Q}$ is a covariance matrix; d) $P(\mathbf{x}(t_0) = 0) = 1$. Note that the formal (as it can not be defined properly) time derivative of the Brownian motion is a zero-mean white Gaussian process, with covariance $\mathbf{Q}(t_i)\delta(t_i - t_j)$. Also note that from c) it follows that the covariance of a Brownian motion with $\mathbf{Q}(t) = \mathbf{Q}_0$ is equal to $\mathbf{Q}_0(t_2 - t_1)$, so that its variance increases linearly with time. [Exercise: write the code to generate a Brownian motion using a random number generator].

3.2 Stochastic differential equations

Let us consider a time-continuous dynamical system with state $\mathbf{x}_t \in \mathbb{R}^n$. Let us assume that the system is described by a set of differential equations given by

$$\frac{d\mathbf{x}_t}{dt} = f(\mathbf{x}_t, t) + \mathbf{G}_t\mathbf{w}_t \tag{88}$$

where $\mathbf{G}_t \in \mathbb{R}^{n\times m}$ and $\mathbf{w}_t \in \mathbb{R}^m$ is a white Gaussian process for $t \geq t_0$ that is independent of the initial conditions $\mathbf{x}_{t_0}$. However, $\mathbf{w}_t$ is not integrable as it is a delta-correlated process, and Eq. 88 cannot be solved. In order to overcome this difficulty, it is possible to express $\mathbf{w}_t$ as $\mathbf{w}_t = d\boldsymbol{\beta}_t dt$, where $\boldsymbol{\beta}_t$ is a Brownian motion and rewrite Eq. 88 as

$$d\mathbf{x}_t = f(\mathbf{x}_t, t)dt + \mathbf{G}_t d\boldsymbol{\beta}_t. \tag{89}$$

From Eq. 89 it follows

$$\mathbf{x}_t = \mathbf{x}_{t_0} + \int_{t_0}^{t} f(\mathbf{x}_s, s)ds + \int_{t_0}^{t} \mathbf{G}_s d\boldsymbol{\beta}_s. \tag{90}$$

The integral $\int_{t_0}^{t} f(\mathbf{x}_s, s)ds$ can be defined as an ordinary integral for a given realization of $\mathbf{x}$. However, for Eqs. 89 and 90 to be meaningful, $\int_{t_0}^{t} \mathbf{G}_s d\boldsymbol{\beta}_s$ must also exist. To this end, we can express it as

$$\int_{t_0}^{t} \mathbf{G}_s d\boldsymbol{\beta}_s = \lim_{j\to\infty} \sum_{j=1}^{n} \mathbf{G}_{t_j^*}(\boldsymbol{\beta}_{t_{j+1}} - \boldsymbol{\beta}_{t_j}). \tag{91}$$

It can be shown that the result of this stochastic integral depends on the choice of t^*. Two choices are usually made: a) $t_j^* = t_j$, defining the *Itô integrals*; b) $t_j^* = (t_j + t_{j+1})/2$, defining the *Stratonovich integrals*. For both choices a) and b) it is possible to prove the existence of the limit at the r.h.s. of Eq. 91, but, in general, the interpretation of the stochastic integral produce different results. In particular, the Stratonovich integral produces solutions $\mathbf{x}_t$ of Eq. 88 that are consistent with the solutions obtained with a deterministic set of ordinary differential equations derived from Eq. 90 for a given realization of $\mathbf{x}_t$. In this course we will use the Itô interpretation of the stochastic integral shown in Eq. 91 so that Eq. 90 is to be intended as a Itô stochastic differential equation. It is possible to show that the solutions $\mathbf{x}_t$ determined by using the Stratonovich interpretation of the stochastic integral are solutions of the *modified Itô equation* given by

$$\mathbf{x}_t = \mathbf{x}_{t_0} + \int_{t_0}^{t} f(\mathbf{x}_s, s)ds + \frac{1}{2}\int_{t_0}^{t} \mathbf{G}_s' \mathbf{G}_s ds + \int_{t_0}^{t} \mathbf{G}_s d\boldsymbol{\beta}_s \tag{92}$$

where $\mathbf{G}_s'$ is the derivative of $\mathbf{G}_s$ with respect to $\mathbf{x}_t$. However, if $\mathbf{G}_s$ does not depend on $\mathbf{x}_t$, the two solutions coincide.

Finally, note that $\mathbf{x}_t$ in Eq. 90 depends on $\mathbf{x}_{t_0}$ and on $\{d\boldsymbol{\beta}_s,\, t_0 \le s \le t\}$. Given that $\boldsymbol{\beta}_s$ is a Brownian motion, $\boldsymbol{\beta}_s$ is independent of $\mathbf{x}_t$ for $s \le t_0$ so that $\mathbf{x}_t$ is Markov.

3.3 Kolmogorov's forward equation

Note that that $\boldsymbol{\beta}_s$ is independent of $\mathbf{x}_\tau$ for $\tau \le t_0$ it follows that the process generated by the stochastic differential equation 89 is a Markov process, with $p(\mathbf{x}_t|\mathbf{x}_\tau, \tau \le t_0) = p(\mathbf{x}_t|\mathbf{x}_{t_0})$. The conditional probability density function $p(\boldsymbol{x}_t|\boldsymbol{x}_{t_0})$ for the process $\mathbf{x}_t$ given in Eq. 90 is given by

$$\frac{\partial p(\boldsymbol{x}_t|\boldsymbol{x}_{t_0})}{\partial t} = -\sum_{i=1}^{n} \frac{\partial}{\partial x_i(t)} p(\boldsymbol{x}_t|\boldsymbol{x}_{t_0}) M_i(\mathbf{x}_t, t) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2}{\partial x_i(t) x_j(t)} [p(\boldsymbol{x}_t|\boldsymbol{x}_{t_0})\{\mathbf{G}_t \mathbf{Q}_t \mathbf{G}_t^T\}_{ij}] \tag{93}$$

where $\mathbf{Q}_t$ is the covariance of the Brownian motion $\boldsymbol{\beta}_t$, M_i is the i-th component of M and where we have assumed the existence of the continuous partial derivatives in Eq. 93. The initial condition at $t = t_0$ is $p(\boldsymbol{x}_{t_0}|\boldsymbol{x}_{t_0}) = \delta(\boldsymbol{x} - \boldsymbol{x}_{t_0})$. This is the Kolmogorov's forward equation or the Fokker-Planck equation. As an example, consider the scalar Brownian motion with constant variance q, defined by the stochastic differential equation $dx_t = \sqrt{q} d\beta_t$ for $t \ge 0$ and $p(\beta_{t_0} = 0) = 1$). The Fokker-Planck equation in this case becomes $\frac{\partial p(x_t|x_{t_0})}{\partial t} = \frac{1}{2}\frac{\partial^2 p(x_t|x_{t_0})}{\partial x_t^2}$ or $t \ge 0$ and $p(x_{t_0}|x_{t_0}) = \delta(x)$. This is the scalar heat equation and this is why a Markov process whose conditional probability density satisfies the Kolmogorov's forward equation is said to be a *diffusion* process. The solution of the equation is in this case given by $p(x_t|x_{t_0}) = (1/2q\pi t)^{1/2} \exp(-x^2/2qt)$, a Gaussian with variance equal to qt, which increases linearly as a function of time consistently with the properties of a Brownian motion. Note that, as the conditional probability $p(\boldsymbol{x}_t|\boldsymbol{x}_{t_0})$

is a random variable, from Eq. 82 we can write $p(\boldsymbol{x}_t) = E_{\mathbf{x_{t_0}}}\{p(\boldsymbol{x}_t|\boldsymbol{x}_{t_0})\}$. This means that $p(\boldsymbol{x}_t)$ also satisfies the Kolmogorov's forward equation, with initial conditions $p(\boldsymbol{x}_{t_0})$ assumed known. The probability density function $p(\boldsymbol{x}_t)$ represents the solution of the prediction problem (in the absence of observations at times $\tau < t$).

4 State estimation

In the previous section we have discussed how it is possible to describe a dynamical system affected by random disturbances or errors. In particular, we have seen that the state can be represented as a stochastic process $\mathbf{x}_t$ characterized by a probability density function $p(\boldsymbol{x}_t)$ that evolves in time according to Eq. 93. Now we want to understand how to solve the problem of determining an "optimal" estimate of the state of this system by using a set of available observations, also affected by errors.

First, assume that $\mathbf{x}_t$, defined as the solution of Eq. 89, is observed at times $t \geq t_0$ so that we can write a stochastic differential equation

$$d\mathbf{z}_t = H(\mathbf{x}_t, t)dt + d\boldsymbol{\beta}_t^o \tag{94}$$

where $\boldsymbol{\beta}_t^o$ is a Brownian motion with formal time derivative $\boldsymbol{\epsilon}_t = d\boldsymbol{\beta}_t^o/dt$, where $cov(\boldsymbol{\epsilon}_t) = \mathbf{R}_t\delta(t-\tau)$. We also assume $\boldsymbol{\beta}_t^o$, $\boldsymbol{\beta}_t$ in Eq. 89 and $\mathbf{x}_{t_0}$ to be independent. If we formally define $\mathbf{y}_t$ as $\mathbf{y}_t = d\mathbf{z}_t/dt$ we can rewrite Eq. 94 as

$$\mathbf{y}_t = H(\mathbf{x}_t, t) + \boldsymbol{\epsilon}_t^o \tag{95}$$

where $\boldsymbol{\epsilon}_t$ is a white Gaussian noise with

$$\begin{aligned} E\{\boldsymbol{\epsilon}_{t_i}\} &= \mathbf{0} \\ E\{\boldsymbol{\epsilon}_{t_i}\boldsymbol{\epsilon}_{t_i}^T\} &= \mathbf{R}_{t_i} \\ E\{\boldsymbol{\epsilon}_{t_i}\boldsymbol{\epsilon}_{t_j}^T\} &= \mathbf{0} \qquad t_i \neq t_j. \end{aligned} \tag{96}$$

Note that the joint $\{\mathbf{x}_t, \mathbf{z}_t\}$ process is Markov. Now consider a set of realizations $Y_\tau = \{z_s, t_0 < s \leq \tau\}$. The solution of the problem of estimating $\mathbf{x}_t$ given Y_τ is given by $p(\boldsymbol{x}_t|Y_\tau)$. In particular, $p(\boldsymbol{x}_t|Y_\tau)$ is the solution of the *smoothing problem* (for $t < \tau$), the *filtering problem* (for $t = \tau$) and the *prediction problem* (for $t > \tau$) when past observations are available. In practice we are interested in determining a specific state $\hat{\mathbf{x}}_t$ that represents the best estimate of $\mathbf{x}_t$ given Y_τ, according to some criterion.

4.1 Minimum variance estimation

The estimation error $\boldsymbol{\epsilon}_t \in (R)^n$ is defined as $\boldsymbol{\epsilon}_t = \hat{\mathbf{x}}_t - \mathbf{x}_t$. Consider the quadratic *cost function*

$$J(\boldsymbol{\epsilon}_t) = \boldsymbol{\epsilon}_t^T\mathbf{S}^{-1}\boldsymbol{\epsilon}_t \tag{97}$$

where $\mathbf{S} \in \mathbb{R}^{n\times n}$ is a positive definite matrix and $J(\boldsymbol{\epsilon}_t) \in \mathbb{R}$. Note that $J(\mathbf{0}) = 0$ and that $J(\boldsymbol{\epsilon}_t)$ increases when the magnitude of the estimation error increases. Also note that $J(\boldsymbol{\epsilon}_t)$ is a scalar random variable, as $\boldsymbol{\epsilon}_t$ is a random process. We now want to find the estimate that minimizes the expected cost $E\{J(\boldsymbol{\epsilon}_t)\}$. Note that when $\mathbf{S} = \mathbf{I}$, $E\{J(\boldsymbol{\epsilon}_t)\}$ is the variance of the estimation error – for a generic $\mathbf{S}$, $E\{J(\boldsymbol{\epsilon}_t)\}$ is the variance of $\mathbf{S}^{-1/2}\boldsymbol{\epsilon}_t$ – so that the estimate that minimizes the expected cost is known as the *minimum variance* estimate.

