Chapter 3

Statistical estimation and sequential data assimilation

The road to wisdom?—Well, it's plain and simple to express:
Err
and err
and err again
but less
and less
and less.
—Piet Hein (1905–1996, Danish mathematician and inventor)

3.1 ▪ Introduction

In this chapter, we present the statistical approach to DA. This approach will be addressed from a Bayesian point of view. But before delving into the mathematical and algorithmic details, we will discuss some ideas about the history of weather forecasting and of the distinction between prediction and forecasting. For a broad, nontechnical treatment of prediction in a sociopolitical-economic context, the curious reader is referred to Silver [2012], where numerous empirical aspects of forecasting are also broached.

3.1.1 ▪ A long history of prediction

From Babylonian times, people have attempted to predict future events, for example in astronomy. Throughout the Renaissance and the Industrial Revolution there were vast debates on predictability.

In 1814, Pierre-Simon Laplace postulated that a perfect knowledge of the actual state of a system coupled with the equations that describe its evolution (natural laws) should provide perfect predictions! This touches on the far-reaching controversy between determinism and randomness/uncertainty … and if we go all the way down to the level of quantum mechanics, then due to Heisenberg's principle there cannot be a perfect prediction. However, weather (and many other physical phenomena) go no further than the molecular (not the atomic) level and as a result molecular chemistry and Newtonian physics are sufficient for weather forecasting. In fact, the (deterministic) PDEs that describe the large-scale circulation of air masses and oceans are

remarkably precise and can reproduce an impressive range of meteorological conditions. This is equally true in a large number of other application domains, as described in Part III.

Weather forecasting is a success story: human and machine combining their efforts to understand and to anticipate a complex natural system. This is true for many other systems thanks to the broad applicability of DA and inverse problem methods and algorithms.

3.1.2 ▪ Stochastic versus deterministic

The simplest statistical approach to forecasting (rather like linear regression, but with a flat line) is to calculate the probability of an event (e.g., rain tomorrow) based on past knowledge and records—i.e., long-term averages. But these purely statistical predictions are of little value—they do not take into account the possibility and potential that we have of modeling the physics—this is where the progress (over the last 30 years) in numerical analysis and high-performance computing can come to the rescue. However, this is not a trivial pursuit, as we often notice when surprised by a rain shower, flood, stock market crash, or earthquake. So what goes wrong and impairs the accuracy/reliability of forecasts?

- The first thing that can go wrong is the resolution (spatial and temporal) of our numerical models … but this is an easier problem: just add more computing power, energy, and money!
- Second, and more important, is *chaos* (see Section 2.5.1), which applies to dynamic, nonlinear systems and is closely associated with the well-posedness issues of Chapter 1—note that this has nothing to do with randomness, but rather is related to the lack of predictability. In fact, in weather modeling, for example, after approximately one week only, chaos theory swamps the dynamic memory of the atmosphere (as "predicted" by the physics), and we are better off relying on climatological forecasts that are based on historical averaged data.
- Finally, there is our imprecise knowledge of the initial (and boundary) conditions for our physical model and hence our simulations—this loops back to the previous point and feeds the chaotic nature of the system. Our measurements are both incomplete and (slightly) inaccurate due to the physical limitations of the instruments themselves.

All of the above needs to be accounted for, as well as possible, in our numerical models and computational analysis. This can best be done with a probabilistic[26] approach.

3.1.3 ▪ Prediction versus forecast

The terms *prediction* and *forecast* are used interchangeably in most disciplines but deserve a more rigorous definition/distinction. Following the philosophy of Silver [2012], a prediction will be considered as a deterministic statement, whereas a forecast will be a probabilistic one. Here are two examples:

- "A major earthquake will strike Tokyo on May 28th" is a prediction, whereas "there is a 60% chance of a major earthquake striking Northern California over the next 25 years" is a forecast.

[26] Equivalently, a statistical or stochastic approach can be used.

- Extrapolation is another example of prediction and is in fact a very basic method that can be useful in some specific contexts but is generally too simplistic and can lead to very bad predictions and decisions.

We notice the UQ in the forecast statement. One way to implement UQ is through Bayesian reasoning—let us explain this now.

3.1.4 ▪ DA is fundamentally Bayesian

Thomas Bayes[27] believed in a rational world of Newtonian mechanics but insisted that by gathering evidence we can get closer and closer to the truth. In other words, rationality is probabilistic. Laplace, as we saw above, claimed that with perfect knowledge of the present and of the laws governing its evolution, we can attain perfect knowledge of the future. In fact it was Laplace who formulated what is known as Bayes' theorem. He considered probability to be "a waypoint between ignorance and knowledge." This is not bad ... it corresponds exactly to our endeavor and what we are trying to accomplish throughout this book: use models and simulations to reproduce and then predict (or, more precisely, forecast or better understand) the actual state and future evolutions of a complex system. For Laplace it was clear: we need a more thorough understanding of probability to make scientific progress!

Bayes' theorem is a very simple algebraic formula based on conditional probability (the probability of one event, A, occurring, knowing or given that another event, B, has occurred—see Section 3.2 below for the mathematical definitions):

$$p_{A|B} = \frac{p_{B|A}\, p_A}{p_B}.$$

It basically provides us with a reevaluated probability (posterior, $p_{A|B}$) based on the prior knowledge, $p_{B|A}\, p_A$, of the system that is normalized by the total knowledge that we have, p_B. To better understand and appreciate this result, let us consider a couple of simple examples that illustrate the importance of Bayesian reasoning.

