Chapter 20
Data Assimilation

Data assimilation is a powerful methodology for incorporating data into a mathematical model in a way that is consistent with physical processes. In this chapter we present a Bayesian approach to data assimilation, discuss the Kalman filter algorithm for data assimilation for linear processes and linear data models in the presence of Gaussian noise, and describe a simple version of the ensemble Kalman filter, which relies on stochastic simulation.

Keywords: Process model, data model, probability distribution, uncertainty, Bayesian statistics, reanalysis, filtering, prediction, sequential data assimilation, Markov chain, Kalman filter, extended Kalman filter, ensemble Kalman filter, Lorenz system, predictability.

20.1 ▪ Data Assimilation and Climate

To predict the state of the climate system at any future time or to recreate its past, we must necessarily rely on mathematical models and numerical simulations. Throughout this book, we have encountered several types of models, some of them conceptual and others closer to the "real" world of physics, chemistry, and biology. Most of these models are *process models*—models that are described by systems of equations that determine the state variables (also called *process variables*) and their evolution in time. In *data assimilation*, one links such process models with *data models*—models of observational data of these same state variables, together with their uncertainties. The idea is that, by making it consistent with the available observations, a process model becomes better at predicting future states of the system. Data assimilation is an essential technique in any scientific discipline that is data-rich and for which well-founded predictive mathematical models exist. The technique originated in engineering and has found widespread application in many other disciplines, most notably in weather prediction, where it has extended the ability to predict weather more or less accurately from hours to days. Not only does the technique generate estimates of the state of the weather system, but it also produces an assessment of the uncertainties in the prediction, often in the form of probability distributions for process variables or parameters. For an overview of data assimilation techniques in weather prediction and climate science, we refer the reader to [48, 89].

A typical example of a data model is the record of SSTs obtained with a drifting instrument that periodically reports local measurements. The record contains information about the temperature at the location of the instrument (but not anywhere else). The

record may have gaps for times when the instrument is shut down, and there may be uncertainties regarding the drifter's location and due to limited accuracy of the thermometer. Any data model contains an element of randomness.

Data assimilation can be applied to estimate process variables at a certain time using all available observational data, including those made at a later time (*reanalysis* or *smoothing* mode), or to estimate the present state using past and present observations (*analysis* or *filtering* mode), or to estimate process variables that are inaccessible to observations, such as future states or states between measurements (*forecasting* or *predicting* mode). In any of these modes, problems can be approached with a variety of techniques, including optimization methods, which attempt to find the best fit of a parameterized process model to a given set of data; maximum likelihood methods, which work in a similar spirit but have more detailed statistical models for the uncertainties in the process; and Bayesian methods.

20.2 ▪ Example

The following example (adapted from [121]) illustrates the application of data assimilation methodology to reanalysis, filtering, and forecasting.

Consider a process with four real-valued state variables, $\{X_i : i = 1,\dots,4\}$. The state variables are related by the process model

$$X_{i+1} = \alpha X_i + \xi_i, \quad i = 1,2,3, \tag{20.1}$$

where α is a known positive constant and the random process error terms ξ_i are either identically zero or have a standard normal distribution, $\xi_i \sim N(0,1)$. (The convention is to use upper-case letters (X, Y, Z, etc.) for random quantities and the corresponding lower-case letters (x, y, z, etc.) for their realizations. Random error terms are indicated by Greek letters.)

The data model consists of two observations, $\{Y_i : i = 2,3\}$, and is related to the process model through the identities

$$Y_i = X_i + \zeta_i, \quad i = 2,3, \tag{20.2}$$

where the random observational error terms ζ_i are independent and identically distributed with a standard normal distribution, $\zeta_i \sim N(0,\tau^2)$. The entire model is shown schematically in Figure 20.1. The problem of estimating X_1 from the observations Y_2 and Y_3 is a reanalysis problem, estimating X_2 from Y_2 or X_3 from Y_2 and Y_3 is a filtering problem, and estimating X_4 from Y_2 and Y_3 is a forecasting problem.

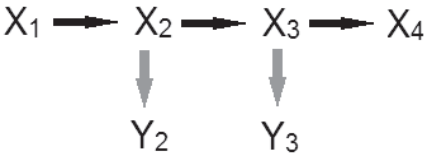

Figure 20.1. *A simple process model (black arrows) together with a data model (gray arrows).*

20.2.1 ▪ Variational Approach

We first demonstrate a variational approach. If the process error terms ξ_i are all zero, we minimize the *cost function* J,

$$J(x_2,x_3;y_2,y_3) = (x_2 - y_2)^2 + (x_3 - y_3)^2. \tag{20.3}$$

Because of the process model relation $x_3 = \alpha x_2$, J reduces to a function J_0 of x_2 alone,

$$J(x_2, \alpha x_2; y_2, y_3) = J_0(x_2; y_2, y_3) = (x_2 - y_2)^2 + (\alpha x_2 - y_3)^2. \tag{20.4}$$

This function reaches its minimum at $x_2^* = (y_2 + \alpha y_3)/(1 + \alpha^2)$, so x_2^* is the reanalysis value of x_2. The corresponding filtering value of x_3 is $x_3^* = \alpha x_2^*$, the forecast value of x_4 is $x_4^* = \alpha x_3^* = \alpha^2 x_2^*$, and the reanalysis value of x_1 is $x_1^* = \alpha^{-1} x_2^*$. We obtain the same values if we use the process model relation $x_3 = \alpha x_2$ to eliminate x_2 from the cost function (20.3) and minimize with respect to x_3.

These solutions do not give any uncertainty estimates (confidence sets) for the reanalysis and filtering values. In this case, such estimates can be inferred from the data model (20.2) and the explicit form of the solution. However, in the general case, the reanalysis and filtering values are obtained numerically and uncertainty estimates are not immediately available. If the process error terms ξ_i do not vanish, the cost function must be extended with additional terms. We refer the reader to [89] for a more detailed discussion of the variational approach and its relation to the probabilistic approaches discussed in the next sections.

20.2.2 ▪ Maximum Likelihood Approach

Next we demonstrate a maximum likelihood approach to the same example. Assume that the process error variables ξ_i are random and have standard normal distributions. Assume for the moment that X_1 has a fixed but unknown value x_1, which we wish to estimate from the observations Y_2 and Y_3. This is therefore a reanalysis problem. Probability theory shows that

$$Y_2 \sim N(\alpha x_1, 1 + \tau^2), \quad Y_3 \sim N(\alpha^2 x_1, 1 + \tau^2 + \alpha^2), \quad \operatorname{cov}(Y_2, Y_3) = \alpha.$$

The joint distribution of (Y_2, Y_3) is Gaussian with density

$$f_{Y_2,Y_3}(y_2, y_3) \propto \exp\left(-\tfrac{1}{2}(y_2 - \alpha x_1, y_3 - \alpha^2 x_1)\Sigma^{-1}(y_2 - \alpha x_1, y_3 - \alpha^2 x_1)^T\right), \tag{20.5}$$

where Σ is the covariance matrix,

$$\Sigma = \begin{pmatrix} 1 + \tau^2 & \alpha \\ \alpha & 1 + \tau^2 + \alpha^2 \end{pmatrix}.$$

The proportionality constant implied in Eq. (20.5) does not depend on the unknown parameter x_1. If y_2 and y_3 are actual observational data, the maximization of the expression on the right-hand side of Eq. (20.5) leads to the maximum likelihood estimate

$$\hat{x}_1 = \frac{\alpha(1 + \tau^2)y_2 + \alpha^2 \tau^2 y_3}{\alpha^2 + \alpha^2 \tau^2 + \tau^4}. \tag{20.6}$$

The reanalysis estimate $\hat{x}_1$ is a linear combination of the observations y_2 and y_3. If $0 < \alpha < 1$, y_2 has the larger weight. If, furthermore, the observation errors ζ_i have very small standard deviations ($\tau \ll 1$), the weight for y_3 is very small, so the reanalysis estimate depends mainly on the observation that was made right after the unknown state. One can show that $\mathcal{E}\hat{x}_1 = x_1$, so the estimate is unbiased, and it is not hard to show that $\hat{x}_1$ has a normal distribution and to find its variance.