Let us define $\boldsymbol{\mu}_t$ as the conditional mean of $\mathbf{x}_t$ given by $\boldsymbol{\mu}_t = E_{\mathbf{x}_t}\{\mathbf{x}_t|Y_\tau\}$. If we express $E\{J(\boldsymbol{\epsilon}_t)\}$ as $E\{J(\boldsymbol{\epsilon}_t)\} = E_{\mathbf{Y}_\tau}\{E_{\mathbf{x}_t}\{J(\mathbf{x}_t - \hat{\mathbf{x}}_t)|Y_\tau\}\}$ we can write

$$E\{J(\boldsymbol{\epsilon}_t)\} = E_{\mathbf{Y}_\tau}\{E_{\mathbf{x}_t}\{[(\mathbf{x}_t - \boldsymbol{\mu}_t) + (\boldsymbol{\mu}_t - \hat{\mathbf{x}}_t)]^T\mathbf{S}^{-1}[(\mathbf{x}_t - \boldsymbol{\mu}_t) + (\boldsymbol{\mu}_t - \hat{\mathbf{x}}_t)]|Y_\tau\}\} \tag{98}$$

Now note that $\boldsymbol{\mu}_t$ and $\hat{\mathbf{x}}_t$ are only function of Y_τ, so that

$$E_{\mathbf{x}_t}\{(\boldsymbol{\mu}_t - \hat{\mathbf{x}}_t)^T\mathbf{S}^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_t)|Y_\tau\} = (\boldsymbol{\mu}_t - \hat{\mathbf{x}}_t)^T\mathbf{S}^{-1}E_{\mathbf{x}_t}\{(\mathbf{x}_t - \boldsymbol{\mu}_t)|Y_\tau\} = \mathbf{0} \tag{99}$$

so that we can write

$$\begin{aligned} E\{J(\boldsymbol{\epsilon}_t)\} &= E_{\mathbf{Y}_\tau}\{E_{\mathbf{x}_t}\{(\mathbf{x}_t - \boldsymbol{\mu}_t)^T\mathbf{S}^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_t)|Y_\tau\}\} + E_{\mathbf{Y}_\tau}\{E_{\mathbf{x}_t}\{(\boldsymbol{\mu}_t - \hat{\mathbf{x}}_t)^T\mathbf{S}^{-1}(\boldsymbol{\mu}_t - \hat{\mathbf{x}}_t)|Y_\tau\}\} \\ &= E_{\mathbf{Y}_\tau}\{J(\mathbf{x}_t - \boldsymbol{\mu}_t)\} + E_{\mathbf{Y}_\tau}\{J(\boldsymbol{\mu}_t - \hat{\mathbf{x}}_t)\}. \end{aligned} \tag{100}$$

As the first term on the r.h.s. of Eq. 100 does not depend on $\hat{\mathbf{x}}_t$, $E\{J(\boldsymbol{\epsilon}_t)\}$ is minimized when $E_{\mathbf{Y}_\tau}\{J(\boldsymbol{\mu}_t - \hat{\mathbf{x}}_t)\}$ is minimum, i.e., for $\hat{\mathbf{x}}_t = \boldsymbol{\mu}_t$. This proves that the minimum variance estimate is given by the conditional mean. Note that this result is independent of the choice of $\mathbf{S}$ (as long as it is positive definite), and this justifies the fact that bmu_t is called the minimum variance estimate (i.e., for $\mathbf{S} = \mathbf{I}$). Also, this result is independent of any assumptions on the probability density associated to $\mathbf{x}_t$. Another important property of the minimum variance estimate $\boldsymbol{\mu}_t$ is that it is unbiased, or that $E_{\mathbf{Y}_\tau}\{\boldsymbol{\mu}_t\} = E_{\mathbf{x}_t}\{\mathbf{x}_t\}$. This follows directly from Eq. 82.

4.2 Conditional mode estimation

When the conditional density is multimodal, the conditional mean is not necessarily the best estimate of the state. In this case a possible alternative is to calculate the *maximum a posteriori* (MAP) estimate, that is the state than maximizes the conditional probability of the state given a set of observational realizations. Nonlinearity and/or non-Gaussianity makes the MAP estimate different from the minimum variance estimate. Also, the MAP estimate is in general a biased estimate. Nonetheless, the MAP estimate may be easier to calculate. Ideally, when the conditional density is multimodal, we should calculate both the mean and the mode(s) of the conditional density.

4.3 Linear system models

Here we restrict our discussion to *linear* dynamical systems, which are described by the linear stochastic differential equation given by

$$d\mathbf{x}_t = \mathbf{F}_t\mathbf{x}_t dt + \mathbf{G}_t d\boldsymbol{\beta}_t. \tag{101}$$

It is possible to show that the stochastic process $\mathbf{x}_t$ that provides the solution of Eq. 101 is given by

$$\mathbf{x}_t = \boldsymbol{\Phi}(t, t_0)\mathbf{x}_{t_0} + \int_{t_0}^{t} \boldsymbol{\Phi}(t, s)\mathbf{G}_s d\boldsymbol{\beta}_s \tag{102}$$

where the state transition matrix $\boldsymbol{\Phi}(t, t_0)$ is the solution of

$$\frac{d\boldsymbol{\Phi}(t, t_0)}{dt} = \mathbf{F}_t\boldsymbol{\Phi}(t, t_0) \tag{103}$$

with initial condition $\boldsymbol{\Phi}(t_0, t_0) = \mathbf{I}$.

4.4 Discrete linear stochastic-dynamic model

To find a numerical solution of Eq. 102 it is necessary to consider a discrete-time stochastic process to represent the state of the linear system under investigation. The solution at time t_{i+1} can be represented as a linear stochastic difference equation given by

$$\mathbf{x}_{t_{i+1}} = \boldsymbol{\Phi}(t_{i+1}, t_i)\mathbf{x}_{t_i} + \int_{t_i}^{t_{i+1}} \boldsymbol{\Phi}(t_{i+1}, s)\mathbf{G}_s d\boldsymbol{\beta}_s = \boldsymbol{\Phi}(t_{i+1}, t_i)\mathbf{x}_{t_i} + \boldsymbol{\eta}_{t_i} \tag{104}$$

where $\boldsymbol{\eta}_{t_i}$ is a white Gaussian discrete-time process with

$$\begin{aligned} E\{\boldsymbol{\eta}_{t_i}\} &= \mathbf{0} \\ E\{\boldsymbol{\eta}_{t_i}\boldsymbol{\eta}_{t_i}^T\} &= \int_{t_i}^{t_{i+1}} \boldsymbol{\Phi}(t_{i+1}, s)\mathbf{G}_s\mathbf{Q}_s\mathbf{G}_s^T\boldsymbol{\Phi}^T(t_{i+1}, s)ds \equiv \overline{\mathbf{Q}}(t_{i+1}, t_i) \\ E\{\boldsymbol{\eta}_{t_i}\boldsymbol{\eta}_{t_j}^T\} &= \mathbf{0} \qquad t_i \neq t_j. \end{aligned} \tag{105}$$

Note that for a given $\mathbf{x}_{t_i}$, here assumed Gaussian, $\mathbf{x}_{t_{i+1}}$ depends only on $\boldsymbol{\eta}_{t_i}$, which is Gaussian and independent of $\mathbf{x}_{t_{i-1}}, \mathbf{x}_{t_{i-2}}, \cdots, \mathbf{x}_{t_0}$ so that $\mathbf{x}_t$ is a Gauss-Markov sequence.

For time steps $\Delta t_i = t_{i+1} - t_i$ that are short with respect to the variation rate of the system so that for $t_i \leq s \leq t_{i+1}$ we can write $\mathbf{F}_s \simeq \mathbf{F}_{t_i}$ and $\mathbf{G}_s\mathbf{Q}_s\mathbf{G}_s^T \simeq \mathbf{G}_{t_i}\mathbf{Q}_{t_i}\mathbf{G}_{t_i}^T$, we can write

$$\boldsymbol{\Phi}(t_{i+1}, t_i) \simeq \mathbf{I} + \mathbf{F}_{t_i}\Delta t_i. \tag{106}$$

By substituting Eq. 106 in Eq. 105 and retaining only first-order terms in Δt_i we get

$$\overline{\mathbf{Q}}(t_{i+1}, t_i) \simeq \mathbf{G}_{t_i}\mathbf{Q}_{t_i}\mathbf{G}_{t_i}^T\Delta t_i \tag{107}$$

so that, at first order, Eq. 102 can be written as

$$\mathbf{x}_{t_{i+1}} = (\mathbf{I} + \mathbf{F}_{t_i}\Delta t_i)\mathbf{x}_{t_i} + \boldsymbol{\eta}_{t_i} \tag{108}$$

where $\boldsymbol{\eta}_{t_i}$ is a zero-mean white Gaussian discrete-time noise with covariance equal to $\mathbf{G}_{t_i}\mathbf{Q}_{t_i}\mathbf{G}_{t_i}^T\Delta t_i$. This result shows how to derive numerical solutions of Eq. 102: although a first-order approximation may be poor for describing the evolution of the deterministic dynamical system – and in this case it may be better to use a higher-order approximation –, it is usually sufficient to describe the stochastic noise term. Note also that the variance of the noise term increases linearly with time, providing a consistent discrete-time approximation of the continuous-time Brownian motion $\boldsymbol{\beta}_t$.

Finally note that, from Eq. 104, the mean and the covariance of $\mathbf{x}_{i+1}$ can be written as

$$\mathbf{m}_x(t_{i+1}) = \mathbf{M}_{t_i}\mathbf{m}_x(t_i) \tag{109}$$

$$\mathbf{P}_{xx}(t_{i+1}) = \mathbf{M}_{t_i}\mathbf{P}_{xx}(t_i)\mathbf{M}_{t_i}^T + \mathbf{Q}_{t_i} \tag{110}$$

where for simplicity of notation we have renamed $\boldsymbol{\Phi}(t_{i+1}, t_i)$ as $\mathbf{M}_{t_i}$ and $\overline{\mathbf{Q}}(t_{i+1}, t_i)$ as $\mathbf{Q}_{t_i}$.

4.5 The discrete-time Kalman filter

Let us assume that the evolution of the state $\mathbf{x}_t$ at discrete times is described by the linear stochastic difference equation

$$\mathbf{x}_{t_{i+1}} = \mathbf{M}_{t_i}\mathbf{x}_{t_i} + \boldsymbol{\eta}_{t_i} \tag{111}$$

and that a set of measurements $\mathbf{y}_{t_{i+1}}$ are available at a number of times t_{i+1}, for $i = 0, \cdots, n$, so that we can write

$$\mathbf{y}_{t_{i+1}} = \mathbf{H}_{t_{i+1}} \mathbf{x}_{t_{i+1}} + \boldsymbol{\epsilon}^o_{t_{i+1}} \tag{112}$$

where Eqs. 96 and 105 hold for $\boldsymbol{\epsilon}^o_{t_{i+1}}$ and $\boldsymbol{\eta}_{t_i}$, respectively. Also, we assume that at initial time t_0, $\mathbf{x}_{t_0}$ is described by a density function $p(\boldsymbol{x}_{t_0})$ that is assumed Gaussian with mean $\hat{\mathbf{x}}_{t_0}$ and covariance $\mathbf{P}_0$ (not that at $t = t_0$ there are no observations). Finally we assume that $\boldsymbol{\eta}_{t_i}$, $\boldsymbol{\epsilon}^o_{t_{i+1}}$ and $\mathbf{x}_{t_0}$ are mutually independent. As they are all Gaussian, this is equivalent to say that they are assumed mutually uncorrelated.