Example 3.1. The famous example of breast cancer diagnosis from mammograms shows the importance and strength of priors. Based on epidemiological studies, the probability that a woman between the ages of 40 and 50 will be afflicted by a cancer of the breast is low, of the order of $p_A = 0.014$ or 1.4%. The question we want to answer is: If a woman in this age range has a positive mammogram (event B), what is the probability that she indeed has a cancer (event A)? Further studies have shown that the false-positive rate of mammograms is $p = 0.1$ or 10% of the time and that the correct diagnosis (true positive) has a rate of $p_{B|A} = 0.75$. So a positive mammogram, taken by itself, would seem to be serious news. However, if we do a Bayesian analysis that factors in the prior information, we get a different picture. Let us do this now. The posterior probability can be computed from Bayes' formula,

$$p_{A|B} = \frac{p_{B|A}\, p_A}{p_B} = \frac{0.75 \times 0.014}{0.75 \times 0.014 + 0.1 \times (1 - 0.014)} = 0.096,$$

and we conclude that the probability is only 10% in this case, which is far less worrisome than the overall 75% true-positive rate. So the false positives have dominated the result thanks to the fact that we have taken into account the prior information of

[27] English clergyman and statistician (1701–1761).

low cancer incidence in this age range. For this reason, there is a tendency in the medical profession today to recommend that women (without antecedents, which would increase the value of p_A) start having regular mammograms starting from age 50 only because, starting from this age, the prior probability is higher. ∎

Example 3.2. Another good example comes from global warming, now called climate change, and we will see why it is so important to quantify uncertainty in the interest of scientific advancement and trust. The study of global warming started around the year 2001. At this time, it was commonly accepted, and scientifically justified, that CO_2 emissions caused and would continue to cause a rise in global temperatures. Thus, we could attribute a high prior probability, $p_A = 0.95$, to the hypothesis of global warming (event A). However, over the subsequent decade from 2001 to 2011, we have observed (event B) that global temperatures have *not* risen as expected—in fact they appeared to have decreased very slightly.[28] So, according to Bayesian reasoning, we should reconsider our estimation of the probability of global warming—the question is, to what extent? If we had a good estimate of the uncertainty in short-term patterns of temperature variations, then the downward revision of the prediction would not be drastic. By analyzing the historical data again, we find that there is a 15% chance that there is no net warming over a decade even if the global warming hypothesis holds—this is due to the inherent variability in the climate. On the other hand, if temperature variations were purely random, and hence unpredictable, then the chance of having a decade in which there is actually a cooling would be 50%. So let us compute the revised estimate for global warming with Bayes' formula. We find

$$p_{A|B} = \frac{p_{B|A}p_A}{p_B} = \frac{0.15 \times 0.95}{0.15 \times 0.95 + 0.5 \times (1-0.95)} = 0.851,$$

so we should revise our probability, in light of the last decade's evidence, from 95% to 85%. This is a truly honest approximation that takes into account the observations and revises the uncertainty. Of course, when we receive a new batch of measurements, we can recompute and obtain an update. This is precisely what DA seeks to achieve. The major difference resides in our possession (in the DA context) of a sophisticated model for actually computing the conditional probability, $p_{B|A}$, the probability of the data, or observations, given the parameters. ∎

3.1.5 ▪ First steps toward a formal framework

Now let us begin to formalize. It can be claimed that a major part of scientific discovery and research deals with questions of this nature: what can be said about the value of an unknown, or inaccurately known, variable θ that represents the parameters of the system, if we have some measured data $\mathcal{D}$ and a model $\mathcal{M}$ of the underlying mechanism that generated the data? But this is precisely the Bayesian context,[29] where we seek a quantification of the uncertainty in our knowledge of the parameters that, according

[28] A recent paper, published in *Science*, has rectified this by taking into account the evolution of instrumentation since the start of the study. Indeed, it now appears that there has been a steady increase! Apparently, the "hiatus" was the result of a double observational artefact [see T.R. Karl et al., *Science Express*, 4 June 2015].

[29] See Barber [2012], where Bayesian reasoning is extensively developed in the context of machine learning.

to Bayes' theorem takes the form

$$p(\theta \mid \mathscr{D}) = \frac{p(\mathscr{D} \mid \theta)p(\theta)}{p(\mathscr{D})} = \frac{p(\mathscr{D} \mid \theta)p(\theta)}{\int_\theta p(\mathscr{D} \mid \theta)p(\theta)}.$$

Here, the physical model is represented by the conditional probability (also known as the *likelihood*) $p(\mathscr{D} \mid \theta)$, and the prior knowledge of the system by the term $p(\theta)$. The denominator is considered as a normalizing factor and represents the total probability of $\mathscr{D}$. From these we can then calculate the resulting posterior probability, $p(\theta \mid \mathscr{D})$.

The most probable estimator, called the maximum a posteriori (MAP) estimator, is the value that maximizes the posterior probability

$$\theta_* = \arg\max_\theta p(\theta \mid \mathscr{D}).$$

Note that for a flat, or uninformative, prior $p(\theta)$, the MAP is just the maximum likelihood, which is the value of θ that maximizes the likelihood $p(\mathscr{D} \mid \theta)$ of the model that generated the data, since in this case neither $p(\theta)$ nor the denominator plays a role in the optimization.

3.1.6 ▪ Concluding remarks (as an opening ...)

There are links between the above and the theories of state space, optimal control, and optimal filtering. We will study KFs, whose original theory was developed in this state space context, below—see Friedland [1986] and Kalman [1960].

The following was the theme of a recent Royal Meteorological Society meeting (Imperial College, London, April 2013): "Should weather and climate prediction models be deterministic or stochastic?"—this is a very important question that is relevant for other physical systems.

In this chapter, we will argue that uncertainty is an inherent characteristic of (weather and most other) predictions and thus that no forecast can claim to be complete without an accompanying estimation of its uncertainty—what we call uncertainty quantification (UQ).