Approaches to estimating x_2 (reanalysis) and x_3 (filtering) from observations using the maximum likelihood method are discussed in the exercises. We next introduce a general Bayesian approach to data assimilation and then return to this example.

20.3 ▪ Bayesian Approach

The contemporary approach to analysis and forecasting problems is based on Bayes' rule. This rule was first formulated by the English mathematician and Presbyterian minister THOMAS BAYES (1701–1761). It yields estimates that are asymptotically correct and does not require an appeal to the law of large numbers, which would make little sense in the climate context. To simplify the following presentation, we assume that all state variables and data are finite-dimensional vectors—a reasonable assumption for data but a substantial simplification for state variables.

Suppose we are interested in estimating a vector $\mathbf{X}$ of process variables with a known or assumed pdf f_X. The distribution may come from long-term observations or from another forecast model and is referred to as the *prior distribution* on $\mathbf{X}$. The data model uses a vector $\mathbf{Y}$ of observations of (the components of) $\mathbf{X}$, which may also include other random effects. We assume that for each possible realization $\mathbf{x}$ of $\mathbf{X}$, the conditional distribution $f_{Y|x}$ of $\mathbf{Y}$ given $\mathbf{x}$ is known. This is essentially the data model.

Now, suppose that the observation of $\mathbf{Y}$ at a particular instance results in the value $\mathbf{y}$. The goal is to incorporate this value in the distribution of $\mathbf{X}$ by constructing the conditional distribution $f_{X|y}$ of $\mathbf{X}$ given $\mathbf{y}$. This conditional distribution is referred to as the *posterior distribution* on $\mathbf{X}$. To find its formula, we note that the joint probability density $f_{X,Y}$ of process variables and observations can be expressed in two ways,

$$f_{X,Y}(\mathbf{x},\mathbf{y}) = f_{Y|x}(\mathbf{y})f_X(\mathbf{x}) = f_{X|y}(\mathbf{x})f_Y(\mathbf{y}). \tag{20.7}$$

The prior distribution f_X is assumed to be known, as is the data model $f_{Y|x}$; the posterior distribution $f_{X|y}$ is sought, and the distribution of the observations f_Y is unknown. After dividing both expressions by $f_Y(\mathbf{y})$, we obtain *Bayes' rule*,

$$f_{X|y}(\mathbf{x}) = \frac{f_{Y|x}(\mathbf{y})f_X(\mathbf{x})}{f_Y(\mathbf{y})}. \tag{20.8}$$

Note that both sides depend on $\mathbf{x}$ and $\mathbf{y}$. The quantity $\mathbf{y}$ is given as an observation; therefore the denominator on the right-hand side is fixed, although unknown. The entire equation has the form

$$f_{X|y}(\mathbf{x}) \propto f_{Y|x}(\mathbf{y})f_X(\mathbf{x}), \tag{20.9}$$

where the implied proportionality constant (depending on f_Y) makes the term on the left a pdf. The constant can be found and equality established by computing an integral, either with analytical techniques or with numerical simulations.

In the case of Gaussian random variables, all integrations that would be required in Eq. (20.9) can, however, be replaced by matrix algebra, as the following lemma shows.

Lemma 20.1. *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be Gaussian random variables such that* $\mathbf{X} \sim N(\mu, P)$ *and* $\mathbf{Y}|\mathbf{X} \sim N(H\mathbf{X}, R)$*, where* H *is a matrix. Then* $\mathbf{X}|\mathbf{Y} \sim N(\mu^*, P^*)$ *with* $\mu^* = \mu + K(\mathbf{Y} - H\mu)$ *and* $P^* = (I - KH)P$*, where* K *is the* gain matrix*,*

$$K = PH^T(R + HPH^T)^{-1}. \tag{20.10}$$

Proof. According to Eq. (20.9), the conditional probability density $f_{X|y}$ satisfies

$$f_{X|y}(\mathbf{x}) \propto \exp\left(-\tfrac{1}{2}\left((\mathbf{x} - H\mathbf{y})^T R^{-1}(\mathbf{x} - H\mathbf{y}) - (\mathbf{y} - \mu)^T P^{-1}(\mathbf{y} - \mu)\right)\right). \tag{20.11}$$

Completing the square, we obtain a Gaussian distribution,

$$f_{X|y}(\mathbf{x}) \propto \exp\left(-\tfrac{1}{2}\left((\mathbf{x}-\mu^*)^T(P^*)^{-1}(\mathbf{x}-\mu^*)\right)\right), \tag{20.12}$$

with

$$\mu^* = \mathcal{E}(\mathbf{X}|\mathbf{y}) = (P^{-1}+H^TR^{-1}H)^{-1}\left(P^{-1}\mu+R^{-1}H^T\mathbf{y}\right) = \mu+K(\mathbf{y}-H\mu),$$
$$P^* = \mathrm{var}(\mathbf{X}|\mathbf{y}) = \left(P^{-1}+H^TR^{-1}H\right)^{-1} = (I-KH)P.$$

We leave the details of these calculations to the reader. □

The distribution $f_{X|y}$ given in the lemma is the posterior distribution. Its covariance matrix P^* does not depend on the value $\mathbf{y}$. The lemma tells us that $P^* = P-C$, where C is a symmetric positive semidefinite matrix. In this sense, the posterior variance is smaller than the prior variance, and we have reduced uncertainty by using the data. The lemma also shows that the mean of the posterior distribution is the prior mean updated by the gain applied to the difference between the observed $\mathbf{Y}$ and its mean value $H\mu$. Recall that the difference between an observed quantity and its temporal mean is called "anomaly" in climate science, so this concept arises naturally in Bayesian data assimilation.

20.3.1 ▪ Example: Bayesian Approach

We return to the example introduced in Section 20.2. Assume that $X_1 \sim N(\mu_0, \sigma^2)$, where μ_0 and σ^2 are known. Then the column vector $\mathbf{X} = (X_1, \ldots, X_4)^T$ of process variables has a multivariate normal distribution $\mathbf{X} \sim N(\mu, \Sigma)$ with mean vector $\mu = (\mu_0, \alpha\mu_0, \alpha^2\mu_0, \alpha^3\mu_0)^T$ and a suitable covariance matrix Σ (see the exercises). This is the prior distribution f_X on $\mathbf{X}$. It does not use any observations. Explicitly,

$$f_X(\mathbf{x}) \propto \exp\left(-\tfrac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right).$$

The column vector $\mathbf{Y} = (Y_2, Y_3)^T$ of observations and the vector $\mathbf{X}$ of process variables are related by the equation $\mathbf{Y} = H\mathbf{X} + (\zeta_2, \zeta_3)^T$, where $H = \left(\begin{smallmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{smallmatrix}\right)$. Given a particular realization $\mathbf{x}$ of the process variables, $\mathbf{Y}$ then has a multivariate normal distribution, $\mathbf{Y}|\mathbf{x} \sim N(H\mathbf{x}, \tau^2 I)$, where I is the 2×2 identity matrix. All quantities can be computed from the gain matrix K, which will also be derived in the exercises.

It is instructive to compare the standard deviations of the prior and posterior distributions for X_1 (reanalysis) and for X_3 (filtering). These quantities are computed in Exercise 4. It turns out that $\mathrm{var}(X_1|\mathbf{y}) < \mathrm{var}(X_1) = \sigma^2$ and $\mathrm{var}(X_3|\mathbf{y}) < \mathrm{var}(X_3) = \sigma^2\alpha^4 + \alpha^2 + 1$, as expected. However, $\mathrm{var}(X_1|\mathbf{y})$ cannot be made arbitrarily small, even if τ is small, while $\mathrm{var}(X_3|\mathbf{y}) = O(\tau^2)$. Details are in the exercises.