Let us now define the following quantities:

$$\begin{aligned}
\mathbf{x}^a_{t_i} &= E\{\mathbf{x}_{t_i} | Y(t_i)\} \\
\mathbf{P}^a_{t_i} &= E\{(\mathbf{x}_{t_i} - \mathbf{x}^a_{t_i})(\mathbf{x}_{t_i} - \mathbf{x}^a_{t_i})^T | Y(t_i)\} \\
\mathbf{x}^f_{t_{i+1}} &= E\{\mathbf{x}_{t_{i+1}} | Y(t_i)\} \\
\mathbf{P}^f_{t_{i+1}} &= E\{(\mathbf{x}_{t_{i+1}} - \mathbf{x}^f_{t_{i+1}})(\mathbf{x}_{t_{i+1}} - \mathbf{x}^f_{t_{i+1}})^T | Y(t_i)\}
\end{aligned} \tag{113}$$

where $\mathbf{Y}(t_i) = \{\mathbf{y}_{t_1}, \mathbf{y}_{t_2}, \cdots, \mathbf{y}_{t_i}\}$ and where the conditional mean $\mathbf{x}^a_{t_i}$ is the *analysis* at time t_i, the conditional covariance $\mathbf{P}^a_{t_i}$ is the analysis error covariance at time t_i, $\mathbf{x}^f_{t_{i+1}}$ is the forecast at time t_{i+1} and $\mathbf{P}^f_{t_{i+1}}$ is the forecast error covariance at time t_{i+1}. Now note that $\boldsymbol{\eta}_{t_i}$ is assumed white Gaussian and independent of $\boldsymbol{\epsilon}^o_{t_{i+1}}$ and $\mathbf{x}_{t_0}$, so that $\boldsymbol{\eta}_{t_i}$ is independent of $\mathbf{Y}(t_i)$. This means that $E\{\boldsymbol{\eta}_{t_i} | Y(t_i)\} = E\{\boldsymbol{\eta}_{t_i}\} = \mathbf{0}$. From Eqs. 111 and 113 it follows that

$$\begin{aligned}
\mathbf{x}^f_{t_{i+1}} &= \mathbf{M}_{t_i} \mathbf{x}^a_{t_i} \\
\mathbf{P}^f_{t_{i+1}} &= E\{[\mathbf{M}_{t_i}(\mathbf{x}_{t_i} - \mathbf{x}^a_{t_i}) + \boldsymbol{\eta}_{t_i}][\mathbf{M}_{t_i}(\mathbf{x}_{t_i} - \mathbf{x}^a_{t_i}) + \boldsymbol{\eta}_{t_i}]^T | Y(t_i)\} \\
&= \mathbf{M}_{t_i} \mathbf{P}^a_{t_i} \mathbf{M}^T_{t_i} + \mathbf{Q}_{t_i}
\end{aligned} \tag{114}$$

This is the so-called *forecast step* of the time-discrete Kalman filter. Note in particular that, when $p(\boldsymbol{x}_{t_i} | Y(t_i))$ is Gaussian, from the linearity of Eq. 111 it follows that $p(\boldsymbol{x}_{t_{i+1}} | Y(t_i))$ is also Gaussian, with mean and covariance given by Eq. 114. By induction, it follows that $p(\boldsymbol{x}_{t_{i+1}} | Y(t_i))$ is Gaussian under the assumption that $p(\boldsymbol{x}_{t_0})$ is Gaussian.

From Eq. 79 we can write

$$\begin{aligned}
p(\boldsymbol{x}_{t_{i+1}} | Y(t_{i+1})) &= \frac{p(\boldsymbol{x}_{t_{i+1}}, Y(t_{i+1}))}{p(Y(t_{i+1}))} = \frac{p(\boldsymbol{x}_{t_{i+1}}, \boldsymbol{y}_{t_{i+1}}, Y(t_i))}{p(\boldsymbol{y}_{t_{i+1}}, Y(t_i))} \\
&= \frac{p(\boldsymbol{y}_{t_{i+1}} | \boldsymbol{x}_{t_{i+1}}, Y(t_i)) p(\boldsymbol{x}_{t_{i+1}}, Y(t_i))}{p(\boldsymbol{y}_{t_{i+1}} | Y(t_i)) p(Y(t_i))} \\
&= \frac{p(\boldsymbol{y}_{t_{i+1}} | \boldsymbol{x}_{t_{i+1}}, Y(t_i)) p(\boldsymbol{x}_{t_{i+1}} | Y(t_i)) p(Y(t_i))}{p(\boldsymbol{y}_{t_{i+1}} | Y(t_i)) p(Y(t_i))} \\
&= \frac{p(\boldsymbol{y}_{t_{i+1}} | \boldsymbol{x}_{t_{i+1}}, Y(t_i)) p(\boldsymbol{x}_{t_{i+1}} | Y(t_i))}{p(\boldsymbol{y}_{t_{i+1}} | Y(t_i))}.
\end{aligned} \tag{115}$$

Now, for a given $\mathbf{x}_{t_{i+1}}$, $\mathbf{y}_{t_{i+1}}$ depends only on $\boldsymbol{\epsilon}^o_{t_{i+1}}$ which is assumed independent of $Y(t_i)$, so that $p(\boldsymbol{y}_{t_{i+1}} | \boldsymbol{x}_{t_{i+1}}, Y(t_i)) = p(\boldsymbol{y}_{t_{i+1}} | \boldsymbol{x}_{t_{i+1}})$ and Eq. 115 becomes

$$p(\boldsymbol{x}_{t_{i+1}} | Y(t_{i+1})) = \frac{p(\boldsymbol{y}_{t_{i+1}} | \boldsymbol{x}_{t_{i+1}}) p(\boldsymbol{x}_{t_{i+1}} | Y(t_i))}{p(\boldsymbol{y}_{t_{i+1}} | Y(t_i))}. \tag{116}$$

Note that this holds also for a nonlinear observation operator. Let us now focus on the density functions in Eq. 116. From Eq. 113 we know that

$$p(\boldsymbol{x}_{t_{i+1}}|Y(t_i)) = N(\mathbf{x}^f_{t_{i+1}}, \mathbf{P}^f_{t_{i+1}}) \tag{117}$$

where $N(\mathbf{m}, \mathbf{P})$ denotes a Gaussian distribution with mean $\mathbf{m}$ and covariance $\mathbf{P}$. From Eq. 112 we have

$$E\{\mathbf{y}_{t_{i+1}}|\mathbf{x}_{t_{i+1}}\} = \mathbf{H}_{t_{i+1}}\mathbf{x}_{t_{i+1}} \tag{118}$$

as $E\{\boldsymbol{\epsilon}^o_{t_{i+1}}|\mathbf{x}_{t_{i+1}}\} = \mathbf{0}$ given $\boldsymbol{\epsilon}^o_{t_{i+1}}$ is state independent and unbiased. Also

$$E\{(\mathbf{y}_{t_{i+1}} - E\{\mathbf{y}_{t_{i+1}}|\mathbf{x}_{t_{i+1}}\})(\mathbf{y}_{t_{i+1}} - E\{\mathbf{y}_{t_{i+1}}|\mathbf{x}_{t_{i+1}}\})^T|\mathbf{x}_{t_{i+1}}\} = \mathbf{R}_{t_{i+1}} \tag{119}$$

so that

$$p(\boldsymbol{y}_{t_{i+1}}|\boldsymbol{x}_{t_{i+1}}) = N(\mathbf{H}_{t_{i+1}}\mathbf{x}_{t_{i+1}}, \mathbf{R}_{t_{i+1}}). \tag{120}$$

We also have (see Eqs. 112 and 113)

$$E\{\mathbf{y}_{t_{i+1}}|Y(t_i)\} = \mathbf{H}_{t_{i+1}}E\{\mathbf{x}_{t_{i+1}}|Y(t_i)\} + E\{\boldsymbol{\epsilon}^o_{t_{i+1}}|Y(t_i)\} = \mathbf{H}_{t_{i+1}}\mathbf{x}^f_{t_{i+1}} \tag{121}$$

$$E\{(\mathbf{y}_{t_{i+1}} - E\{\mathbf{y}_{t_{i+1}}|Y(t_i)\})(\mathbf{y}_{t_{i+1}} - E\{\mathbf{y}_{t_{i+1}}|Y(t_i)\})^T|Y(t_i)\} \tag{122}$$

$$= \mathbf{H}_{t_{i+1}}E\{(\mathbf{x}_{t_{i+1}} - \mathbf{x}^f_{t_{i+1}})(\mathbf{x}_{t_{i+1}} - \mathbf{x}^f_{t_{i+1}})^T|Y(t_i)\}\mathbf{H}^T_{t_{i+1}} + E\{\boldsymbol{\epsilon}^o_{t_{i+1}}\boldsymbol{\epsilon}^{oT}_{t_{i+1}}|Y(t_i)\} \tag{123}$$

$$= \mathbf{H}_{t_{i+1}}\mathbf{P}^f_{t_{i+1}}\mathbf{H}^T_{t_{i+1}} + \mathbf{R}_{t_{i+1}} \tag{124}$$

so that

$$p(\boldsymbol{y}_{t_{i+1}}|Y(t_i)) = N(\mathbf{H}_{t_{i+1}}\mathbf{x}^f_{t_{i+1}}, \mathbf{H}_{t_{i+1}}\mathbf{P}^f_{t_{i+1}}\mathbf{H}^T_{t_{i+1}} + \mathbf{R}_{t_{i+1}}). \tag{125}$$

Eq. 116 can then be written as

$$p(\boldsymbol{x}_{t_{i+1}}|Y(t_{i+1})) = \frac{|\mathbf{H}_{t_{i+1}}\mathbf{P}^f_{t_{i+1}}\mathbf{H}^T_{t_{i+1}} + \mathbf{R}_{t_{i+1}}|^{1/2}}{(2\pi)^{n/2}|\mathbf{P}^f_{t_{i+1}}|^{1/2}|\mathbf{R}_{t_{i+1}}|^{1/2}} \exp(-J(\mathbf{x}_{t_{i+1}})/2) \tag{126}$$

where $|\mathbf{P}|$ is the determinant of $\mathbf{P}$ and $J(\mathbf{x}_{t_{i+1}})$ is defined as

$$\begin{aligned} J(\mathbf{x}_{t_{i+1}}) &= (\mathbf{x}_{t_{i+1}} - \mathbf{x}^f_{t_{i+1}})^T(\mathbf{P}^f_{t_{i+1}})^{-1}(\mathbf{x}_{t_{i+1}} - \mathbf{x}^f_{t_{i+1}}) + (\mathbf{y}_{t_{i+1}} - \mathbf{H}_{t_{i+1}}\mathbf{x}_{t_{i+1}})^T\mathbf{R}^{-1}_{t_{i+1}}(\mathbf{y}_{t_{i+1}} - \mathbf{H}_{t_{i+1}}\mathbf{x}_{t_{i+1}}) \\ &- (\mathbf{y}_{t_{i+1}} - \mathbf{H}_{t_{i+1}}\mathbf{x}^f_{t_{i+1}})^T(\mathbf{H}_{t_{i+1}}\mathbf{P}^f_{t_{i+1}}\mathbf{H}^T_{t_{i+1}} + \mathbf{R}_{t_{i+1}})^{-1}(\mathbf{y}_{t_{i+1}} - \mathbf{H}_{t_{i+1}}\mathbf{x}^f_{t_{i+1}}). \end{aligned} \tag{127}$$

Note that Eq. 127 is quadratic in $\mathbf{x}_{t_{i+1}}$ so that we can write

$$J(\mathbf{x}_{t_{i+1}}) = (\mathbf{x}_{t_{i+1}} - \mathbf{x}^a_{t_{i+1}})^T(\mathbf{P}^a_{t_{i+1}})^{-1}(\mathbf{x}_{t_{i+1}} - \mathbf{x}^a_{t_{i+1}}) + c \tag{128}$$

where c is independent of $\mathbf{x}_{t_{i+1}}$. By equating the second-order terms in Eqs. 127 and 128 we find

$$\mathbf{x}^T_{t_{i+1}}(\mathbf{P}^f_{t_{i+1}})^{-1}\mathbf{x}_{t_{i+1}} + \mathbf{x}^T_{t_{i+1}}\mathbf{H}^T_{t_{i+1}}\mathbf{R}^{-1}_{t_{i+1}}\mathbf{H}_{t_{i+1}}\mathbf{x}_{t_{i+1}} = \mathbf{x}^T_{t_{i+1}}(\mathbf{P}^a_{t_{i+1}})^{-1}\mathbf{x}_{t_{i+1}} \tag{129}$$

so that we have

$$(\mathbf{P}^a_{t_{i+1}})^{-1} = (\mathbf{P}^f_{t_{i+1}})^{-1} + \mathbf{H}^T_{t_{i+1}}\mathbf{R}^{-1}_{t_{i+1}}\mathbf{H}_{t_{i+1}}. \tag{130}$$

For the first-order terms we can write

$$-2\mathbf{x}^T_{t_{i+1}}(\mathbf{P}^f_{t_{i+1}})^{-1}\mathbf{x}^f_{t_{i+1}} - 2\mathbf{x}^T_{t_{i+1}}\mathbf{H}^T_{t_{i+1}}\mathbf{R}^{-1}_{t_{i+1}}\mathbf{y}_{t_{i+1}} = -2\mathbf{x}^T_{t_{i+1}}(\mathbf{P}^a_{t_{i+1}})^{-1}\mathbf{x}^a_{t_{i+1}} \tag{131}$$

$$\mathbf{x}^T_{t_{i+1}}[(\mathbf{P}^f_{t_{i+1}})^{-1}\mathbf{x}^f_{t_{i+1}} + \mathbf{H}^T_{t_{i+1}}\mathbf{R}^{-1}_{t_{i+1}}\mathbf{y}_{t_{i+1}} - (\mathbf{P}^a_{t_{i+1}})^{-1}\mathbf{x}^a_{t_{i+1}}] = \mathbf{0} \tag{132}$$

and given the result must be independent of $\mathbf{x}_{t_{i+1}}$ we have

$$(\mathbf{P}^f_{t_{i+1}})^{-1}\mathbf{x}^f_{t_{i+1}} + \mathbf{H}^T_{t_{i+1}}\mathbf{R}^{-1}_{t_{i+1}}\mathbf{y}_{t_{i+1}} - (\mathbf{P}^a_{t_{i+1}})^{-1}\mathbf{x}^a_{t_{i+1}} = \mathbf{0}. \tag{133}$$