3.2 ▪ Statistical estimation theory

In statistical modeling, the concepts of sample space, probability, and random variable play key roles. Readers who are already familiar with these concepts can skip this section. Those who require more background on probability and statistics should definitely consult a comprehensive treatment, such as DeGroot and Schervisch [2012] or the excellent texts of Feller [1968], Jaynes [2003], McPherson [2001], and Ross [1997].

A sample space, $\mathscr{S}$, is the set of all possible outcomes of a random, unpredictable experiment. Each outcome is a point (or an element) in the sample space. Probability provides a means for quantifying how likely it is for an outcome to take place. Random variables assign numerical values to outcomes in the sample space. Once this has been done, we can systematically work with notions such as average value, or mean, and variability.

It is customary in mathematical statistics to use capital letters to denote random variables (r.v.'s) and corresponding lowercase letters to denote values taken by the r.v. in its range. If $X : \mathscr{S} \to \mathbb{R}$ is an r.v., then for any $x \in \mathbb{R}$, by $\{X \le x\}$ we mean $\{s \in \mathscr{S} \mid X(s) \le x\}$.

Definition 3.3. *A* probability space $(\mathcal{S}, \mathcal{B}, \mathcal{P})$ *consists of a set* $\mathcal{S}$ *called the* sample space, *a collection* $\mathcal{B}$ *of (Borel) subsets of* $\mathcal{S}$, *and a* probability function $\mathcal{P} : \mathcal{B} \to \mathbb{R}_+$ *for which*

- $\mathcal{P}(\emptyset) = 0$,
- $\mathcal{P}(\mathcal{S}) = 1$, and
- $\mathcal{P}\left(\bigcup_i S_i\right) = \sum_i \mathcal{P}(S_i)$ for any disjoint, countable collection of sets $S_i \in \mathcal{B}$.

A *random variable* X is a measurable function $X : \mathcal{S} \to \mathbb{R}$. Associated with the r.v. X is its *distribution function*,

$$F_X(x) = \mathcal{P}\{X \le x\}, \quad x \in \mathbb{R}.$$

The distribution function is nondecreasing and right continuous and satisfies

$$\lim_{x \to -\infty} F_X(x) = 0, \quad \lim_{x \to +\infty} F_X(x) = 1.$$

Definition 3.4. *A random variable* X *is called* discrete *if there exist countable sets* $\{x_i\} \subset \mathbb{R}$ *and* $\{p_i\} \subset \mathbb{R}_+$ *for which*

$$p_i = \mathcal{P}\{X = x_i\} > 0$$

for each i, *and*

$$\sum_i p_i = 1.$$

In this case, the PDF *for* X *is the real-valued function with discrete support*

$$p_X(x) = \begin{cases} p_i & \text{if } x = x_i, \quad i = 1, 2, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

The x_i*'s are the points of discontinuity of the distribution function,*

$$F_X(x) = \sum_{\{i | x_i \le x\}} p_X(x_i).$$

Definition 3.5. *A random variable* X *is called* continuous *if its distribution function,* F_X, *is absolutely continuous. In this case,*

$$F_X(x) = \int_{-\infty}^{x} p_X(u)\, \mathrm{d}u,$$

and there exists a derivative of F_X,

$$p_X(x) = \frac{\mathrm{d}F_X}{\mathrm{d}x},$$

that is called the probability density function (PDF) *for* X.

Definition 3.6. *The* mean, *or* expected value, *of an r.v.* X *is given by the integral*

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x\, \mathrm{d}F_X(x).$$

This is also known as the first moment of the random variable. If X is a continuous r.v., then

$$\mathrm{d}F_X(x) = p_X(x)\,\mathrm{d}x,$$

and, in the discrete case,

$$\mathrm{d}F_X(x) = p_X(x_i)\delta(x - x_i).$$

In the latter case,

$$\mathrm{E}(X) = \sum_i x_i\, p_X(x_i).$$

The expectation operator, E, *is a linear operator.*

Definition 3.7. *The* variance *of an r.v. X is given by*

$$\sigma^2 = \mathrm{E}\left[(X-\mu)^2\right] = \mathrm{E}(X^2) - (\mathrm{E}(X))^2,$$

where

$$\mu = \mathrm{E}(X).$$

Definition 3.8. *The* mode *is the value of x for which the PDF $p_X(x)$ attains its maximal value.*

Definition 3.9. *Two r.v.'s, X and Y, are* jointly distributed *if they are both defined on the same probability space, $(\mathcal{S}, \mathcal{B}, \mathcal{P})$.*

Definition 3.10. *A* random vector, $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, *is a mapping from $\mathcal{S}$ into $\mathbb{R}^n$ for which all the components X_i are jointly distributed. The* joint distribution function *of $\mathbf{X}$ is given by*

$$F_{\mathbf{X}}(\mathbf{x}) = \mathcal{P}\{X_1 \le x_1, \ldots, X_n \le x_n\}, \quad \mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

The components X_i are independent *if the joint distribution function is the product of the distribution functions of the components,*

$$F_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} F_{X_i}(x_i).$$

Definition 3.11. *A random vector $\mathbf{X}$ is* continuous *with* joint PDF $p_{\mathbf{X}}$ *if*

$$F_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} p_{\mathbf{X}}(\mathbf{u})\,\mathrm{d}u_1 \ldots \mathrm{d}u_n.$$