20.4 ▪ Sequential Data Assimilation

We now focus on reanalysis, filtering, and forecasting for time-dependent processes. The data arrive as a *time series*—a sequence of realizations of a *discrete-time stochastic process*.

Suppose the process variables are $\mathbf{X}(k)$, $k = 0, 1, \ldots$, and the observations are $\mathbf{Y}(k)$, $k = 1, 2, \ldots$. A starting value $\mathbf{X}(0)$ for the process variables is allowed to incorporate a background state for which no observations are available. We use the notation $\mathbf{X}(0:N)$ for a sequence $\{\mathbf{X}(k) : k = 0, 1, \ldots, N\}$ of vector-valued random variables and $\mathbf{x}(0:N)$ for a sequence of its realizations, and similarily for $\mathbf{Y}$. We are interested in estimating $\mathbf{X}(n)$, given $\mathbf{Y}(1:N)$ (reanalysis), or given $\mathbf{Y}(1:n)$ (filtering), or given $\mathbf{Y}(1:n-1)$ (forecasting).

Joint pdfs (also called probability mass functions) and conditional pdfs are identified by suitable indices. For example, $f_{\mathbf{X}(0:n),\mathbf{Y}(1:n)}(\mathbf{x}(0:n),\mathbf{y}(1:n))$ is the joint density function of the process variables $\mathbf{X}(0),\dots,\mathbf{X}(n)$ and observations $\mathbf{Y}(1),\dots,\mathbf{Y}(n)$, and $f_{\mathbf{X}(2:n)|\mathbf{y}(1:n-1)}(\mathbf{x}(2:n))$ is the conditional pdf for $\mathbf{X}(2),\dots,\mathbf{X}(n)$ given observations $\mathbf{y}(1),\dots,\mathbf{y}(n-1)$. The latter is a function of $\mathbf{y}(1:n-1)$ and $\mathbf{x}(2:n)$. Since observations arrive sequentially, one can try to find forecast and filter estimates also sequentially.

20.4.1 ▪ Filtering and Forecasting for Markov Chains

Definition 20.1. *A discrete-time stochastic process* $\mathbf{X}(0:N)$ *has the* Markov property *if its pdfs satisfy*

$$f_{\mathbf{X}(n:N)|\mathbf{x}(0:n-1)}(\mathbf{x}(n:N)) = f_{\mathbf{X}(n:N)|\mathbf{x}(n-1)}(\mathbf{x}(n:N)), \quad n = 1,\dots,N, \tag{20.13}$$

for all $\mathbf{x}(n:N)$. *A discrete-time stochastic process that has the Markov property is called a* Markov chain.

Intuitively, the Markov property says that, to predict future observations $\mathbf{X}(n:N)$ of the stochastic process, it is sufficient to know the immediate past $\mathbf{X}(n-1)$. Additional knowledge of the more distant past $\mathbf{X}(0:n-2)$ does not change the predictions of the future. An induction argument shows that the joint distribution of $\mathbf{X}(0:n)$ can then be written as a product of conditional distributions,

$$f_{\mathbf{X}(0:n)}(\mathbf{x}(0:n)) = f_{\mathbf{X}(0)}(\mathbf{x}(0))\prod_{j=1}^{n} f_{\mathbf{X}(j)|\mathbf{x}(j-1)}(\mathbf{x}(j)), \quad n = 1,2,\dots. \tag{20.14}$$

The functions $f_{\mathbf{X}(j)|\mathbf{x}(j-1)}(\mathbf{x}(j))$ are called *transition probabilities* of the Markov chain. We shall also assume throughout that

$$f_{\mathbf{Y}(1:n)|\mathbf{x}(0:n)}(\mathbf{y}(1:n)) = \prod_{j=1}^{n} f_{\mathbf{Y}(j)|\mathbf{x}(j)}(\mathbf{y}(j)), \quad n = 1,2,\dots. \tag{20.15}$$

This identity implies that, given the sequence of process variables $\mathbf{X}(0:n)$, the observations $\mathbf{Y}(1:n)$ are independent of one another and their distributions do not depend on $\mathbf{X}(0)$. In particular, it follows that given $\mathbf{x}(i)$, the observation $\mathbf{Y}(i)$ is independent of all other observations $\mathbf{Y}(j)$ $(j \neq i)$.

We can now formulate a basic algorithm for filtering and forecasting of Markov chains.

Algorithm 20.1 (Filtering and prediction). *Given*
(i) *a prior distribution* $f_{\mathbf{X}(0)}$,
(ii) *transition probabilities* $f_{\mathbf{X}(i)|\mathbf{x}(i-1)}$ $(i = 1,\dots,N)$ *for the process variables, and*
(iii) *conditional distributions* $f_{\mathbf{Y}(i)|\mathbf{x}(i)}$ $(i = 1,\dots,N)$ *for the observations,*
the following algorithm gives the filtering distributions $f_{\mathbf{X}(i)|\mathbf{y}(1:i)}$ *and forecasting distributions* $f_{\mathbf{X}(i)|\mathbf{y}(1:i-1)}$:

Step 1. *Set*

$$f_{\mathbf{X}(1)}(\mathbf{x}(1)) = \int f_{\mathbf{X}(1)|\mathbf{x}(0)}(\mathbf{x}(1))f_{\mathbf{X}(0)}(\mathbf{x}(0))d\mathbf{x}(0),$$
$$f_{\mathbf{X}(1)|\mathbf{y}(1)}(\mathbf{x}(1)) \propto f_{\mathbf{Y}(1)|\mathbf{x}(1)}(\mathbf{y}(1))f_{\mathbf{X}(1)}(\mathbf{x}(1)),$$

where the proportionality constant is chosen such that a pdf with respect to $\mathbf{x}(1)$ *is generated.*

Step 2. *Suppose* $i \in \{2,3,\dots,N\}$ *and the filtering pdf* $f_{\mathbf{X}(i-1)|\mathbf{y}(1:i-1)}$ *is given. Set*

$$f_{\mathbf{X}(i)|\mathbf{y}(1:i-1)}(\mathbf{x}(i)) = \int f_{\mathbf{X}(i)|\mathbf{x}(i-1)}(\mathbf{x}(i)) f_{\mathbf{X}(i-1)|\mathbf{y}(1:i-1)}(\mathbf{x}(i-1)) d\mathbf{x}(i-1),$$

$$f_{\mathbf{X}(i)|\mathbf{y}(1:i)}(\mathbf{x}(i)) \propto f_{\mathbf{Y}(i)|\mathbf{x}(i)}(\mathbf{y}(i)) f_{\mathbf{X}(i)|\mathbf{y}(1:i-1)}(\mathbf{x}(i)),$$

where the proportionality constant is chosen such that a pdf with respect to $\mathbf{x}(i)$ *is generated.*

The correctness of this algorithm can be proved with induction. Step 1 is just the law of total probability (to obtain $f_{\mathbf{X}(1)}$) and Bayes' rule (to obtain $f_{\mathbf{X}(1)|\mathbf{y}(1)}$). For the induction step, the formula for the forecasting distribution is again just the law of total probability, and the filtering distribution can be obtained from

$$\begin{aligned} f_{\mathbf{X}(i)|\mathbf{y}(1:i)}(\mathbf{x}(i)) &= f_{\mathbf{X}(i)|\mathbf{y}(i),\mathbf{y}(1:i-1)}(\mathbf{x}(i)) \\ &\propto f_{\mathbf{Y}(i)|\mathbf{x}(i),\mathbf{y}(1:i-1)}(\mathbf{y}(i)) f_{\mathbf{X}(i)|\mathbf{y}(1:i-1)}(\mathbf{x}(i)) \\ &= f_{\mathbf{Y}(i)|\mathbf{x}(i)}(\mathbf{y}(i)) f_{\mathbf{X}(i)|\mathbf{y}(1:i-1)}(\mathbf{x}(i)). \end{aligned}$$

The last equation follows because the observations $\mathbf{Y}(i)$ were assumed to be independent of the other observations conditioned on the $\mathbf{x}(i)$ in Eq. (20.15).