From Eqs. 130 and 133 we have

$$\begin{aligned}
\mathbf{x}^a_{t_{i+1}} &= [(\mathbf{P}^f_{t_{i+1}})^{-1} + \mathbf{H}^T_{t_{i+1}}\mathbf{R}^{-1}_{t_{i+1}}\mathbf{H}_{t_{i+1}}]^{-1}[\mathbf{H}^T_{t_{i+1}}\mathbf{R}^{-1}_{t_{i+1}}\mathbf{y}_{t_{i+1}} + (\mathbf{P}^f_{t_{i+1}})^{-1}\mathbf{x}^f_{t_{i+1}}] &(134)\\
&= \mathbf{x}^f_{t_{i+1}} + [(\mathbf{P}^f_{t_{i+1}})^{-1} + \mathbf{H}^T_{t_{i+1}}\mathbf{R}^{-1}_{t_{i+1}}\mathbf{H}_{t_{i+1}}]^{-1}\mathbf{H}^T_{t_{i+1}}\mathbf{R}^{-1}_{t_{i+1}}(\mathbf{y}_{t_{i+1}} - \mathbf{H}_{t_{i+1}}\mathbf{x}^f_{t_{i+1}}) &(135)\\
&= \mathbf{x}^f_{t_{i+1}} + (\mathbf{P}^a_{t_{i+1}})^{-1}\mathbf{H}^T_{t_{i+1}}\mathbf{R}^{-1}_{t_{i+1}}(\mathbf{y}_{t_{i+1}} - \mathbf{H}_{t_{i+1}}\mathbf{x}^f_{t_{i+1}}). &(136)
\end{aligned}$$

The constant c has to be of the form

$$c = 2\ln((2\pi)^{n/2}|\mathbf{P}^a_{t_{i+1}}|^{1/2}) \tag{137}$$

so that the density function is normalized. This means that we can write

$$p(\boldsymbol{x}_{t_{i+1}}|Y(t_{i+1})) = N(\mathbf{x}^a_{t_{i+1}}, \mathbf{P}^a_{t_{i+1}}) \tag{138}$$

where $\mathbf{x}^a_{t_{i+1}}$ and $\mathbf{P}^a_{t_{i+1}}$ are given in Eqs. 130 and 135. Now from the Sherman-Morrison-Woodbury formula we can write

$$\begin{aligned}
\mathbf{P}^a &= \mathbf{P}^f - \mathbf{P}^f\mathbf{H}^T(\mathbf{R} + \mathbf{H}\mathbf{P}^f\mathbf{H}^T)^{-1}\mathbf{H}\mathbf{P}^f &(139)\\
&= (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P}^f &(140)
\end{aligned}$$

where

$$\mathbf{K} \equiv \mathbf{P}^f\mathbf{H}^T(\mathbf{R} + \mathbf{H}\mathbf{P}^f\mathbf{H}^T)^{-1} \tag{141}$$

is the *Kalman gain*. Now, from Eqs. 139 and 141 we have that

$$\mathbf{P}^a\mathbf{H}^T\mathbf{R}^{-1} = \mathbf{K} \tag{142}$$

so that an alternative form of Eq. 135 is given by

$$\mathbf{x}^a_{t_{i+1}} = \mathbf{x}^f_{t_{i+1}} + \mathbf{K}(\mathbf{y}_{t_{i+1}} - \mathbf{H}_{t_{i+1}}\mathbf{x}^f_{t_{i+1}}). \tag{143}$$

Eqs. 143, 141 and 140 define the *analysis update* for the time-discrete Kalman filter.

Note that $\mathbf{K}$, $\mathbf{P}^f$ and $\mathbf{P}^a$ do not depend on the observations, so that $\mathbf{P}^f$ and $\mathbf{P}^a$ are unconditional covariances. This means that $\mathbf{K}$, $\mathbf{P}^f$ and $\mathbf{P}^a$ can be precomputed, so that it is possible to choose the appropriate $\mathbf{R}_t$ for the required accuracy of the state. We now define the *innovation* $\mathbf{d}_{t_{i+1}}$ as

$$\mathbf{d}_{t_{i+1}} = \mathbf{y}_{t_{i+1}} - \mathbf{H}_{t_{i+1}}\mathbf{x}^f_{t_{i+1}} = \boldsymbol{\epsilon}^o_{t_{i+1}} - \mathbf{H}_{t_{i+1}}\boldsymbol{\epsilon}^f_{t_{i+1}} \tag{144}$$

where $\boldsymbol{\epsilon}^f_{t_{i+1}} = \mathbf{x}^f_{t_{i+1}} - \mathbf{x}_{t_{i+1}}$. We have that

$$\begin{aligned}
E\{\mathbf{d}_{t_{i+1}}\} &= \mathbf{0} &(145)\\
E\{\mathbf{d}_{t_{i+1}}\mathbf{d}^T_{t_{i+1}}\} &= \mathbf{H}_{t_{i+1}}\mathbf{P}^f_{t_{i+1}}\mathbf{H}^T_{t_{i+1}} + \mathbf{R}_{t_{i+1}} &(146)
\end{aligned}$$

and it can be shown that $E\{\mathbf{d}_{t_i}\mathbf{d}^T_{t_j}\} = \mathbf{0}$ for $i \neq j$. This means that the innovation sequence is a white Gaussian process. Innovations and their statistics are readily available and they can be used to assess the performance of the filter. In particular, by monitoring the innovations it is possible to check for the problem of *filter divergence*. This problem

may presents itself when the uncertainty associated to the dynamical system model and to the observations are small, so that the filter quickly converges to the estimate and subsequent observations have a small impact on the estimate. However, given the model is "wrong", the estimate of the state and the state diverge. In this case, the innovations become biased with bias magnitude that is larger than the analysis error standard deviation. A possible solution to this problem is to increase the typical magnitude of model error, so that forecast errors increase in magnitude and new observation can correct the estimate. More generally, it is possible to prove that conservative estimates of both model and observation errors generates conservative estimates of both forecast and analysis errors, so that we can determine upper bound to the actual error covariances.

In the case when the distribution of the initial state and of the model and observation errors are not Gaussian, it is still possible to prove that the analysis for a linear observation operator is given by the Kalman filter update. The analysis is still the minimum variance solution, with unbiased error and is often denoted as the best linear unbiased estimate (BLUE). However, the analysis is not anymore the conditional expectation of the posterior density function. This means that when densities are non-Gaussian the conditional expectation is a nonlinear function of the state, which may have a smaller uncertainty (remember that when the density is non-Gaussian, the mean and the variance are not enough to determine the density function).

As seen in Eqs. 141 and 143, the Kalman filter requires the inversion of an $m \times m$ matrix, where m is the dimension of the measurement vector. However, if the measurement errors are uncorrelated, it is possible to process each observation separately, so that we need instead to invert m scalars. This has practical advantages as it enables the processing of each observation separately, so that the filter is easier to process numerically in a parallel way.

4.6 Observability and controllability

It is of fundamental importance to be able to establish whether it is possible to determine the state of the system from the measurements. In order to investigate this issue, let us consider $\mathbf{Y}(t_N) = (\mathbf{y}_{t_1}^T, \mathbf{y}_{t_2}^T, \cdots, \mathbf{y}_{t_i}^T, \cdots, \mathbf{y}_{t_N}^T)^T$, where

$$\mathbf{y}_{t_1} = \mathbf{H}_{t_1}\mathbf{x}_{t_1} + \boldsymbol{\epsilon}_{t_1}^o = \mathbf{H}_{t_1}\mathbf{M}_{t_0}\mathbf{x}_{t_0} + \boldsymbol{\epsilon}_{t_1}^o \tag{147}$$

$$\mathbf{y}_{t_2} = \mathbf{H}_{t_2}\mathbf{x}_{t_2} + \boldsymbol{\epsilon}_{t_2}^o = \mathbf{H}_{t_2}\mathbf{M}_{t_1}\mathbf{M}_{t_0}\mathbf{x}_{t_0} + \boldsymbol{\epsilon}_{t_2}^o \tag{148}$$

$$\vdots \tag{149}$$

$$\mathbf{y}_{t_i} = \mathbf{H}_{t_i}\mathbf{x}_{t_i} + \boldsymbol{\epsilon}_{t_i}^o = \mathbf{H}_{t_i}\mathbf{M}_{t_i}\cdots\mathbf{M}_{t_1}\mathbf{M}_{t_0}\mathbf{x}_{t_0} + \boldsymbol{\epsilon}_{t_2}^o \tag{150}$$

$$\vdots \tag{151}$$

$$\mathbf{y}_{t_N} = \mathbf{H}_{t_N}\mathbf{x}_{t_N} + \boldsymbol{\epsilon}_{t_N}^o = \mathbf{H}_{t_N}\mathbf{M}_{t_N}\cdots\mathbf{M}_{t_1}\mathbf{M}_{t_0}\mathbf{x}_{t_0} + \boldsymbol{\epsilon}_{t_N}^o \tag{152}$$

This can be written as

$$\mathbf{Y}(t_N) = \mathbf{H}(t_1, t_N)\mathbf{x}_{t_0} + \boldsymbol{\epsilon}^o \tag{153}$$

where the *observability matrix* $\mathbf{H}(t_1,t_N)$ can be written as

$$\mathbf{H}(t_1,t_N)=\begin{pmatrix}\mathbf{H}_{t_1}\mathbf{M}_{t_0}\\ \mathbf{H}_{t_2}\mathbf{M}_{t_1}\mathbf{M}_{t_0}\\ \vdots\\ \mathbf{H}_{t_i}\mathbf{M}_{t_i}\cdots\mathbf{M}_{t_1}\mathbf{M}_{t_0}\\ \vdots\\ \mathbf{H}_{t_N}\mathbf{M}_{t_N}\cdots\mathbf{M}_{t_1}\mathbf{M}_{t_0}\end{pmatrix} \tag{154}$$

and where $\boldsymbol{\epsilon}^o=((\boldsymbol{\epsilon}^o_{t_1})^T,\cdots,(\boldsymbol{\epsilon}^o_{t_N})^T)^T$, with covariance $\mathbf{R}=Diag\{\mathbf{R}_1,\mathbf{R}_2,\cdots,\mathbf{R}_i,\cdots,\mathbf{R}_N\}$, where $\mathbf{R}_i=E\{\boldsymbol{\epsilon}^o_{t_i}(\boldsymbol{\epsilon}^o_{t_i})^T\}$. In absence of any prior information, it is possible to determine $\mathbf{x}_{t_0}$ as the solution of the least squares problem:

$$\mathbf{x}_{t_0}=(\mathbf{H}^T(t_1,t_N)\mathbf{R}^{-1}\mathbf{H}(t_1,t_N))^{-1}\mathbf{H}^T(t_1,t_N)\mathbf{R}^{-1}\mathbf{Y}(t_N) \tag{155}$$

provided $\mathbf{H}^T(t_1,t_N)\mathbf{R}^{-1}\mathbf{H}(t_1,t_N)=\sum_{i=0}^{N}\mathbf{M}^T(i,0)\mathbf{H}^T_{t_i}\mathbf{R}^{-1}_{t_i}\mathbf{H}_{t_i}\mathbf{M}(i,0)\in\mathbb{R}^{n\times n}$ is full rank, where $\mathbf{M}(i,0)=\mathbf{M}_{t_i}\cdots\mathbf{M}_{t_1}\mathbf{M}_{t_0}$. When this is the case, the system is said to be *completely observable*, i.e., it is possible to infer all of the system's initial conditions. From Eqs. 114 and 130 with $\mathbf{Q}_{t_i}=\mathbf{0}$ we can write