Definition 3.12. *The mean, or* expected value, *of a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is the n-vector $\mathrm{E}(\mathbf{X})$ with components*

$$[\mathrm{E}(\mathbf{X})]_i = \mathrm{E}(X_i), \quad i = 1, \ldots, n.$$

The covariance *of $\mathbf{X}$ is the $n \times n$ matrix $\mathrm{cov}(\mathbf{X})$ with components*

$$[\mathrm{cov}(\mathbf{X})]_{ij} = \mathrm{E}\left[(X_i - \mu_i)(X_j - \mu_j)\right] = \sigma_{ij}^2, \quad 1 \le i, j \le n,$$

where

$$\mu_i = \mathrm{E}(X_i).$$

3.2.1 ▪ Gaussian distributions

A continuous random vector $\mathbf{X}$ has a Gaussian distribution if its joint PDF has the form

$$p_{\mathbf{X}}(\mathbf{x};\mu,\Sigma)=\frac{1}{\sqrt{(2\pi)^n\det(\Sigma)}}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right),$$

where $\mathbf{x},\mu\in\mathbb{R}^n$ and Σ is an $n\times n$ symmetric positive definite matrix. The mean is given by

$$\mathrm{E}(\mathbf{X})=\mu,$$

and the covariance matrix is

$$\mathrm{cov}(\mathbf{X})=\Sigma.$$

These two parameters completely characterize the distribution, and we indicate this situation by

$$\mathbf{X}\sim\mathcal{N}(\mu,\Sigma).$$

Note that in the *scalar* case we have the familiar *bell curve*,

$$p(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/(2\sigma^2)}.$$

3.2.2 ▪ Estimators and their properties

We now present some fundamental concepts of statistical estimation. A far more complete treatment can be found in Garthwaite et al. [2002], for example. Let us begin by defining estimation and estimators.

We suppose that we are in possession of a random sample $(x_1,x_2,\ldots,x_n)$ (of measurements, say) of the corresponding r.v.'s $X_1,X_2,\ldots,X_n$, whose PDF is $p_{\mathbf{X}}(\mathbf{x};\theta)$. We want to use the observed values $x_1,x_2,\ldots,x_n$ to estimate the parameter θ, which is either unknown or imprecisely known. We then calculate (see methods below) an estimate $\hat{\theta}$ of θ as a function of $(x_1,x_2,\ldots,x_n)$. The corresponding function $\hat{\theta}(X_1,X_2,\ldots,X_n)$, which is an r.v. itself, is an *estimator* for θ. In a given situation, there can exist a number of possible estimators (see example below), and thus the questions of how to choose the best one and what we mean by "best," have to be answered.

The first criterion, considered as indispensable in most circumstances, is that of *unbiasedness*.

Definition 3.13. *The estimator $\hat{\theta}$ for θ is an* unbiased *estimator if the expected value*

$$\mathrm{E}(\hat{\theta})=\theta.$$

The bias of $\hat{\theta}$ is the quantity $\mathrm{E}(\hat{\theta})-\theta$.

The notion of unbiasedness implies that the distribution of $\hat{\theta}$ (recall that $\hat{\theta}$ is an r.v.) is centered exactly at the value θ and that thus there is no tendency to either under- or overestimate this parameter.

Example 3.14. To estimate the mean of a (scalar-valued) normal distribution $\mathcal{N}(\mu,\sigma)$ from a sample of n values, the most evident estimator is the *sample mean*,

$$\hat{X}=\frac{1}{n}\sum_{i=1}^{n}X_i,$$

which is unbiased (this is also the case for other distributions, not only for Gaussians). However, there are numerous other unbiased estimators, for example, the median, the mid-range, and even X_1. ∎

We conclude that unbiasedness is usually not enough for choosing an estimator and that we need some other criteria to settle this issue. The classical ones (in addition to unbiasedness) are known as consistency, efficiency, and sufficiency. We will not go into the details here but will concentrate on some optimality conditions.

3.2.3 ▪ Maximum likelihood estimation

Suppose a random vector $\mathbf{X}$ has a joint PDF $p_{\mathbf{X}}(\mathbf{x};\theta)$, where θ is an unknown parameter vector that we would like to estimate. Suppose also that we have a data vector $\mathbf{d}=(d_1,\dots,d_n)$, a given realization of $\mathbf{X}$ (an outcome of a random experiment).

Definition 3.15. *A maximum likelihood estimator (MLE) for θ given $\mathbf{d}$ is a parameter vector $\hat{\theta}$ that maximizes the likelihood function*

$$L(\theta)=p_{\mathbf{X}}(\mathbf{d};\theta),$$

which is the joint PDF, considered as a function of θ. The MLE is also a maximizer of the log-likelihood function,

$$l(\theta)=\log p_{\mathbf{X}}(\mathbf{d};\theta).$$

3.2.4 ▪ Bayesian estimation

Now we can formalize the notions of Bayesian probability that were introduced in Sections 3.1.4 and 3.1.5. To this end, we must begin by discussing and defining conditional probability and conditional expectation.

Let $\mathbf{X}=(X_1,X_2,\dots,X_n)$ and $\mathbf{Y}=(Y_1,Y_2,\dots,Y_n)$ be jointly distributed discrete random vectors. Then $(\mathbf{X},\mathbf{Y})$ is also a discrete random vector.

Definition 3.16. *The joint PDF for $(\mathbf{X},\mathbf{Y})$ is given by*

$$p_{(\mathbf{X},\mathbf{Y})}(\mathbf{x},\mathbf{y})=\mathscr{P}\{\mathbf{X}=\mathbf{x},\quad \mathbf{Y}=\mathbf{y}\},\quad (\mathbf{x},\mathbf{y})\in\mathbb{R}^n\times\mathbb{R}^n.$$

The marginal PDF *of $\mathbf{X}$ is then defined as*

$$p_{\mathbf{X}}(\mathbf{x})=\sum_{\mathscr{P}\{\mathbf{Y}=\mathbf{y}\}>0} p_{(\mathbf{X},\mathbf{Y})}(\mathbf{x},\mathbf{y}),\quad \mathbf{x}\in\mathbb{R}^n. \tag{3.1}$$

The conditional PDF *for $\mathbf{Y}$ given $\mathbf{X}=\mathbf{x}$ is then defined as*

$$p_{(\mathbf{Y}|\mathbf{X})}(\mathbf{y}\mid\mathbf{x})=\frac{p_{(\mathbf{X},\mathbf{Y})}(\mathbf{x},\mathbf{y})}{p_{\mathbf{X}}(\mathbf{x})}, \tag{3.2}$$

where the denominator is nonzero.