Once the filtering distributions are known, the reanalysis distributions $f_{\mathbf{X}(i)|\mathbf{y}(1:N)}(\mathbf{x}(i))$ can be obtained recursively for $i = N, N-1, \dots, 0$.

Algorithm 20.2 (Reanalysis). *Given*
(i) *transition probabilities* $f_{\mathbf{X}(i)|\mathbf{x}(i-1)}$ $(i = 1,\dots,N)$ *for the process variables and*
(ii) *filtering distributions* $f_{\mathbf{X}(i)|\mathbf{y}(1:i)}$ $(i = 1,\dots,N)$,
the following algorithm gives the reanalysis distributions $f_{\mathbf{X}(i)|\mathbf{y}(1:N)}$*:*

Step 1. *If* $i = N$*, the filtering distribution and the reanalysis distribution coincide.*

Step 2. *Suppose* $i \in \{1,2,\dots,N-1\}$ *and the reanalysis pdf* $f_{\mathbf{X}(i+1)|\mathbf{y}(1:N)}$ *is given. Set*

$$f_{\mathbf{X}(i)|\mathbf{x}(i+1),\mathbf{y}(1:i)}(\mathbf{x}(i)) \propto f_{\mathbf{X}(i+1)|\mathbf{x}(i)}(\mathbf{x}(i+1)) f_{\mathbf{X}(i)|\mathbf{y}(1:i)}(\mathbf{x}(i)),$$

$$f_{\mathbf{X}(i)|\mathbf{y}(1:N)}(\mathbf{x}(i)) = \int f_{\mathbf{X}(i)|\mathbf{x}(i+1),\mathbf{y}(1:i)}(\mathbf{x}(i)) f_{\mathbf{X}(i+1)|\mathbf{y}(1:N)}(\mathbf{x}(i+1)) d\mathbf{x}(i+1),$$

where the proportionality constant is chosen such that a pdf with respect to $\mathbf{x}(i)$ *is generated.*

The correctness of this algorithm is proved with backwards induction.

Algorithms 20.1 and 20.2 provide a general framework for reanalysis, filtering, and forecasting. However, they are usually impossible to implement, due to multiple difficulties. For general process models, it is often not possible to obtain all the required transition probabilities, even in very simple cases. In the case of a process described by differential equations, a closed form solution would be required, followed by a complicated change of variables. Even if the transition probabilities were known, each step would require the computation of many integrals, one given explicitly in the algorithm and another to determine the proportionality constants. These integrations are usually impossible to do in closed form, and in the case of five or more state space dimensions they are also difficult to do numerically.

20.5 ▪ Kalman Filtering

If the prior distribution on $\mathbf{X}(0)$ is Gaussian, if the process is described by linear equations, and if the data model is also Gaussian with means that depend linearly on the pro-

cess variables, then all probability distributions in the reanalysis and filtering algorithms are also Gaussian, and all integrations reduce to matrix manipulations. The result is the famous *Kalman filtering algorithm*, first proposed by the Hungarian-American engineer and mathematician RUDOLF (RUDY) EMIL KÁLMÁN (b. 1930).

Assume that the process variables $\mathbf{X}(i) \in \mathbb{R}^n$ form a linear process model,

$$\mathbf{X}(i) = M_i \mathbf{X}(i-1) + \xi_i, \quad i = 1, \ldots, N, \tag{20.16}$$

with $\mathbf{X}(0) \sim N(\mu, \Sigma)$. (The more general process model $\mathbf{X}(i) = M_i \mathbf{X}(i-1) + \mathbf{b}(i) + \xi_i$ can be reduced to Eq. (20.16); see the exercises.) Assume, furthermore, that the data variables $\mathbf{Y}(i) \in \mathbb{R}^m$ are linearly related to the process variables,

$$\mathbf{Y}(i) = H_i \mathbf{X}(i) + \zeta_i, \quad i = 1, \ldots, N. \tag{20.17}$$

The M_i and H_i are matrices of suitable dimensions, and the $\xi_i \in \mathbb{R}^n$ and $\zeta_i \in \mathbb{R}^m$ are random variables, independent of each other and of $\mathbf{X}(0)$, distributed as $\xi_i \sim N(0, Q_i)$ and $\zeta_i \sim N(0, R_i)$. The assumptions cover the case where the dimension m_i of the ith observation $\mathbf{Y}(i)$ depends on i or where some of the $\mathbf{Y}(i)$ are absent.

Then the theory of multivariate normal distributions implies that the $\mathbf{X}(i)$ and $\mathbf{Y}(i)$ are also Gaussian, as are all conditional variables. In particular, $\mathbf{X}(1) \sim N(M_1\mu, Q_1 + M_1 \Sigma M_1^T)$. It is therefore sufficient to describe the means and covariance matrices of these variables. We use the following abbreviations:

$$\begin{aligned}
\mu_{i|i-1} &= \mathscr{E}(\mathbf{X}(i)|\mathbf{y}(1:i-1)), \quad \Sigma_{i|i-1} = \operatorname{var}(\mathbf{X}(i)|\mathbf{y}(1:i-1)), \\
\mu_{i|i} &= \mathscr{E}(\mathbf{X}(i)|\mathbf{y}(1:i)), \quad \Sigma_{i|i} = \operatorname{var}(\mathbf{X}(i)|\mathbf{y}(1:i)), \\
\mu_i &= \mathscr{E}(\mathbf{X}(i)|\mathbf{y}(1:N)), \quad \Sigma_i = \operatorname{var}(\mathbf{X}(i)|\mathbf{y}(1:N)),
\end{aligned}$$

with the convention $\mu_{0|0} = \mu$, $\mu_{1|0} = \mathscr{E}\mathbf{X}(1) = M_1\mu$ and, similarly, $\Sigma_{0|0} = \Sigma$, $\Sigma_{1|0} = Q_1 + M_1 \Sigma M_1^T$. The quantities μ are conditional means and the quantities Σ conditional covariance matrices. The subscript $i|i-1$ refers to a forecasting quantity, the subscript $i|i$ to a filtering quantity (estimate the current state based on current and past observations), and the simple subscript i to a reanalysis quantity (estimate a past state from all available data). The goal is to obtain recursions for all these quantities. Straightforward computations show that

$$\mu_{i|i-1} = M_i \mu_{i-1|i-1}, \quad \Sigma_{i|i-1} = Q_i + M_i \Sigma_{i-1|i-1} M_i^T, \quad i = 1, \ldots, N. \tag{20.18}$$

As expected, the forecasting distribution does not depend on new data.