$$\begin{aligned}(\mathbf{P}^a_{t_{i+1}})^{-1} &= \mathbf{M}^{-T}_{t_i}(\mathbf{P}^a_{t_i})^{-1}\mathbf{M}^{-1}_{t_i}+\mathbf{H}^T_{t_{i+1}}\mathbf{R}^{-1}_{t_{i+1}}\mathbf{H}_{t_{i+1}} &(156)\\ &= \mathbf{M}^{-T}_{t_i}\mathbf{M}^{-T}_{t_{i-1}}(\mathbf{P}^a_{t_{i-1}})^{-1}\mathbf{M}^{-1}_{t_{i-1}}\mathbf{M}^{-1}_{t_i}+\mathbf{M}^{-T}_{t_i}\mathbf{H}^T_{t_i}\mathbf{R}^{-1}_{t_i}\mathbf{H}_{t_i}\mathbf{M}^{-1}_{t_i}+\mathbf{H}^T_{t_{i+1}}\mathbf{R}^{-1}_{t_{i+1}}\mathbf{H}_{t_{i+1}}\\ &= \mathbf{M}^T(0,i+1)\mathbf{P}^{-1}_{t_0}\mathbf{M}(0,i+1)+\sum_{k=0}^{i+1}\mathbf{M}^T(0,k)\mathbf{H}^T_{t_k}\mathbf{R}^{-1}_{t_k}\mathbf{H}_{t_k}\mathbf{M}(0,k)\\ &= \mathbf{M}^T(0,i+1)\mathbf{P}^{-1}_{t_0}\mathbf{M}(0,i+1)+\mathcal{I}(i+1,0)\end{aligned}$$

where $\mathbf{M}(0,i+1)=\mathbf{M}^{-1}(i+1,0)$ is the transition matrix for propagating the state backward in time and where $\mathcal{I}(i+1,0)$ is the *Fisher information matrix*, which quantifies the reduction of uncertainty on the estimate of the state due to the measurements. We can also see that

$$\mathcal{I}(i+1,0)=\mathbf{M}^T(i,i+1)\mathcal{I}(i,0)\mathbf{M}^T(i,i+1)+\mathbf{H}^T_{t_{i+1}}\mathbf{R}^{-1}_{t_{i+1}}\mathbf{H}_{t_{i+1}} \tag{157}$$

so that $\mathbf{H}^T_{t_{i+1}}\mathbf{R}^{-1}_{t_{i+1}}\mathbf{H}_{t_{i+1}}$ represents the information added by the measurement at time t_{i+1}. Also note that the observability and the information matrices have the same rank, so that the system is said to be completely observable with respect to $\mathbf{Y}(t_i)$ if $\mathcal{I}(i,0)$ is positive definite, for $i\geq 0$.

The concept of controllability describes the property of a system model regarding its ability to map any initial state to any final state in a finite time. From Eq. 111 we can write

$$\begin{aligned}\mathbf{x}_{t_{i+1}} &= \mathbf{M}_{t_i}\mathbf{x}_{t_i}+\boldsymbol{\eta}_{t_i} &(158)\\ &= \mathbf{M}_{t_i}\mathbf{M}_{t_{i-1}}\mathbf{x}_{t_{i-1}}+\mathbf{M}_{t_{i-1}}\boldsymbol{\eta}_{t_{i-1}}+\boldsymbol{\eta}_{t_i} &(159)\\ \vdots & &(160)\\ &= \mathbf{M}(i,0)\mathbf{x}_{t_0}+\sum_{k=0}^{i}\mathbf{M}(i,k)\boldsymbol{\eta}_{t_k}. &(161)\end{aligned}$$

If we define $\Delta_k = \mathbf{x}_{t_{i+1}} - \mathbf{M}(i,0)\mathbf{x}_{t_0}$ we have that $E\{\Delta_k\} = 0$ and

$$E\{\Delta_k \Delta_k^T\} = \sum_{k=0}^{i} \mathbf{M}(i,k)\mathbf{Q}_k \mathbf{M}^T(i,k) \equiv \mathcal{C}(i,0) \tag{162}$$

where $\mathcal{C}(i,0)$ is the *controllability matrix*. The discrete dynamical system described by Eqs. 111 and 112 is said to be completely controllable if $\mathcal{C}(i,0)$ is positive definite and bounded, for $k > 0$. We now rewrite Eq. 143 as

$$\mathbf{x}^a_{t_{i+1}} = (\mathbf{I} - \mathbf{K}\mathbf{H}_{t_{i+1}})\mathbf{x}^f_{t_{i+1}} + \mathbf{K}\mathbf{y}_{t_{i+1}}. \tag{163}$$

It is possible to prove the following: if the system model is both observable and controllable then the Kalman filter is *asymptotically stable*, i.e.,

$$\lim_{i\to\infty} \|(\mathbf{I} - \mathbf{K}\mathbf{H}_{t_{i+1}})\mathbf{x}^f_{t_{i+1}}\| = 0 \tag{164}$$

This means that the effect of an increasingly large number of measurements make the impact of the prior on the estimate negligible (the prior is "forgotten"). This is a desirable property of the filter, as bounded inputs $\boldsymbol{\eta}_{t_i}$ produce bounded outputs $\mathbf{x}_{t_{i+1}}$ for any initial conditions.

Finally, note that from Eq. 163 we can write an alternative form for $\mathbf{P}^a_{t_{i+1}}$ known as the Joseph form:

$$\mathbf{P}^a_{t_{i+1}} = (\mathbf{I} - \mathbf{K}_{t_{i+1}}\mathbf{H}_{t_{i+1}})\mathbf{P}^f_{t_{i+1}}(\mathbf{I} - \mathbf{K}_{t_{i+1}}\mathbf{H}_{t_{i+1}})^T + \mathbf{K}_{t_{i+1}}\mathbf{R}_{t_{i+1}}\mathbf{K}^T_{t_{i+1}}. \tag{165}$$

Note that $\mathbf{P}^a_{t_{i+1}}$ is now the sum of two symmetric matrices and positive definite or semidefinite matrices, so that the Joseph form is better suited for numerical computations.

4.7 Square root filter

It is possible to increase the numerical precision of the conventional Kalman filter algorithm by estimating and propagating the square root of the analysis error covariance (as well as the state) rather than the error covariance. In this way, it is possible to double the accuracy of the calculations and ensure positive definiteness of the error covariances. At time t_{i+1} we have seen we can write

$$\mathbf{P}^f_{t_{i+1}} = \mathbf{M}_{t_i}\mathbf{P}^a_{t_i}\mathbf{M}^T_{t_i} + \mathbf{Q}_{t_i} \tag{166}$$

$$\mathbf{P}^a_{t_{i+1}} = (\mathbf{I} - \mathbf{K}_{t_{i+1}}\mathbf{H}_{t_{i+1}})\mathbf{P}^f_{t_{i+1}}. \tag{167}$$

We now express $\mathbf{P}^f_{t_{i+1}}$ and $\mathbf{P}^a_{t_{i+1}}$ as $\mathbf{P}^f_{t_{i+1}} = \mathbf{Z}^f(\mathbf{Z}^f)^T$ and $\mathbf{P}^a_{t_{i+1}} = \mathbf{Z}^a(\mathbf{Z}^a)^T$, where $\mathbf{Z}^f$ and $\mathbf{Z}^a$ are matrix square roots. From Eqs. 167 and 141 we can write (dropping the time index)

$$\mathbf{P}^a = [\mathbf{I} - \mathbf{P}^f\mathbf{H}^T(\mathbf{R} + \mathbf{H}\mathbf{P}^f\mathbf{H}^T)^{-1}\mathbf{H}]\mathbf{P}^f \tag{168}$$

$$= \mathbf{Z}^f[\mathbf{I} - \mathbf{Z}^{fT}\mathbf{H}^T(\mathbf{H}\mathbf{Z}^f\mathbf{Z}^{fT}\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{H}\mathbf{Z}^f]\mathbf{Z}^{fT} \tag{169}$$

$$= \mathbf{Z}^f(\mathbf{I} - \mathbf{S}^T\mathbf{C}^{-1}\mathbf{S})\mathbf{Z}^{fT} \tag{170}$$

where $\mathbf{S} = \mathbf{H}\mathbf{Z}^f$ and $\mathbf{C} = \mathbf{S}\mathbf{S}^T + \mathbf{R}$. We can then write

$$\mathbf{Z}^a_{t_{i+1}} = \mathbf{Z}^f_{t_{i+1}}\mathbf{X}_{t_{i+1}}\mathbf{U} \tag{171}$$

where $\mathbf{X}\mathbf{X}^T = (\mathbf{I}-\mathbf{S}^T\mathbf{C}^{-1}\mathbf{S})$ and $\mathbf{U}$ is an arbitrary orthogonal matrix such that $\mathbf{U}\mathbf{U}^T = \mathbf{I}$. The factor $\mathbf{Z}^f_{t_{i+1}}$ can be computed from $\mathbf{Z}^a_{t_i}$ as

$$\mathbf{Z}^f_{t_{i+1}} = [\mathbf{M}_{t_i}\mathbf{Z}^a_{t_i}, \mathbf{Z}^q_{t_i}] \tag{172}$$

where $\mathbf{Q}_{t_i} = \mathbf{Z}^q_{t_i}\mathbf{Z}^{q_{t_i}T}$. The analysis update is given as (see Eq. 143)

$$\mathbf{x}^a_{t_{i+1}} = \mathbf{x}^f_{t_{i+1}} + \mathbf{Z}^f_{t_{i+1}}\mathbf{S}^T\mathbf{C}^{-1}(\mathbf{y}_{t_{i+1}} - \mathbf{H}_{t_{i+1}}\mathbf{x}^f_{t_{i+1}}) \tag{173}$$

where $\mathbf{x}^f_{t_{i+1}} = \mathbf{M}_{t_i}\mathbf{x}^a_{t_i}$. Note that from Eq. 172 it follows that the size of $\mathbf{Z}^f_{t_{i+1}}$ increases at each time step due to model error. However, it is possible to reduce the number of elements by discarding the components of $\mathbf{Z}^f_{t_{i+1}}$ that lead to small forecast error variances, or by considering $\mathbf{Z}^{f\prime}_{t_{i+1}} = \mathbf{Z}^f_{t_{i+1}}\mathbf{U}$ instead, where $\mathbf{U}$ is the orthogonal matrix determined as the $\mathbf{QR}$-decomposition of $\mathbf{Z}^{fT}_{t_{i+1}}$.

5 Nonlinear state estimation

In the preceding sections we discussed the solution of the state estimation problem when both the system and the observation model were linearly related to the state. We now extend the discussion to the case when the state evolves according to a nonlinear stochastic differential equation and the observations are nonlinearly related to the state.

5.1 Nonlinear stochastic dynamic model with discrete-time measurements

Let us assume the stochastic process $\mathbf{x}_t$ $(t \geq t_0)$is described by (see also Eq. 89)

$$d\mathbf{x}_t = f(\mathbf{x}_t, t)dt + \mathbf{G}_t d\boldsymbol{\beta}_t \tag{174}$$

and that the observation vector at time $t_{i+1} \geq t_0$ can be written as (see Eq. 95)

$$\mathbf{y}_{t_{i+1}} = H(\mathbf{x}_{t_{i+1}}, t_{i+1}) + \boldsymbol{\epsilon}^o_{t_{i+1}}. \tag{175}$$

As seen for the linear case, Eq. 174 can be rewritten as

$$\mathbf{x}_{t_{i+1}} = M(\mathbf{x}_{t_i}, t_i) + \boldsymbol{\eta}_{t_i} \tag{176}$$

We want to determine the conditional density for $\mathbf{x}_{t_{i+1}}$ given all observation up to (and including) time t_{i+1}, which we denote as $p(\boldsymbol{x}_{t_{i+1}}|Y(t_{i+1}))$, solution of the filtering problem. If we are successful, we would, at least in principle, be able to to determine optimal estimators such as the minimum variance estimate (conditional mean) and the maximum a posteriori estimate (conditional mode).