So, the conditional probability $p(A|B)$ is the revised probability of an event A after learning that the event B has occurred.

Remark 3.17. *If* $\mathbf{X}$ *and* $\mathbf{Y}$ *are independent random vectors, then the conditional density function of* $\mathbf{Y}$ *given* $\mathbf{X} = \mathbf{x}$ *does not depend on* $\mathbf{x}$ *and satisfies*

$$p_{(\mathbf{Y}|\mathbf{X})}(\mathbf{y} \mid \mathbf{x}) = \frac{p_{\mathbf{X}}(\mathbf{x}) p_{\mathbf{Y}}(\mathbf{y})}{p_{\mathbf{X}}(\mathbf{x})} = p_{\mathbf{Y}}(\mathbf{y}). \tag{3.3}$$

Definition 3.18. *Let* $\phi : \mathbb{R}^n \to \mathbb{R}^k$ *be a measurable mapping. The* conditional expectation *of* $\phi(\mathbf{Y})$ *given* $\mathbf{X} = \mathbf{x}$ *is*

$$\mathrm{E}(\phi(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}) = \sum_{\mathscr{P}\{\mathbf{Y}=\mathbf{y}\}>0} \phi(\mathbf{y}) p_{(\mathbf{Y}|\mathbf{X})}(\mathbf{y} \mid \mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^n. \tag{3.4}$$

Remark 3.19. *For continuous random vectors* $\mathbf{X}$ *and* $\mathbf{Y}$*, we can define the analogous concepts by replacing the summations in* (3.1)–(3.4) *with appropriate integrals:*

$$p_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{\infty} p_{(\mathbf{X},\mathbf{Y})}(\mathbf{x}, \mathbf{y}) \, \mathrm{d}F_{\mathbf{Y}}(\mathbf{y}),$$

$$\mathrm{E}(\phi(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}) = \int_{-\infty}^{\infty} \phi(\mathbf{y}) p_{(\mathbf{Y}|\mathbf{X})}(\mathbf{y} \mid \mathbf{x}) \, \mathrm{d}F_{\mathbf{Y}}(\mathbf{y}).$$

We are now ready to state Bayes' law, which relates the conditional random vector $\mathbf{X}|_{\mathbf{Y}=\mathbf{y}}$ to the inverse conditional random vector $\mathbf{Y}|_{\mathbf{X}=\mathbf{x}}$.

Theorem 3.20 (Bayes' law). *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be jointly distributed random vectors. Then*

$$p_{(\mathbf{X}|\mathbf{Y})}(\mathbf{x} \mid \mathbf{y}) = \frac{p_{(\mathbf{Y}|\mathbf{X})}(\mathbf{y} \mid \mathbf{x}) p_{\mathbf{X}}(\mathbf{x})}{p_{\mathbf{Y}}(\mathbf{y})}. \tag{3.5}$$

Proof. By the definition of conditional probability (3.2),

$$p_{(\mathbf{X}|\mathbf{Y})}(\mathbf{x} \mid \mathbf{y}) = \frac{p_{(\mathbf{X},\mathbf{Y})}(\mathbf{x}, \mathbf{y})}{p_{\mathbf{Y}}(\mathbf{y})},$$

and the numerator is exactly equal to that of (3.5), once again by definition. ∎

Definition 3.21. *In the context of Bayes' law* (3.5)*, suppose that* $\mathbf{X}$ *represents the variable of interest and that* $\mathbf{Y}$ *represents an observable (measured) quantity that depends on* $\mathbf{X}$*. Then,*

- $p_{\mathbf{X}}(\mathbf{x})$ *is called the* a priori *PDF, or the* prior*;*
- $p_{(\mathbf{X}|\mathbf{Y})}(\mathbf{x} \mid \mathbf{y})$ *is called the* a posteriori *PDF, or the* posterior*;*
- $p_{(\mathbf{Y}|\mathbf{X})}(\mathbf{y} \mid \mathbf{x})$*, considered as a function of* $\mathbf{x}$*, is the* likelihood *function;*
- *the denominator, called the* evidence*,* $p_{\mathbf{Y}}(\mathbf{y})$*, can be considered as a normalization factor; and*
- *the posterior distribution is thus proportional to the product of the likelihood and the prior distribution or, in applied terms,*

$$p(\text{parameter} \mid \text{data}) \propto p(\text{data} \mid \text{parameter}) p(\text{parameter}).$$

Remark 3.22. *A few fundamental remarks are in order here. First, Bayes' law plays a central role in probabilistic reasoning since it provides us with a method for inverting probabilities, going from* $p(\mathbf{y} \mid \mathbf{x})$ *to* $p(\mathbf{x} \mid \mathbf{y})$*. Second, conditional probability matches perfectly our intuitive notion of uncertainty. Finally, the laws of probability combined with Bayes' law constitute a complete reasoning system for which traditional deductive reasoning is a special case [Jaynes,* 2003*].*

3.2.5 ▪ Linear least-squares estimation: BLUE, minimum variance linear estimation

In this section we define the two estimators that form the basis of statistical DA. We show that these are optimal, which explains their widespread use.

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_m)$ be two jointly distributed, real-valued random vectors with finite expected squared components:

$$\mathrm{E}(X_i^2) < \infty, \quad i = 1, \ldots, n, \qquad \mathrm{E}(Z_j^2) < \infty, \quad j = 1, \ldots, m.$$

This is an existence condition and is necessary to have a rigorous, functional space setting for what follows.