Next, we use induction, applying Lemma 20.1 with $\mathbf{X} = \mathbf{X}(i)|\mathbf{y}(1:i-1)$ and $\mathbf{Y} = \mathbf{Y}(i)|\mathbf{y}(1:i-1)$. The data model has the property that $\mathbf{Y}(i)$ is independent of $\mathbf{Y}(k)$ for $k < i$, so $\mathbf{Y}(i)|\mathbf{y}(1:i-1) = \mathbf{Y}(i)$. According to Lemma 20.1,

$$\mu_{i|i} = \mu_{i|i-1} + K_i \left(\mathbf{y}(i) - H_i \mu_{i|i-1} \right), \tag{20.19}$$

where the *Kalman gain matrix* K_i is given by

$$K_i = \Sigma_{i|i-1} H_i^T \left(H_i^T \Sigma_{i|i-1} H_i + R_i \right)^{-1}, \tag{20.20}$$

provided the matrix inverses all exist. The term $\mathbf{y}(i) - H_i \mu_{i|i-1}$ is called the *innovation* and is conceptually similar to an anomaly in climate science. Also from Lemma 20.1, the filtering covariance matrix is

$$\Sigma_{i|i} = (I - K_i H_i) \Sigma_{i|i-1}. \tag{20.21}$$

The *Kalman filter algorithm* is obtained by alternating the forecasting and filtering step, just as in the general Algorithm 20.1.

Algorithm 20.3 (Kalman filter). *Given*
(i) *a prior distribution* $\mathbf{X}(0) \sim N(\mu, \Sigma)$,
(ii) *the process model* (20.16), *and*
(iii) *the linear data model* (20.17),
the following algorithm gives the forecasting distributions $f_{\mathbf{X}(i)|\mathbf{y}(1:i-1)}$ *and filtering distributions* $f_{\mathbf{X}(i)|\mathbf{y}(1:i)}$:

Step 1. *The forecasting and filtering distributions of* $\mathbf{X}(1)$ *are given by*

$$\mathbf{X}(1) \sim N(\mu_{1|0}, \Sigma_{1|0}),$$
$$\mathbf{X}(1)|\mathbf{y}(1) \sim N(\mu_{1|1}, \Sigma_{1|1}).$$

Step 2. *Suppose* $i \in \{2, 3, \ldots, N\}$ *and the filtering distribution at time step* $i-1$ *is known,* $\mathbf{X}(i-1)|\mathbf{y}(1:i-1) \sim N(\mu_{i-1|i-1}, \Sigma_{i-1|i-1})$. *Then the forecasting and filtering distributions at time step* i *are given by*

$$\mathbf{X}(i)|\mathbf{y}(1:i-1) \sim N(\mu_{i|i-1}, \Sigma_{i|i-1}),$$
$$\mathbf{X}(i)|\mathbf{y}(1:i) \sim N(\mu_{i|i}, \Sigma_{i|i}),$$

where $\mu_{i|i-1}$, $\mu_{i|i}$, $\Sigma_{i|i-1}$, *and* $\Sigma_{i|i}$ *are computed from Eqs.* (20.18), (20.19), *and* (20.21).

The data influence only the forecasting and filtering means but not the variance matrices which, in principle, can all be computed in advance. In practical applications, the problem of computing or estimating the covariance matrices becomes important. For those i for which no observations are available, the filtering distribution and the forecasting distribution agree.

The basic reanalysis Algorithm 20.2 can also be rewritten in terms of matrix operations for this situation [121].

20.6 ▪ Numerical Example

The following numerical example illustrates the results of Kalman filtering and reanalysis. The example uses the one-dimensional process model $x_i = \alpha x_{i-1} + \xi_i$ for $i = 1, \ldots, 30$, where $\alpha = 0.8$ and

$$\xi_i \sim N(0, q^2), \quad q = 0.4.$$

The value x_0 is drawn from a standard normal distribution. A typical sequence is plotted in black in Figure 20.2. The data model is $y_i = h_i x_i + \zeta_i$, where

$$\zeta_i \sim N(0, r^2), \quad r = 0.1; \quad h_i = \begin{cases} 0.1, & i = 11, \ldots, 20, \\ 1, & i = 1, \ldots, 10, 21, \ldots, 30. \end{cases}$$

The data model is set up so that for $i = 11, \ldots, 20$, there is a period of "low observability." Figure 20.2 shows a single realization of the true process, the filtering estimates, and the reanalysis estimates. It is clear that these estimates are not the same. As expected, all estimates are much closer to the true values where $h_i = 1$.

Figure 20.3 shows computed standard deviations (bold lines) for the forecasting estimates $x_{i|i-1}$ (black), filtering estimates $x_{i|i}$ (blue), and reanalysis estimates $x_{i|N}$ (red), together with their negative values. Forecasting standard deviations are always larger than

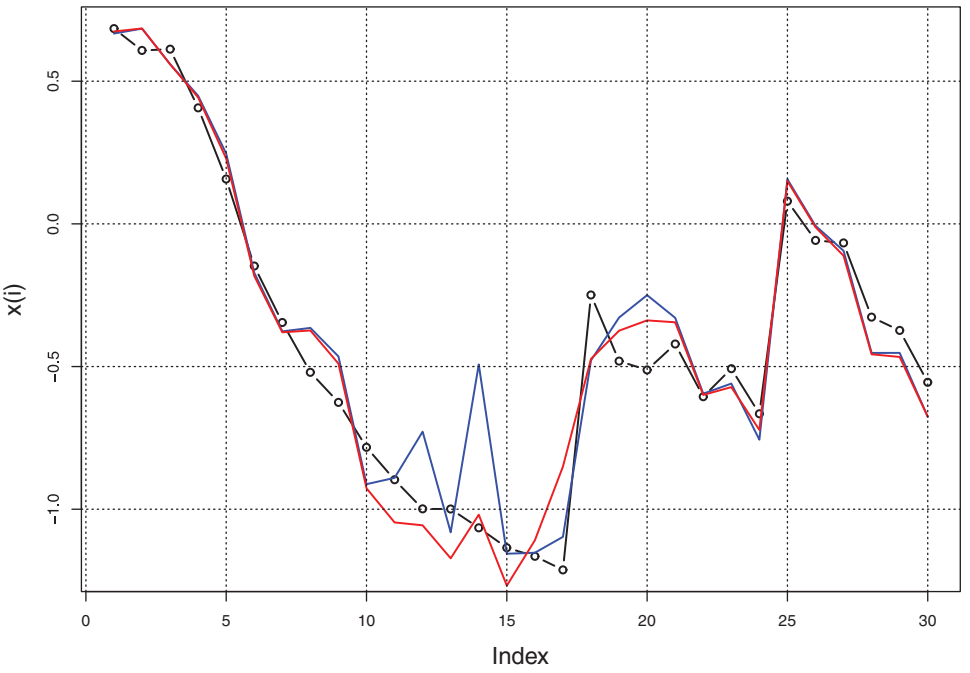

Figure 20.2. *Simulated Kalman filter example; true process (black), filtering estimate (blue), and reanalysis estimate (red).*

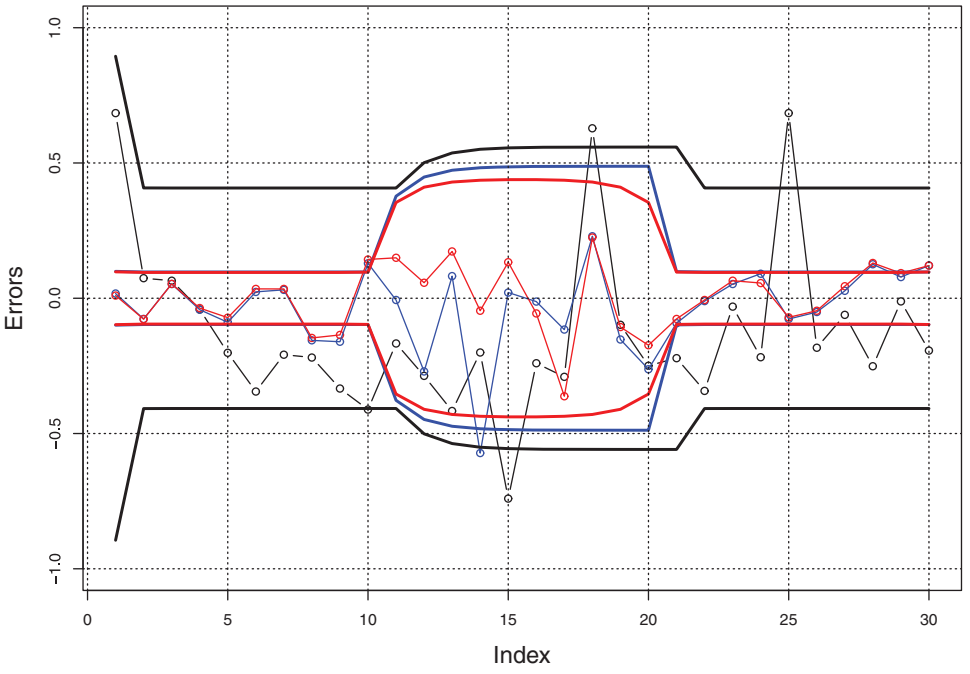

Figure 20.3. *Errors and error standard deviations for Kalman filter example; forecasting error (black), filtering error (blue), and reanalysis error (red).*