In section 3.3 we have seen that $p(\boldsymbol{x}_t)$ is given by the Kolmogorov's forward equation with initial conditions $p(\boldsymbol{x}_{t_0})$. This means that $p(\boldsymbol{x}_{t_{i+1}}|Y(t_{i+1}))$ can also be determined from Eq. 93 between observation times (e.g., $t_i \leq t \leq t_{i+1}$). The problem is then to determine the relationship between $p(\boldsymbol{x}_{t_{i+1}}|Y(t_{i+1}))$ and $p(\boldsymbol{x}_{t_{i+1}}|Y(t_i))$. We can write

$$p(\boldsymbol{x}_{t_{i+1}}|Y(t_i)) = \frac{p(\boldsymbol{x}_{t_{i+1}}, Y(t_i))}{p(Y(t_i))} \tag{177}$$

and

$$p(\boldsymbol{y}_{t_{i+1}}|\boldsymbol{x}_{t_{i+1}},Y(t_i))p(\boldsymbol{x}_{t_{i+1}},Y(t_i)) = p(\boldsymbol{y}_{t_{i+1}},\boldsymbol{x}_{t_{i+1}},Y(t_i)) \tag{178}$$

so that

$$\int_{-\infty}^{+\infty} p(\boldsymbol{y}_{t_{i+1}}|\boldsymbol{x}_{t_{i+1}},Y(t_i))p(\boldsymbol{x}_{t_{i+1}}|Y(t_i))d\boldsymbol{x}_{t_{i+1}} \tag{179}$$

$$= \frac{1}{p(Y(t_i))}\int_{-\infty}^{+\infty} p(\boldsymbol{y}_{t_{i+1}},\boldsymbol{x}_{t_{i+1}},Y(t_i))d\boldsymbol{x}_{t_{i+1}} \tag{180}$$

$$= \frac{p(\boldsymbol{y}_{t_{i+1}},Y(t_i))}{p(Y(t_i))} \tag{181}$$

$$= p(\boldsymbol{y}_{t_{i+1}}|Y(t_i)). \tag{182}$$

Finally, we can write

$$p(\boldsymbol{x}_{t_{i+1}}|Y(t_{i+1})) = \frac{p(\boldsymbol{y}_{t_{i+1}}|\boldsymbol{x}_{t_{i+1}},Y(t_i))p(\boldsymbol{x}_{t_{i+1}}|Y(t_i))}{\int_{-\infty}^{+\infty} p(\boldsymbol{y}_{t_{i+1}}|\boldsymbol{x}_{t_{i+1}},Y(t_i))p(\boldsymbol{x}_{t_{i+1}}|Y(t_i))d\boldsymbol{x}_{t_{i+1}}}. \tag{183}$$

Now, as discussed in section 4.5, for a given $\mathbf{x}_{t_{i+1}}$, $\mathbf{y}_{t_{i+1}}$ depends only on $\boldsymbol{\epsilon}^o_{t_{i+1}}$ which is assumed independent of $Y(t_i)$, so that $p(\boldsymbol{y}_{t_{i+1}}|\boldsymbol{x}_{t_{i+1}},Y(t_i)) = p(\boldsymbol{y}_{t_{i+1}}|\boldsymbol{x}_{t_{i+1}})$. From our assumptions on $\boldsymbol{\epsilon}^o_t$ we can write

$$p(\boldsymbol{y}_{t_{i+1}}|\boldsymbol{x}_{t_{i+1}}) = N(H(\mathbf{x}_{t_{i+1}},t_{i+1}),\mathbf{R}_{t_{i+1}}). \tag{184}$$

For linear filters, we have seen that the conditional density $p(\boldsymbol{x}_{t_{i+1}}|Y(t_{i+1}))$ is Gaussian, so that the density is completely determined by the mean and the covariance. In the nonlinear case, we actually need to know the whole density to be able to calculate its moments: the problem of calculating the minimum variance estimate is infinite dimensional. However, it is always possible to consider the maximum a posteriori estimate instead, as explained below.

Let us assume that $p(\boldsymbol{x}_{t_{i+1}}|Y(t_i))$ is Gaussian, so that we can write (see also Eq. 117)

$$p(\boldsymbol{x}_{t_{i+1}}|Y(t_i)) = N(\mathbf{x}^f_{t_{i+1}},\mathbf{P}^f_{t_{i+1}}). \tag{185}$$

Note that, differently from the linear case, this assumption may not be justified as the nonlinear evolution of the density makes the density non-Gaussian even when the initial density is Gaussian. With this approximation, Eq. 183 becomes

$$p(\boldsymbol{x}_{t_{i+1}}|Y(t_{i+1})) = \frac{\exp(-J(\boldsymbol{x}_{t_{i+1}}))}{\int_{-\infty}^{+\infty}\exp(-J(\boldsymbol{x}_{t_{i+1}}))d\boldsymbol{x}_{t_{i+1}}} \tag{186}$$

where

$$J(\mathbf{x}_{t_{i+1}}) = \frac{1}{2}(\mathbf{x}_{t_{i+1}}-\mathbf{x}^f_{t_{i+1}})^T(\mathbf{P}^f_{t_{i+1}})^{-1}(\mathbf{x}_{t_{i+1}}-\mathbf{x}^f_{t_{i+1}})+(\mathbf{y}_{t_{i+1}}-H(\mathbf{x}_{t_{i+1}}))^T\mathbf{R}^{-1}_{t_{i+1}}(\mathbf{y}_{t_{i+1}}-H(\mathbf{x}_{t_{i+1}})). \tag{187}$$

To evaluate $p(\boldsymbol{x}_{t_{i+1}}|Y(t_{i+1}))$ in Eq. 186 we need to evaluate the integral at the denominator, and this may be impractical when the dimensionality of the system model is large. However, as the denominator does not depend on $\mathbf{x}_{t_{i+1}}$, it is possible to estimate the maximum of $p(\boldsymbol{x}_{t_{i+1}}|Y(t_{i+1}))$ rather than its mean. This can be found by setting the gradient of J to zero, that is

$$\frac{\partial J(\mathbf{x}_{t_{i+1}})}{\partial \mathbf{x}_{t_{i+1}}} = (\mathbf{P}^f_{t_{i+1}})^{-1}(\mathbf{x}_{t_{i+1}}-\mathbf{x}^f_{t_{i+1}}) + \mathbf{H}^T_{t_{i+1}}\mathbf{R}^{-1}_{t_{i+1}}(H(\mathbf{x}_{t_{i+1}})-\mathbf{y}_{t_{i+1}}) = \mathbf{0} \tag{188}$$

where $\mathbf{H}^T_{t_{i+1}} \equiv (\partial H(\mathbf{x}_{t_{i+1}})/\partial \mathbf{x}_{t_{i+1}})$. To determine the minimum of J it is possible to use an iterative method such as the Newton method, which at iteration k determines $\mathbf{x}_{t_{i+1}}$ as (we drop the time index)

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \left(\frac{\partial^2 J(\mathbf{x}_k)}{\partial \mathbf{x}_k^2} \right)^{-1} \frac{\partial J(\mathbf{x}_k)}{\partial \mathbf{x}_k} \tag{189}$$

where

$$\frac{\partial^2 J(\mathbf{x}_k)}{\partial \mathbf{x}_k^2} = (\mathbf{P}^f)^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} + (\frac{\partial \mathbf{H}}{\partial \mathbf{x}_k})^T \mathbf{R}^{-1} (H(\mathbf{x}_k) - \mathbf{y}). \tag{190}$$

Note that $\frac{\partial \mathbf{H}}{\partial \mathbf{x}_k}$ is an order-3 array, that is a vector whose i-th component is the Hessian matrix of the i-th component of $H(\mathbf{x}_k)$. If we neglect the third term on the r.h.s. of Eq. 190, which is small when the residual $H(\mathbf{x}_k) - \mathbf{y}$ or the nonlinearity of $H(\mathbf{x}_k)$ are small, we get the Gauss-Newton iteration

$$\frac{\partial^2 J(\mathbf{x}_k)}{\partial \mathbf{x}_k^2} = (\mathbf{P}^f)^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} = [(\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P}^f]^{-1}. \tag{191}$$

Note that convergence to a local minimum is guaranteed as long as $\mathbf{P}^f$ is positive definite. The iteration in Eq. 189 becomes

$$\mathbf{x}_{k+1} = \mathbf{x}^f + \mathbf{K}_k [\mathbf{y} - H(\mathbf{x}_k) + \mathbf{H}_k (\mathbf{x}_k - \mathbf{x}^f)] \tag{192}$$

where $\mathbf{x}_0 = \mathbf{x}^f$, until convergence, when $\mathbf{x}_{k+1} \simeq \mathbf{x}_k \equiv \mathbf{x}_\infty$. Note that Eq. 192 reduces to the Kalman filter analysis update equation when the observation operator is linear: this nonlinear version is the *iterated extended Kalman filter*. An approximation for the error covariance can be defined as $\mathbf{P}_a = (\mathbf{I} - \mathbf{K}_\infty \mathbf{H}_\infty)\mathbf{P}^f$, which gives $\mathbf{P}_a$ when the observation operator is truncated at the first order (tangent linear approximation). Finally, $\mathbf{x}^f$ and $\mathbf{P}^f$ can be defined as (cf the linear case)

$$\mathbf{x}^f_{t_{i+1}} = M(\mathbf{x}^a_{t_i}) \tag{193}$$

$$\mathbf{P}^f_{t_{i+1}} = \mathbf{M}(t_i, \mathbf{x}^a_{t_i}) \mathbf{P}^a_{t_i} (\mathbf{M}(t_i, \mathbf{x}^a_{t_i}))^T + \mathbf{Q}_{t_i} \tag{194}$$

where $\mathbf{M}(t_i, \mathbf{x}^a_{t_i}) = (\partial M(t, \mathbf{x}_t)/\partial \mathbf{x}_t)(\mathbf{x}_t = \mathbf{x}^a_{t_i})$.

6 Ensemble-based data assimilation

Evaluating the moments of a probability density function associated to a random vector with a number of components over a 3D grid by numerical integration is computationally very challenging. Consider a one-dimensional example. Assume we sample the probability to find each of the 5 components of a random vector between 0 and 1 with a 0.01 resolution. It follows that we need to know the density over 101 sample points (say 100 for simplicity) for each component of the vector, for a total of $100^5 = 10^{10}$ sample points. This problem is a manifestation of the so-called *curse of dimensionality*. An alternative way to provide an estimate of the moments of a density function $p(x)$ (for simplicity we consider the scalar case), which does not require the use of an ordered partition of the integration domain, is through a Monte Carlo integration, also denoted as Monte Carlo method. In this way the expectation $E_{\mathrm{x}}\{x\}$ defined as

$$E_{\mathrm{x}}\{x\} = \int_{-\infty}^{+\infty} x p(x) dx = \mu \tag{195}$$

is approximated by the sample mean, calculated from a set of $(x_1, x_2, \cdots, x_n)$ that are generated from the density $p(x)$ and defined as

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{196}$$

For any set of x_i that are independent and identically distributed (*iid*), with finite expectation μ and variance σ^2, the expectation of $\overline{x}$, $E\{\overline{x}\}$ is equal to $\frac{1}{n}\sum_{i=1}^{n} E\{x_i\} = E\{x\} = \mu$, so that the sample mean is unbiased. This is good news, as a given estimate of a quantity is *accurate* when the expectation of the estimate is close to the quantity. The expectation of the sample mean actually coincides with μ. Also, a given estimate is *precise* when its variance is small. For any x_i that are *iid*, the variance of $\overline{x}$ is given by

$$E\{(\frac{1}{n}\sum_{i=1}^{n} x_i - \mu)^2\} = E\{\frac{1}{n^2}(\sum_{i=1}^{n} -n\mu)^2\} = \frac{1}{n^2}E\{[\sum_{i=1}^{n}(x_i - \mu)]^2\} \tag{197}$$

$$= \frac{1}{n^2}E\{\sum_{i=1}^{n}(x_i - \mu)^2\} = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n} \tag{198}$$

$$\tag{199}$$

where σ^2 is the variance of x_i given by $\sigma^2 = E\{(x_i - \mu)^2\}$ and where by independence of the x_i we have that $E\{(x_i - \mu)(x_j - \mu)\} = 0$ for $i \neq j$. Now, the *central limit theorem* states that $Y_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)$ is normally distributed for $n \to \infty$, with zero mean variance σ^2 so that we can write

$$\lim_{n\to\infty} Y_n = Y \sim N(0, \sigma^2). \tag{200}$$

Given that $\overline{x} = \mu + Y_n/\sqrt{n}$ we have that

$$\lim_{n\to\infty} \overline{x} = \overline{x}_\infty \sim N(\mu, \sigma^2/n). \tag{201}$$

In summary, we can state that $\overline{x}$ is a *consistent* estimate as the probability of finding $|\overline{x} - \mu| < \epsilon$ for a given $\epsilon > 0$ tends to 1 for $n \to \infty$, regardless how the x_i are distributed (as long as they are *iid*). In other words, $\overline{x}$ converges almost surely to μ. Also, we know that the magnitude of error goes to zero as $1/\sqrt{n}$. In a similar way it is possible to prove that the sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ is unbiased, with

$$\lim_{n\to\infty} s^2 = s^2_\infty \sim N(\sigma^2, \mathrm{var}[(x_i - \mu)^2]/n = [E\{(x_i - \mu)^4\} - \sigma^4]/n). \tag{202}$$

so that s^2 is also a consistent estimate. Note that when $x_i \sim N(\mu, \sigma^2)$ we have that $E\{(x_i - \mu)^4\} = 3\sigma^4$ so that $\mathrm{var}[(x_i - \mu)^2] = 2\sigma^4$. In summary, for large n it is reasonable to approximate the mean and the variance of a random variable with the sample mean and the sample variance, respectively, of a set of realizations of the same random variable characterized by an arbitrary distribution. The approximation error tends to zero as $1/\sqrt{n}$.