Definition 3.23. *The* cross-correlation *matrix for* $\mathbf{X}$ *and* $\mathbf{Z}$ *is the* $n \times m$ *matrix* $\Gamma_{\mathbf{XZ}} = \mathrm{E}(\mathbf{X}\mathbf{Z}^T)$ *with entries*

$$[\Gamma_{\mathbf{XZ}}]_{ij} = \mathrm{E}(X_i Z_j), \quad i = 1, \ldots, n, \; j = 1, \ldots, m.$$

The autocorrelation *matrix for* $\mathbf{X}$ *is* $\Gamma_{\mathbf{XX}} = \mathrm{E}(\mathbf{X}\mathbf{X}^T)$*, with entries*

$$[\Gamma_{\mathbf{XX}}]_{ij} = \mathrm{E}(X_i X_j), \quad 1 \le i, j \le n.$$

Remark 3.24. *Note that* $\Gamma_{\mathbf{ZX}} = \Gamma_{\mathbf{XZ}}^T$ *and that* $\Gamma_{\mathbf{XX}}$ *is symmetric and positive semidefinite, i.e.,* $\forall \mathbf{x}$, $\mathbf{x}^{\mathrm{T}} \Gamma_{\mathbf{Xx}} \mathbf{x} \ge 0$*. Also, if* $\mathrm{E}(\mathbf{X}) = 0$*, then the autocorrelation reduces to the covariance,* $\Gamma_{\mathbf{XX}} = \mathrm{cov}(\mathbf{X})$.

We can relate the trace of the autocorrelation matrix to the second moment of the random vector $\mathbf{X}$.

Proposition 3.25. *If a random vector* $\mathbf{X}$ *has finite expected squared components, then*

$$\mathrm{E}\left(||\mathbf{X}||^2\right) = \mathrm{trace}(\Gamma_{\mathbf{XX}}).$$

We are now ready to formally define the BLUE. We consider a linear model,

$$\mathbf{z} = \mathbf{K}\mathbf{x} + \mathbf{N},$$

where $\mathbf{K}$ is an $m \times n$ matrix, $\mathbf{x} \in \mathbb{R}^n$ is deterministic, and $\mathbf{N}$ is a random (noise) n-vector with

$$\mathrm{E}(\mathbf{N}) = 0, \quad \mathbf{C}_{\mathbf{N}} = \mathrm{cov}(\mathbf{N}),$$

and $\mathbf{C}_{\mathbf{N}}$ is a known, nonsingular, $n \times n$ covariance matrix.

Definition 3.26. *The* best linear unbiased estimator (BLUE) *for* $\mathbf{x}$ *from the linear model* $\mathbf{z}$ *is the vector* $\hat{\mathbf{x}}_{\mathrm{BLUE}}$ *that minimizes the quadratic cost function*

$$J(\hat{\mathbf{x}}) = \mathrm{E}\left(||\hat{\mathbf{x}} - \mathbf{x}||^2\right)$$

subject to the constraints of linearity,

$$\hat{\mathbf{x}} = \mathbf{B}\mathbf{z}, \quad B \in \mathbb{R}^{n\times m},$$

and unbiasedness,

$$\mathrm{E}(\hat{\mathbf{x}}) = \mathbf{x}.$$

In the case of a full-rank matrix $\mathbf{K}$, the *Gauss–Markov theorem* [Sayed, 2003; Vogel, 2002] gives us an explicit form for the BLUE.

Theorem 3.27 (Gauss–Markov). *If* $\mathbf{K}$ *has full rank, then the BLUE is given by*

$$\hat{\mathbf{x}}_{\mathrm{BLUE}} = \hat{\mathbf{B}}\mathbf{z},$$

where

$$\hat{\mathbf{B}} = \left(\mathbf{K}^T\mathbf{C}_N^{-1}\mathbf{K}\right)^{-1}\mathbf{K}^T\mathbf{C}_N^{-1}.$$

Remark 3.28. *If the noise covariance matrix* $\mathbf{C}_N = \sigma^2 I$ *(white, uncorrelated noise), and* $\mathbf{K}$ *has full rank, then*

$$\hat{\mathbf{x}}_{\mathrm{BLUE}} = \left(\mathbf{K}^T\mathbf{K}\right)^{-1}\mathbf{K}^T\mathbf{z} = \mathbf{K}^\dagger\mathbf{z},$$

where $\mathbf{K}^\dagger$ *is called the pseudoinverse of* $\mathbf{K}$*. This corresponds, in the deterministic case, to the* least-squares problem

$$\min_{\mathbf{x}} ||\mathbf{K}\mathbf{x} - \mathbf{z}||.$$

Due to the dependence of the BLUE on the inverse of the noise covariance matrix, it is unsuitable for the solution of noisy, ill-conditioned linear systems. To remedy this situation, we assume that $\mathbf{x}$ is a realization of a random vector $\mathbf{X}$, and we formulate a linear least-squares analogue of Bayesian estimation.

Definition 3.29. *Suppose that* $\mathbf{x}$ *and* $\mathbf{z}$ *are jointly distributed, random vectors with finite expected squares. The* minimum variance linear estimator (MVLE) *of* $\mathbf{x}$ *from* $\mathbf{z}$ *is given by*

$$\hat{\mathbf{x}}_{\mathrm{MVLE}} = \hat{\mathbf{B}}\mathbf{z},$$

where

$$\hat{\mathbf{B}} = \arg\min_{B\in\mathbb{R}^{n\times m}} \mathrm{E}\left(||\mathbf{B}\mathbf{z} - \mathbf{x}||^2\right).$$

Proposition 3.30. *If* $\Gamma_{\mathbf{ZZ}}$ *is nonsingular, then the MVLE of* $\mathbf{x}$ *from* $\mathbf{z}$ *is given by*

$$\hat{\mathbf{x}}_{\mathrm{MVLE}} = \left(\Gamma_{\mathbf{XZ}}\Gamma_{\mathbf{ZZ}}^{-1}\right)\mathbf{z}.$$

3.3 ▪ Examples of Bayesian estimation

In this section we provide some calculated examples of Bayesian estimation.