filtering standard deviations which, in turn, are larger than reanalysis standard deviations. During the interval of low observability ($i = 11, \ldots, 20$), all these standard deviations are larger. Also plotted in the same figure (thin lines) are the forecasting errors $x_i - x_{i|i-1}$ (black), filtering errors $x_i - x_{i|i}$ (blue), and reanalysis errors $x_i - x_{i|N}$ (red) for a single realization.

20.7 ▪ Extensions

The Kalman filter and reanalysis algorithms have theoretical and practical limitations if the error distributions are not Gaussian or if the process model is nonlinear. In the latter case, even Gaussian errors typically become immediately non-Gaussian, and biases (systematic errors) appear. Another practical difficulty arises because in weather forecasting or climate science temporal and spatial features may lead to state variables with millions of components, with error covariance matrices with 10^{12} or more entries.

20.7.1 ▪ Extended Kalman Filter

Nonlinear process and data models are often handled by linearization. The resulting algorithm is known as the *extended Kalman filter*. In essence, it uses the nonlinear process model (without noise terms) to compute the forecasting estimate and uses linearization about the most recent filtering estimate to compute the forecasting covariance and the gain matrix. The filtering estimate and its covariance are then computed from these quantities more or less in the same way as in Algorithm 20.3. This approach works well in engineering applications with a modest number of process variables but quickly becomes infeasible in high-dimensional situations. Variational methods for data assimilation that were mentioned earlier can avoid these difficulties but do not readily produce an assessment of errors.

20.7.2 ▪ Ensemble Kalman Filter

The *ensemble Kalman filter* (EnKf), introduced in the 1990s, uses stochastic simulation techniques known as *Monte Carlo* methods; [21] is a review article by one of the inventors of EnKf. The archetypical use for a Monte Carlo approach is the computation of an expected value $\mathscr{E}F(\mathbf{X})$ of a random variable $\mathbf{X}$ with density $f_X(\mathbf{x})$. Formally, the definition is $\mathscr{E}F(\mathbf{X}) = \int F(\mathbf{x}) f_X(\mathbf{x})\, d\mathbf{x}$, where the integral is over the space in which $\mathbf{X}$ takes its values. Numerical computation of the integral is essentially impossible if the dimension of $\mathbf{X}$ exceeds 10 or so. In a Monte Carlo approach, one draws m independent random samples $\mathbf{x}(1:m)$ from the distribution of $\mathbf{X}$ and approximates the expected value,

$$\mathscr{E}F(\mathbf{X}) \approx \frac{1}{m} \sum_{i=1}^{m} F(\mathbf{x}(i)). \tag{20.22}$$

By the law of large numbers, the right-hand side converges almost surely to the correct expected value as $m \to \infty$, and the speed of convergence can be determined (it is always slow).

For a simple version of EnKf, we assume a nonlinear process model,

$$\mathbf{X}(i) = \mathscr{M}_i(\mathbf{X}(i-1)) + \xi_i, \tag{20.23}$$

where the $\mathscr{M}_i$ are functions from the range of the $\mathbf{X}(i)$ to itself. All other assumptions are left unchanged—the model errors ξ_i are Gaussian, there is a linear data model (20.17),

and the background state is Gaussian, $\mathbf{X}(0) \sim N(\mu, \Sigma)$. Suppose we are given an estimate $\hat{\mathbf{x}}_{i-1|i-1}$ of the filtering mean at $i-1$, an estimate $\hat{\Sigma}_{i-1|i-1}$ of the filtering covariance matrix at this time step, and an estimate $N(\hat{\mathbf{x}}_{i-1|i-1}, \hat{\Sigma}_{i-1|i-1})$ of the distribution of $\mathbf{X}(i-1)|\mathbf{y}(1:i-1)$. For the simulation, draw m independent samples $\mathbf{x}^j_{i-1|i-1}$ $(j = 1, \ldots, m)$ from this distribution, propagate them forward with the process, and add simulated model errors $\eta_j \sim N(0, Q_i)$ that have the same distribution as the model errors ξ_i at this time step. The result is an *ensemble* of simulated forecasts,

$$\mathbf{x}^j_{i|i-1} = \mathcal{M}_i(\mathbf{x}^j_{i|i-1}) + \eta_j, \quad j = 1, \ldots, m. \tag{20.24}$$

The forecasting estimate is now computed as the sample mean of this ensemble,

$$\hat{\mathbf{x}}_{i|i-1} = \frac{1}{m} \sum_{j=1}^{m} \mathbf{x}^j_{i|i-1}, \tag{20.25}$$

and the sample covariance matrix provides an estimate $\hat{\Sigma}_{j|j-1}$ of the forecasting covariance matrix. Next, compute the gain matrix, just as in (20.20) but with the estimated covariance,

$$\hat{K}_j = \hat{\Sigma}_{j|j-1} H_j^T \left(H_j^T \hat{\Sigma}_{j|j-1} H_j + R_j \right)^{-1}. \tag{20.26}$$

To compute a filtering estimate, the ensemble of forecasts is adjusted using an innovation term and gain matrix as in Eq. (20.19). There is only one observation $\mathbf{y}(i)$ available, but it turns out that using it unchanged for all innovations in the ensemble tends to underestimate the variability of the filtering distribution. Hence, the innovation term $\mathbf{y}(i) - H_i \mu_{i|i-1}$ in the ordinary Kalman filter is replaced by an innovation ensemble $\mathbf{y}(i) + e_j - H_i \mathbf{x}^j_{i|i}$, where the perturbations e_j are simulated observation errors with the same distribution as the errors in the data model (20.17). One can therefore compute an ensemble of simulated filtering states,

$$\mathbf{x}^j_{i|i} = \mathbf{x}^j_{i|i-1} + \hat{K}_i \left(\mathbf{y}(i) + e_j - H_i \mathbf{x}^j_{i|i-1} \right). \tag{20.27}$$

This ensemble is used to produce estimates $\hat{\mathbf{x}}_{i|i}$ of the filtering mean and $\hat{\Sigma}_{i|i}$ of the filtering covariance matrix. One then uses $N(\hat{\mathbf{x}}_{i|i}, \hat{\Sigma}_{i|i})$ as an estimate of the filtering distribution at time step i, and the algorithm has completed a step. There is also a reanalysis version of this method.

If the process model is actually linear, such that Eq. (20.23) reduces to Eq. (20.16), then the ensemble Kalman forecasting and filtering estimates converge in the limit of large ensemble size to those of the ordinary Kalman filter algorithm. However, if the $\mathcal{M}_i$ are nonlinear, then the forecasting and filtering distributions become non-Gaussian, and there will be biases from the ensemble approach that do not disappear with large ensemble size. These biases must be assessed or possibly corrected separately.