When we have two sets of *iid* variables, x_i and y_i, we can also define the sample covariance as $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$. It is important to note that if we define a matrix $\mathbf{X}'$ whose first row is represented by an *ensemble* of $x_i - \overline{x}$ and whose second row is represented by an ensemble of $y_i - \overline{y}$, we can define the sample covariance matrix $\mathbf{P}_e$ as

$$\mathbf{P}_e = \frac{1}{n-1}\mathbf{X}'\mathbf{X}'^T \tag{203}$$

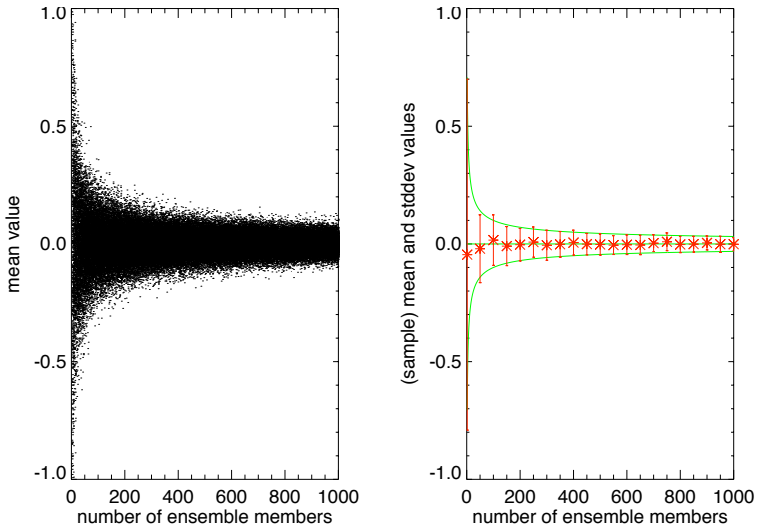

Figure 2: Black dots: mean values of N ensemble members drawn from a Gaussian with zero mean and unit variance, for a total of 100 mean values for each N value In red: sample mean of 100 means and sample standard deviation of the mean In green: expected value of the mean (= 0) and of the standard deviation of the mean (= 1/sqrt(n))

which is a consistent estimate of $\mathbf{P} = \text{cov}(\mathbf{x}, \mathbf{y}) = E\{(\mathbf{x} - \boldsymbol{\mu}_x)(\mathbf{y} - \boldsymbol{\mu}_y)^T\}$, where $\mathbf{x} = (x_1, \cdots, x_n)^T$ and $\mathbf{y} = (y_1, \cdots, y_n)^T$. It is obvious to generalize to the multidimensional case.

6.1 Construction of random vectors

Let us assume that at time t_0 our best estimate of the state of a system is represented by $\mathbf{x}_0^a \in \mathbb{R}^n$, where each of its n components represents the best estimate of a given quantity (e.g., atmospheric temperature) at a given location. We want to generate a

random matrix $\mathbf{X}_0 \in \mathbb{R}^{n\times N}$ with each element representing a realization from a Gaussian distribution with known mean $\mathbf{x}_0^a$ and covariance $\mathbf{P}_0^a$. In other words, the columns of $\mathbf{X}_0$ represent N realizations of $\mathbf{x}_0 \sim N(\mathbf{x}_0^a, \mathbf{P}_0^a)$. Let us assume to have a random number generator to generate each component of $\boldsymbol{\xi}_0 \in \mathbb{R}^n$ from a Gaussian distribution with zero mean and unit variance. Let us know express $\mathbf{P}_0^a$ as its Cholesky decomposition: $\mathbf{P}_0^a = \mathbf{Z}_0^a(\mathbf{Z}_0^a)^T$. The random vector $\mathbf{x}_0$ we are seeking can be calculated as

$$\mathbf{x}_0 = \mathbf{x}_0^a + \mathbf{Z}_0^a\boldsymbol{\xi}_0, \tag{204}$$

with $E\{\mathbf{x}_0\} = \mathbf{x}_0^a$ and $E\{(\mathbf{x}_0 - \mathbf{x}_0^a)(\mathbf{x}_0 - \mathbf{x}_0^a)^T\} = \mathbf{P}_0^a$. If we now construct the matrix $\mathbf{X_0}' \in \mathbb{R}^{n\times N}$ where each column is a vector of realizations $(\mathbf{x}_0 - \mathbf{x}_0^a)_i$, we can state that the matrix $(\mathbf{P}_0^a)_e = \mathbf{X}_0'\mathbf{X}_0'^T/(N-1)$ is a consistent estimate of $\mathbf{P}_0^a$.

In order to determine $\mathbf{x}_0$ we need to rely on a $\mathbf{P}_0^a$ that is suitable for the system whose variability we want to describe. Let us consider the (i,j) element of $\mathbf{P}_0^a$ divided by the product of the standard deviation at $\mathbf{r}_i$ and at $\mathbf{r}_j$, given by the correlation element $C(\mathbf{r}_i, \mathbf{r}_j) \equiv E\{[(x_0)_i - (x_0^a)_i][(x_0)_j - (x_0^a)_j]^T\}/\sqrt{E\{(x_0)_i^2\}E\{(x_0)_j^2\}}$. The correlation function is said to be separable when we can write

$$C(\mathbf{r}_i, \mathbf{r}_j) = C_h(x_i, y_i; x_j, y_j)C_v(z_i, z_j) \tag{205}$$

where C_h and C_v are the horizontal and vertical error correlation functions, respectively. Note that for an hydrostatically balanced atmosphere it is possible to show that horizontal temperature and geopotential height (or pressure) error correlations are in general not separable. However, in practice it is usually assumed that the error correlation function is separable under specific conditions. Let us now assume for simplicity that the function $C(\mathbf{r}_i, \mathbf{r}_j)$ is separable. Let us now define $\mathbf{r}_i = (x_i, y_i)^T$, the horizontal position vector. Another common assumption is to consider horizontal correlation functions that are homogeneous (function only of $\mathbf{r}_i - \mathbf{r}_j$) and also isotropic (function only of $r_i - r_j$). Note that the concept of homogeneity in space is equivalent to stationarity in time.

Examples of homogeneous and isotropic correlation functions are the exponential (or first-order autoregressive) function

$$C(r) = \exp(-\frac{r}{L}) \tag{206}$$

the second-order autoregressive function (SOAR)

$$C(r) = (1 + \frac{r}{L})\exp(-\frac{r}{L}) \tag{207}$$

and the Gaussian

$$C(r) = \exp(-\frac{r^2}{2L^2}). \tag{208}$$

It is also possible to define homogeneous and isotropic correlations that are compactly supported in $[-c, c]$. Once we have specified the covariance function we can construct a correlation matrix $\mathbf{C}$ (if the dimensions are not too large!) and a covariance matrix $\mathbf{P}_0^a = \boldsymbol{\Sigma}\mathbf{C}\boldsymbol{\Sigma}$ so as to determine the initial ensemble as in Eq. 204.

The forecast step for the i-th ensemble member is given by

$$\mathbf{x}_i^f(t_{j+1}) = M(\mathbf{x}_i^a(t_j)) + \sqrt{\Delta t}\boldsymbol{\eta}_i(t_j) \tag{209}$$

where $\boldsymbol{\eta}_i(t_j)$ is a white Gaussian discrete-time process with $E\{\boldsymbol{\eta}_i(t_j)\boldsymbol{\eta}_i(t_j)^T\} = \mathbf{Q}_{t_j}$, $E\{\boldsymbol{\eta}_i(t_j)\} = \mathbf{0}$ and $E\{\boldsymbol{\eta}_i(t_j)\boldsymbol{\eta}_i(t_j)^T\} = \mathbf{0}$.

The analysis update or data assimilation step can be performed by using either a stochastic or a deterministic filtering scheme. In stochastic (or fully Monte Carlo) schemes, an ensemble of observation vectors is considered, where the i-th virtual observation is defined as

$$\mathbf{y}_i(t_{j+1}) = \mathbf{y}(t_{j+1}) + \boldsymbol{\epsilon}_i^o(t_{j+1}) \tag{210}$$

where in the case of a linear observation operator we have

$$\mathbf{y}(t_{j+1}) = \mathbf{H}\mathbf{x}(t_{j+1}) + \boldsymbol{\epsilon}^o(t_{j+1}) \tag{211}$$

where $\boldsymbol{\epsilon}^o(t_{j+1})$ is a white Gaussian discrete-time process defined as in Eq. 96. The update can be written as

$$\mathbf{x}_i^a(t_{j+1}) = \mathbf{x}_i^f(t_{j+1}) + \mathbf{K}_e(\mathbf{y}_i(t_{j+1}) - \mathbf{H}\mathbf{x}_i^f(t_{j+1})) \tag{212}$$

where $\mathbf{K}_e$ is analogous to the Kalman gain (see Eq. 141) but with $\mathbf{P}^f$ substituted with $\mathbf{P}_e^f$ and $\mathbf{R}$ substituted with $\mathbf{R}_e$. In this way, the analysis $\overline{\mathbf{x}^a}(t_{j+1})$ and its error covariance $\mathbf{P}_e^a$ can be found as

$$\overline{\mathbf{x}^a}(t_{j+1}) = \frac{1}{N}\sum_{i=1}^{N}\overline{\mathbf{x}_i^a}(t_{j+1}) = \overline{\mathbf{x}^f}(t_{j+1}) + \mathbf{K}_e(\mathbf{y}(t_{j+1}) + \mathbf{H}\overline{\mathbf{x}^f}(t_{j+1})) \tag{213}$$

$$\mathbf{P}_e^a = \frac{1}{N-1}\mathbf{X}^a(t_{j+1})(\mathbf{X}^a(t_{j+1}))^T \tag{214}$$

$$\tag{215}$$

where the i-th column of $\mathbf{X}^a(t_{j+1})$ is given by $\boldsymbol{\epsilon}_i^a \equiv \mathbf{x}_i^a(t_{j+1}) - \overline{\mathbf{x}^a}(t_{j+1})$. We have

$$\boldsymbol{\epsilon}_i^a = (\mathbf{I} - \mathbf{K}_e\mathbf{H})(\mathbf{x}_i^f - \overline{\mathbf{x}^f}) + \mathbf{K}_e(\mathbf{y}_i - \mathbf{y}) = (\mathbf{I} - \mathbf{K}_e\mathbf{H})\boldsymbol{\epsilon}_i^f + \mathbf{K}_e\boldsymbol{\epsilon}_i^o. \tag{216}$$

It is important to note that $\lim_{N\to\infty}\boldsymbol{\epsilon}_i^a = \mathbf{0}$, so that $\overline{\mathbf{x}^a}$ tends to the Kalman filter analysis. Also, note that we can write

$$\mathbf{P}_e^a = \frac{1}{N-1}\sum_{i=1} N\boldsymbol{\epsilon}_i^a(\boldsymbol{\epsilon}_i^a)^T = (\mathbf{I} - \mathbf{K}_e\mathbf{H})\mathbf{P}_e^f(\mathbf{I} - \mathbf{K}_e\mathbf{H})^T + \mathbf{K}_e\mathbf{R}_e\mathbf{K}_e^T. \tag{217}$$

so that $\lim_{N\to\infty}\mathbf{P}_e^a = \mathbf{P}^a$, where $\mathbf{P}^a$ is the analysis error covariance determined with the Kalman filter.

Note that for calculating the analysis ensemble is not necessary to calculate $\mathbf{P}_e^f$, which is a $n \times n$ matrix. If we define $\mathbf{S} = \mathbf{H}\mathbf{X}^{f\prime} \in \mathbb{R}^{m\times N}$ and $\mathbf{C}_e = \mathbf{S}\mathbf{S}^T + (N-1)\mathbf{R}_e \in \mathbb{R}^{m\times m}$ (where m is the number of observations), we can write Eq. 212 as

$$\mathbf{x}_i^a = \mathbf{x}_i^f + \mathbf{X}^{f\prime}\mathbf{S}^T\mathbf{C}_e^{-1}(\mathbf{y}_i - \mathbf{H}\mathbf{x}^f) \tag{218}$$

so that to calculate the analysis we need to invert a $m \times m$ matrix $\mathbf{C}_e$ and compute the $m \times N$ matrix $\mathbf{S}$, but we never need to calculate $\mathbf{P}_e^f$. Note that in the case when observation errors are mutually uncorrelated, observations can be processed serially so that $\mathbf{C}_e$ becomes a scalar. Eq. 218 represents the ensemble Kalman filter (EnKF).