3.3.1 ▪ Scalar Gaussian distribution example

In this simple, but important, example we will derive in detail the parameters of the posterior distribution when the data and the prior are normally distributed. This will provide us with a richer understanding of DA.

We suppose that we are interested in forecasting the value of a *scalar* state variable, x, which could be a temperature, a wind velocity component, an ozone concentration, etc. We are in possession of a Gaussian prior distribution for x,

$$x \sim \mathcal{N}(\mu_X, \sigma_X^2),$$

with expectation μ and variance σ^2, which could come from a forecast model, for example. We are in possession of n independent, noisy observations,

$$\mathbf{y} = (y_1, y_2, \ldots, y_n),$$

each with conditional distribution

$$y_i \mid x \sim \mathcal{N}(x, \sigma^2),$$

that are conditioned on the true value of the parameter/process x. Thus, the conditional distribution of the data/observations is a product of Gaussian laws,

$$\begin{aligned} p(\mathbf{y} \mid x) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - x)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x)^2\right\}. \end{aligned}$$

But from Bayes' law (3.5),

$$p(x \mid \mathbf{y}) \propto p(\mathbf{y} \mid x) p(x),$$

so using the data and the prior distributions/models, we have

$$\begin{aligned} p(x \mid \mathbf{y}) &\propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(y_i - x)^2/\sigma^2 + (x - \mu_X)^2/\sigma_X^2\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[x^2\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_X^2}\right) - 2\left(\sum_{i=1}^{n}\frac{y_i}{\sigma^2} + \frac{\mu_X}{\sigma_X^2}\right)x\right]\right\}. \end{aligned}$$

Notice that this is the product of two Gaussians, which, by completing the square, can be show to be Gaussian itself. This produces the posterior distribution,

$$x \mid \mathbf{y} \sim \mathcal{N}\left(\mu_{x|\mathbf{y}}, \sigma_{x|\mathbf{y}}^2\right), \tag{3.6}$$

where

$$\mu_{x|\mathbf{y}} = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_X^2}\right)^{-1} \left(\sum_{i=1}^{n} \frac{y_i}{\sigma^2} + \frac{\mu_X}{\sigma_X^2}\right)$$

and

$$\sigma_{x|\mathbf{y}}^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_X^2}\right)^{-1}.$$

Let us now study more closely these two parameters of the posterior law. We first remark that the inverse of the posterior variance, called the posterior *precision*, is equal to the sum of the prior precision, $1/\sigma_X^2$, and the data precision, n/σ^2. Second, the posterior mean, or conditional expectation, can also be written as a sum of two terms:

$$\begin{aligned} \mathrm{E}(x \mid \mathbf{y}) &= \frac{\sigma^2 \sigma_X^2}{\sigma^2 + n\sigma_X^2} \left(\frac{n}{\sigma^2}\bar{y} + \frac{\mu_X}{\sigma_X^2}\right) \\ &= w_y \bar{y} + w_{\mu_X} \mu_X, \end{aligned}$$

where the sample mean,

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i,$$

and the two weights,

$$w_y = \frac{n\sigma_X^2}{\sigma^2 + n\sigma_X^2}, \quad w_{\mu_X} = \frac{\sigma^2}{\sigma^2 + n\sigma_X^2},$$

add up to

$$w_y + w_{\mu_X} = 1.$$

We observe immediately that the posterior mean is the weighted sum/average of the data mean ($\bar{y}$) and the prior mean (μ_X). Now let us examine the weights themselves. If there is a large uncertainty in the prior, then $\sigma_X^2 \to \infty$ and hence $w_y \to 1$, $w_{\mu_X} \to 0$ and the likelihood dominates the prior, leading to what is known as the sampling distribution for the posterior:

$$p(x \mid \mathbf{y}) \to \mathcal{N}(\bar{y}, \sigma^2/n).$$

If we have a large number of observations, then $n \to \infty$ and the posterior now tends to the sample mean, whereas if we have few observations, then $n \to 0$ and the posterior

$$p(x \mid \mathbf{y}) \to \mathcal{N}(\mu_X, \sigma_X^2)$$

tends to the prior. In the case of equal uncertainties between data and prior, $\sigma^2 = \sigma_X^2$, and the prior mean has the weight of a single additional observation. Finally, if the uncertainties are small, either the prior is infinitely more precise than the data ($\sigma_X^2 \to 0$) or the data are perfectly precise ($\sigma^2 \to 0$).

We end this example by rewriting the posterior mean and variance in a special form. Let us start with the mean:

$$\begin{aligned} \mathrm{E}(x \mid \mathbf{y}) &= \mu_X + \frac{n\sigma_X^2}{\sigma^2 + n\sigma_X^2} (\bar{y} - \mu_X) \\ &= \mu_X + G(\bar{y} - \mu_X). \end{aligned} \tag{3.7}$$