If the space of process variables is high-dimensional (10^6 or 10^8 is not uncommon), the ensemble approach successfully avoids the problem of high-dimensional integration and the manipulation of huge covariance matrices. However, typically the ensemble size is much smaller, perhaps $m = \mathcal{O}(10^2)$. Then the sample covariance matrices have rank at most m and cannot possibly give all covariances correctly. On the other hand, in such situations the process vector $\mathbf{x}(i)$ may describe physical quantities at different locations across a region or around the globe. Then one often multiplies $\hat{\Sigma}_{i|i-1}$ or $\hat{\Sigma}_{i|i}$ elementwise with a "cut-off" matrix C whose entries are small far from the diagonal (for component pairs that have little to do with each other). This trick eliminates spurious large correlations at distant locations and at the same time tends to restore full rank to the estimated covariance matrices. Care must be taken to avoid destroying teleconnections.

The EnKf is now widely used. Many versions have been developed, and the literature has grown quite large; [49] is a recent overview by one of the leading experts.

20.8 ▪ Data Assimilation for the Lorenz System

To illustrate the EnKf technique, we apply the algorithm to the Lorenz model (7.1) with the fixed parameter values $\sigma = 10$, $\beta = \frac{8}{3}$, and $\rho = 28$. The attractor is shown in Figure 7.2.

Let $\mathbf{x} = (x, y, z) : t \mapsto \mathbf{x}(t) = (x(t), y(t), z(t))$ be the solution of the Lorenz system (7.1) which satisfies the initial data $\mathbf{x}(0) = \mathbf{x}_0 = (x_0, y_0, z_0) \in \mathbb{R}^3$. The process model associated with the Lorenz equations is the map $\mathcal{M} : \mathbb{R}^3 \to \mathbb{R}^3$ defined by

$$\mathcal{M}(\mathbf{x}_0) = \mathbf{x}(1) \in \mathbb{R}^3, \quad \mathbf{x}_0 \in \mathbb{R}^3. \tag{20.28}$$

There is no closed formula for $\mathcal{M}$, so $\mathcal{M}$ must be computed numerically by solving the system of differential equations for each $\mathbf{x}_0$.

Figure 20.4 shows the x-component of a trajectory for $0 \le t \le 6$ for initial data $x_0 = -7.3$, $y_0 = -11.5$, $z_0 = 17.8$. This trajectory switches from one "sheet" of the attractor to the other near $t = 2$. Also shown in Figure 20.4 are the x-components of ten trajectories from a solution ensemble with initial data $\mathbf{x}_0' = \mathbf{x}_0 + (\xi_1, \xi_2, \xi_3)$, where the ξ_i are $N(0,1)$ random variables, as well as the ensemble mean computed by averaging 1000 trajectories of this ensemble. For $t > 1.5$ or so, the ensemble mean is seen to differ dramatically from the true solution. The specific reason in this case is that trajectories from the ensemble switch from one leaf of the attractor to the other at times that can be very different from $t = 2$. Essentially, if the initial data are known only up to random errors that have standard normal distributions, then the true trajectory becomes unpredictable for $t > 1.5$ or so. Averaging over many ensemble trajectories does not eliminate this bias. This situation is typical for nonlinear systems of differential equations; for linear systems, the ensemble mean always equals the true trajectory.

Figure 20.5 again shows the x-component of the true trajectory, together with the corresponding results of the EnKf, for a ensemble of size 50. The data model is given by Eq. (20.17), with $H = \left(\begin{smallmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{smallmatrix}\right)$. Therefore, only a two-dimensional projection of the true trajectory, corrupted by random noise, is observed at each time. The error terms ζ_i are standard normal $N(0,1)$. Each assimilated trajectory starts at a filtering state at $t = i$ (red circle) and ends at a forecasting state at $t = i + 1$ (green circle). Evidently, the assimilation processes is only partially successful in approximating the correct trajectory. For example, on the intervals $1.5 < t < 2$ and $4.5 < t < 5$, the assimilated trajectories are not even qualitatively correct. Nevertheless, it is remarkable that a significant portion of the true trajectory is recreated correctly over the entire interval of interest.

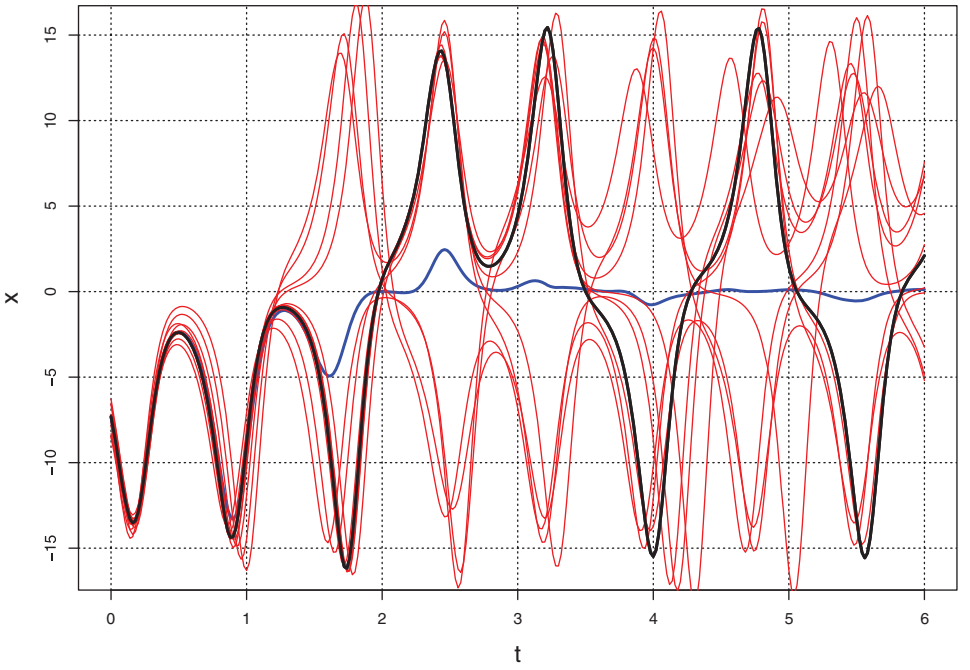

Figure 20.4. *Lorenz system: x-component without data assimilation; exact trajectory (black), some trajectories from the solution ensemble (red), and ensemble mean (blue).*

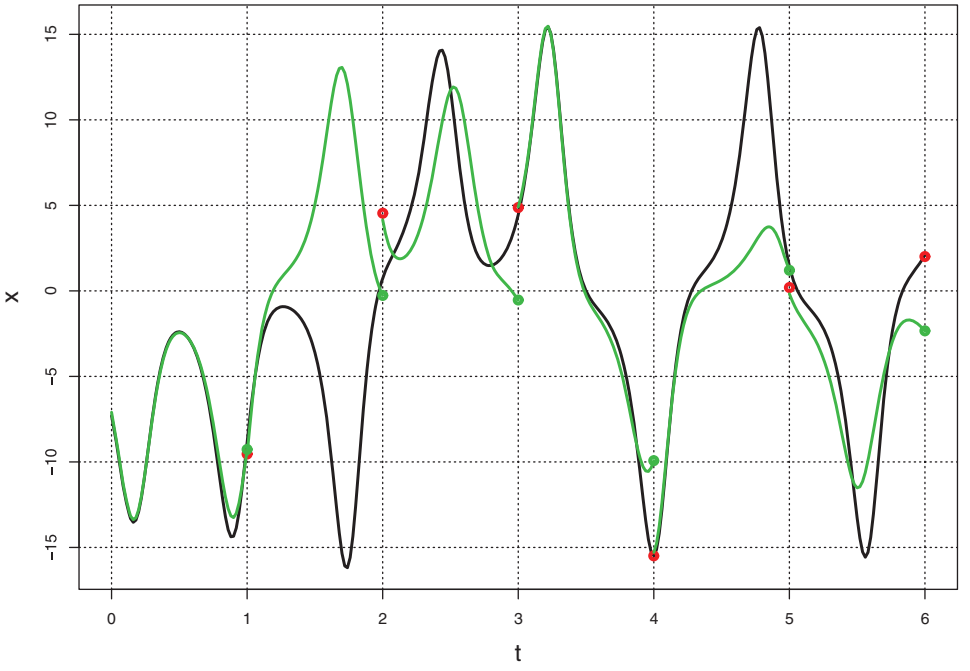

Figure 20.5. *Lorenz system: x-component with data assimilation; exact trajectory (black) and assimilated trajectories (green), starting with filtering states (red circles) and ending in forecasting states (green circles).*