6.2 Ensemble square-root filter

The main difficulty with the stochastic algorithms is the fact that the observation error covariance estimated with a set of N ensemble members may not reflect the correct observation error characteristics of the instrument and provides an additional source of sampling error (i.e., error due to the use of a finite size ensemble). It is possible to avoid perturbing the observations by using a so-called *deterministic* formulation of the ensemble filter. This is based on the square-root formulation of the Kalman filter seen in section 4.7. If we rewrite Eq. 170 with the ensemble representation $\mathbf{P}_e^f$ and $\mathbf{P}_e^a$ we can write

$$\mathbf{X}^{a\prime}\mathbf{X}^{a\prime T} = \mathbf{X}^{f\prime}(\mathbf{I} - \mathbf{S}^T\mathbf{C}^{-1}\mathbf{S})\mathbf{X}^{f\prime T} \tag{219}$$

where now $\mathbf{C}$ is defined as $\mathbf{C} = \mathbf{S}\mathbf{S}^T + (N-1)\mathbf{R}$ and $\mathbf{R}$ is the full-rank observation error covariance. The analysis update can be determined by averaging the analysis ensemble member in Eq. 218, using $\mathbf{C}$ rather than $\mathbf{C}_e$:

$$\overline{\mathbf{x}^a} = \overline{\mathbf{x}^f} + \mathbf{X}^{f\prime}\mathbf{S}^T\mathbf{C}^{-1}(\mathbf{y} - \mathbf{H}\overline{\mathbf{x}^f}). \tag{220}$$

When $\mathbf{R}$ is full rank, $\mathbf{C}$ is also full rank so that it can be written as $\mathbf{C} = \mathbf{Z}\mathbf{\Lambda}^{-1}\mathbf{Z}^T$, where $\mathbf{\Lambda}$ is full rank. Eq 219 can written as

$$\begin{aligned}
\mathbf{X}^{a\prime}\mathbf{X}^{a\prime T} &= \mathbf{X}^{f\prime}(\mathbf{I} - \mathbf{S}^T\mathbf{Z}\mathbf{\Lambda}^{-1}\mathbf{Z}^T\mathbf{S})\mathbf{X}^{f\prime T} &(221)\\
&= \mathbf{X}^{f\prime}[\mathbf{I} - (\mathbf{\Lambda}^{-1/2}\mathbf{Z}^T\mathbf{S})^T(\mathbf{\Lambda}^{-1/2}\mathbf{Z}^T\mathbf{S})]\mathbf{X}^{f\prime T} &(222)\\
&= \mathbf{X}^{f\prime}(\mathbf{I} - \mathbf{X}_2^T\mathbf{X}_2)\mathbf{X}^{f\prime T}. &(223)
\end{aligned}$$

We can now consider the singular value decomposition of $\mathbf{X}_2$, as $\mathbf{X}_2 = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, so that we can write

$$\mathbf{X}^{a\prime}\mathbf{X}^{a\prime T} = \mathbf{X}^{f\prime}\mathbf{V}(\mathbf{I} - \mathbf{\Sigma}^T\mathbf{\Sigma})\mathbf{V}^T\mathbf{X}^{f\prime T}. \tag{224}$$

This means that the analysis perturbation matrix can be written as

$$\mathbf{X}^{a\prime} = \mathbf{X}^{f\prime}\mathbf{V}\sqrt{\mathbf{I} - \mathbf{\Sigma}^T\mathbf{\Sigma}}\mathbf{W}^T \tag{225}$$

where $\mathbf{W}$ is an orthonormal matrix. Note that when $\mathbf{W} = \mathbf{V}$ the expression multiplying $\mathbf{X}^{f\prime}$ on the right is symmetric: it is possible to show that this choice leads to an unbiased ensemble analysis error.

6.3 Nonlinear observation operator

When the observation operator is nonlinear, Eq. 211 can be written as

$$\mathbf{y}(t_{j+1}) = H(\mathbf{x}(t_{j+1})) + \boldsymbol{\epsilon}^o(t_{j+1}) \tag{226}$$

where $\boldsymbol{\epsilon}^o(t_{j+1})$ is still assumed to be a white Gaussian discrete-time process defined as in Eq. 96. By analogy with the linear case, it is possible to define a matrix $\mathbf{S}_{\text{nl}}$ whose i-th column is given by $H(\mathbf{x}_i) - \overline{H(\mathbf{x}_i)}$. In this way, the i-th analysis ensemble member can be defined as (see Eq. 218)

$$\mathbf{x}_i^a = \mathbf{x}_i^f + \mathbf{X}^{f\prime}\mathbf{S}_{\text{nl}}^T\mathbf{C}_e^{-1}(\mathbf{y}_i - \mathbf{H}\mathbf{x}^f) \tag{227}$$

with $\mathbf{C}_e = \mathbf{S}_{\text{nl}}\mathbf{S}_{\text{nl}}^T + (N-1)\mathbf{R}_e$. However, the mean of $\mathbf{x}_i^a$ in general does not converge for $N \to \infty$ to the maximum a posteriori solution estimate (i.e., the conditional mode of the

posterior density) given by the extended Kalman filter, unless the observation operator is linear. In order to achieve a consistent estimate with the ensemble Kalman filter and a nonlinear observation operator we need to abandon the strategy of seeking a solution in closed form and adopt an iterative minimization strategy similar to that used to derive the extended Kalman filter. In this case, Eq. 227 can provide a useful first guess.

6.4 Model error

When model error is given by a white Gaussian discrete time process, it can be included in the EnKF forecast step in a straightforward way (see Eq. 209). However, it is also possible to consider model error in the form of coloured noise. To this end, let us first consider again the stochastic differential equation 174, where now $d\boldsymbol{\beta}_t$ is written as

$$d\boldsymbol{\beta}_t = -\theta(\boldsymbol{\beta}_t - \boldsymbol{\mu})dt + \sigma d\boldsymbol{\beta}'_t \tag{228}$$

where $\boldsymbol{\beta}'_t$ is a Brownian motion. Eq. 228 represents a *Ornstein Uhlenbeck process*. In the case when the initial conditions $\boldsymbol{\beta}_0$ are such that $E\{\boldsymbol{\beta}_0\} = \boldsymbol{\mu}$ and $E\{(\boldsymbol{\beta}_0 - \boldsymbol{\mu})^2\} = \sigma^2/(2\theta)$, it is possible to show that

$$E\{\boldsymbol{\beta}_t\} = \boldsymbol{\mu} \tag{229}$$

$$\mathrm{cov}(\boldsymbol{\beta}_s, \boldsymbol{\beta}_t) = \frac{\sigma^2}{2\theta} e^{(-\theta|s-t|)} \tag{230}$$

$$. \tag{231}$$

This means that under the aforementioned initial conditions requirements, the Ornstein Uhlenbeck process is a stationary process. When the initial conditions are constant, it is possible to show (by computing the conditional mean and covariance of $\boldsymbol{\beta}_t$) that the process is asymptotically stationary. This means that a coloured noise can be thought of as a Ornstein Uhlenbeck process. The discrete time equivalent of an a Ornstein Uhlenbeck process is given by the first-order autoregressive process. In the discrete case, model error at time t_j can be written as (see Eq. 209)

$$\boldsymbol{\eta}(t_j) = \mu + \rho(\boldsymbol{\eta}(t_{j-1} - \mu) + \sqrt{1-\rho^2}\mathbf{w}_j. \tag{232}$$

Over the time interval $[0, T]$, the components of $\mathbf{w}_j$ are Gaussian random variables with zero mean and variance $\sigma^2/(2\theta)$ and $\boldsymbol{\eta}(t_0) = \boldsymbol{\mu} + \mathbf{w}_0$. For $T = N\Delta t$ we define $\rho = \exp(-\theta\Delta t)$. This is a coloured noise in the time interval $[0, T]$.

6.5 Localization and inflation

The rank of the sample covariance matrix defined in Eq. 203 is at most equal to the minimum between the size of the state space and the number of ensemble members minus one. When the number of ensemble members is much less than the size of the state space, the sample covariance may not provide an accurate representation of long range correlations, leading to spurious analysis increments. In Fig. 3 (a), spurious correlation that are distant from a given grid point are evident, in comparison to those arising from the use of a considerably larger number of ensemble members, as in Fig. 3 (b).

6.6 Discussion

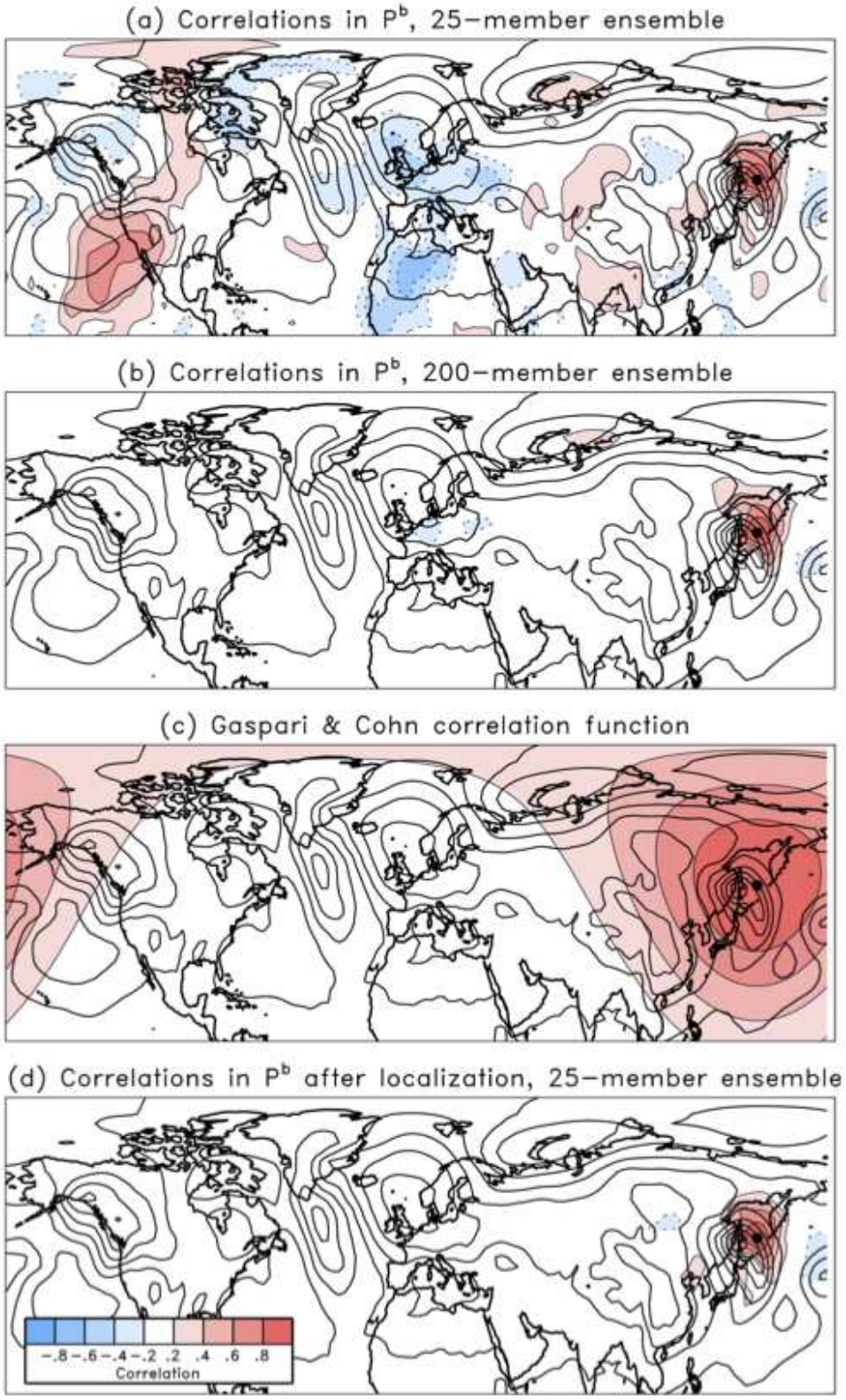

Figure 3: (a) $\mathbf{P}_e^f$ ($N = 25$); (b) $\mathbf{P}_e^f$ ($N = 100$); (c) Correlation function with compact support; (d) localized $\mathbf{P}_e^f$ ($N = 25$); From Fig. 6.4 of Hamill, 2006.