20.9 ▪ Concluding Remarks

Data assimilation has been very important for weather prediction, where it has extended both the range and reliability of forecasts. In the best-case scenario, there is a correct dynamical model with sufficiently high resolution and a well-designed observational network that can provide accurate data. In the next-best-case scenario, there may be a somewhat deficient dynamical model, but the observational network is still good enough to provide adequate data. Then it is still possible to obtain estimates in agreement with nature, as long as dynamical model errors are recognized and incorporated adequately.

In climate science, data assimilation has been used since the late 1980s. It is having an increasingly significant impact, since it allows improved and faster estimation of both internal and external parameters. Parameter estimation is particularly promising in the biogeochemistry of the climate system, where it is often difficult to measure rates directly *in situ*. Realistic dynamical models and feasible measurements of quantities such as concentration fields of plankton, in combination with advanced data assimilation techniques, can lead to surprisingly realistic estimates of these rates.

Data assimilation techniques have been applied to create reliable uniform background data for the past (reanalysis) and to obtain information about unobservable quantities such as ocean upwelling and chemical reaction rates in the ocean. A typical example is described in [11], where an EnKf approach is used to reconstruct the ocean climate for the period 1958–2001.

Interestingly, there are specific problems with heterogeneous data sources associated with all reanalysis exercises for ocean circulation for the 20th century. Since about 1980, satellites with circumpolar orbits have provided reliable records of planet-wide uniformly distributed ocean measurements. But prior to about 1980, observations came mainly from shipboard observations that were concentrated along major shipping routes. There are also other more subtle trends in the data collection efforts that are known to introduce spurious trends during data assimilation. For example, the typical height at which shipboard anemometers are mounted has increased since the middle of the last century, leading to biased results for wind strengths and patterns. In the past, when data assimilation was used mainly for weather prediction, such slow trends did not matter.

Ridgwell et al. [88] describe a similar project for biogeochemistry data. Here, an EnKf is employed in an iterative fashion to obtain steady-state information about geochemical parameters such as uptake rates and concentrations of phosphate and calcium carbonate in the ocean at preindustrial times.

The EnKf is known to run into problems when the underlying process model is highly nonlinear and has many unstable equilibria. Other methods have been proposed for such situations. For example, Apte, Jones, and Stuart [2] develop a Bayesian approach to address data assimilation questions coming from drifting buoys, which are known to have trajectories that are very sensitive to changes in the initial data.

20.10 ▪ Exercises

1. Consider a physical process in which real-valued state variables x_i, $i = 1, 2, \ldots$, are generated according to the rule $x_{i+1} = f(x_i)$, where f is a given real-valued function. The states are observed according to the data model $y_i = x_i + \xi_i$, where the ξ_i are independent random variables with an $N(0, \sigma^2)$-distribution.

   (i) Assume that observations $y_1, y_2, \ldots, y_N$ are available and that you want to estimate x_1. Describe in general terms how a cost function should be set up whose minimum is expected to give an estimate for x_1.

(ii) Consider the special case $f(x) = 4x(1-x)$. Begin with the case $N = 3$, $\sigma = 0.1$, $x_1 = 0.2$. Construct this cost function for several realizations of the y_i and plot it, using a computer. Is the minimum of the cost function where you expect it to be?

(iii) Repeat part (ii) with a larger σ, for example $\sigma = 0.5$. Repeat with the previous σ and a larger N, for example $N = 6$. Describe your observations. Does it help to have more observations available?

The cost function may have multiple local minima, which can lead to serious numerical difficulties.

2. Consider the process model (20.1) together with the data model (20.2). Derive the joint distribution of Y_2, Y_3 as a function of the unknown parameter x_2 and use it to obtain the maximum likelihood estimate $\hat{x}_2$. Interpret your result in the case where $\tau \ll 1$.

3. Consider again the process model (20.1) together with the data model (20.2). Derive the joint distribution of Y_2, Y_3 as a function of the unknown parameter x_3 and use it to obtain the maximum likelihood estimate $\hat{x}_3$. Interpret your result in the case where $\tau \ll 1$. Use $x_2 = \alpha^{-1}(x_3 - \xi_2)$.

4. Consider the process model (20.1) and assume that $X_1 \sim N(\mu_0, \sigma^2)$.

   (i) Compute the covariance matrix Σ of the random vector $\mathbf{X} = (X_1, \ldots, X_4)^T$.

   (ii) Compute the gain matrix K from the definition (20.10).

   (iii) Show that

$$\operatorname{var}(X_1|\mathbf{y}) = \frac{\sigma^2\left(\tau^4 + \left(\alpha^2+2\right)\tau^2 + 1\right)}{\tau^4 + \left(\left(\alpha^2+1\right)\sigma^2\alpha^2 + \alpha^2 + 2\right)\tau^2 + \alpha^2\sigma^2 + 1}$$

   and

$$\operatorname{var}(X_3|\mathbf{y}) = \frac{\tau^2\left(\alpha^2\sigma^2 + \left(\sigma^2\alpha^4 + \alpha^2 + 1\right)\tau^2 + 1\right)}{\tau^4 + \left(\left(\alpha^2+1\right)\sigma^2\alpha^2 + \alpha^2 + 2\right)\tau^2 + \alpha^2\sigma^2 + 1}.$$

5. Consider the general process model $\mathbf{X}(i) = M_i\mathbf{X}(i-1) + \mathbf{b}(i) + \xi_i$, $i = 1, \ldots, N$, where $\mathbf{X}(i), \mathbf{b}(i), \xi_i \in \mathbb{R}^n$ and M_i are $n \times n$ matrices. Define $\overline{\mathbf{X}}(i)$ by

$$\overline{\mathbf{X}}(0) = 0; \quad \overline{\mathbf{X}}(i) = M_i\overline{\mathbf{X}}(i-1) + \mathbf{b}(i), \quad i = 1, \ldots, N, \tag{20.29}$$

   and set $\tilde{\mathbf{X}}(i) = \mathbf{X}(i) - \overline{\mathbf{X}}(i)$. Show that the $\tilde{\mathbf{X}}(i)$ satisfy the recursion (20.16).

6. Use the MATLAB code for the Lorenz equations given in Section C.1 to explore the behavior of the ensemble mean of the Lorenz equations for other initial data $\mathbf{x}_0$. Can you find initial data such that the ensemble mean stays close to the correct solution for an interval of length $T = 5$?

7. Consider the linear process model (20.16) together with the data model (20.17), and assume that all matrices are independent of i, so $M_i = M$, $H_i = H$, $Q_i = Q$, and $R_i = R$ for all i. Assume that the forecasting covariates matrices $\Sigma_{i|i-1}$ converge to a limit Σ_0. Derive the equation

$$\Sigma_0 = Q + M\left(\Sigma_0 - \Sigma_0 H^T(H\Sigma_0 H^T + R)^{-1}H\Sigma_0\right)M^T,$$

and derive similar equations for the limits of the matrices $\Sigma_{i|i}$ and K_i. The above equation is known as a matrix RICCATI equation. It reduces to an ordinary quadratic equation if the dimension of the state space is 1.