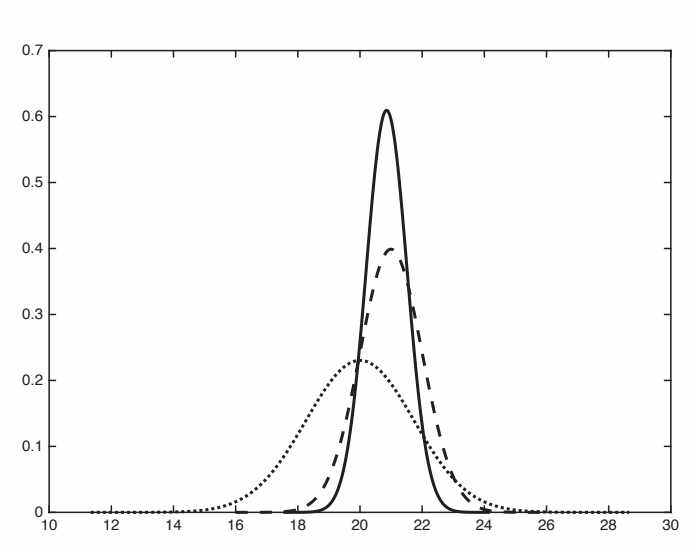

Figure 3.1. *Scalar Gaussian distribution example. Prior* $\mathcal{N}(20, 3)$ *(dotted), instrument* $\mathcal{N}(x, 1)$ *(dashed), and posterior* $\mathcal{N}(20.86, 0.43)$ *(solid) distributions.*

We conclude that the prior mean μ_X is adjusted toward the sample mean $\bar{y}$ by a gain (or amplification factor) of $G = 1/(1+\sigma^2/n\sigma_X^2)$, multiplied by the innovation $\bar{y}-\mu_X$, and we observe that the variance ratio, between data and prior, plays an essential role. In the same way, the posterior variance can be reformulated as

$$\sigma_{x|\mathrm{y}}^2 = (1-G)\sigma_X^2, \tag{3.8}$$

and the posterior variance is thus updated from the prior variance according to the same gain G. These last two equations, (3.7) and (3.8), are fundamental for a good understanding of DA, since they clearly express the interplay between prior and data and the effect that each has on the posterior.

Let us illustrate this with two initial numerical examples. Suppose we have a prior distribution $x \sim \mathcal{N}(\mu_X, \sigma_X^2)$ with mean 20 and variance 3. Suppose that our data model has the conditional law $y_i \mid x \sim \mathcal{N}(x, \sigma^2)$ with variance 1. Here the data are relatively precise compared to the prior. Say we have acquired two observations, $\mathbf{y} = (19, 23)'$. Now we can use (3.7) and (3.8) to compute the posterior distribution:

$$\mathrm{E}(x \mid \mathbf{y}) = 20 + \frac{6}{1+6}(21-20) = 20.86,$$
$$\sigma_{x|\mathrm{y}}^2 = \left(1-\frac{6}{7}\right)3 = 0.43,$$

thus yielding the posterior distribution

$$y_i \mid x \sim \mathcal{N}(20.86, 0.43),$$

which represents the update of the prior according to the observations and takes into account all the uncertainties available—see Figure 3.1. In other words, we have obtained a complete forecast at a given point in time.

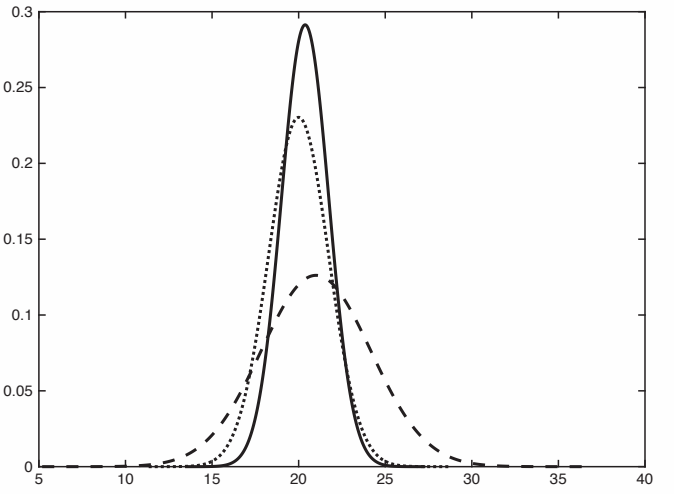

Figure 3.2. *Scalar Gaussian distribution example. Prior* $\mathcal{N}(20, 3)$ *(dotted), instrument* $\mathcal{N}(x, 10)$ *(dashed), and posterior* $\mathcal{N}(20.375, 1.875)$ *(solid) distributions.*

Now consider the same prior, $x \sim \mathcal{N}(20, 3)$, but with a relatively uncertain/imprecise observation model, $y_i \mid x \sim \mathcal{N}(x, 10)$, and the same two measurements, $\mathbf{y} = (19, 23)'$. Redoing the above calculations, we now find

$$\mathrm{E}(x \mid \mathbf{y}) = 20 + \frac{6}{16}(21 - 20) = 20.375,$$
$$\sigma^2_{x|\mathbf{y}} = \left(1 - \frac{6}{16}\right)3 = 1.875,$$

thus yielding the new posterior distribution,

$$y_i \mid x \sim \mathcal{N}(20.375, 1.875),$$

which has virtually the same mean but a much larger variance—see Figure 3.2, where the scales on both axes have changed.

3.3.2 ▪ Estimating a temperature

Suppose that the outside temperature measurement gives 2°C and the instrument has an error distribution that is Gaussian with mean $\mu = 2$ and variance $\sigma^2 = 0.64$—see the dashed curve in Figure 3.3. This is the model/data distribution. We also suppose that we have a prior distribution that estimates the temperature, with mean $\mu = 0$ and variance $\sigma^2 = 1.21$. The prior comes from either other observations, a previous model forecast, or physical and climatological constraints—see the dotted curve in Figure 3.3. By combining these two, using Bayes' formula, we readily compute the posterior distribution of the temperature given the observations, which has mean $\mu = 1.31$ and variance $\sigma^2 = 0.42$. This is the update or the analysis—see the solid curve in Figure 3.3.

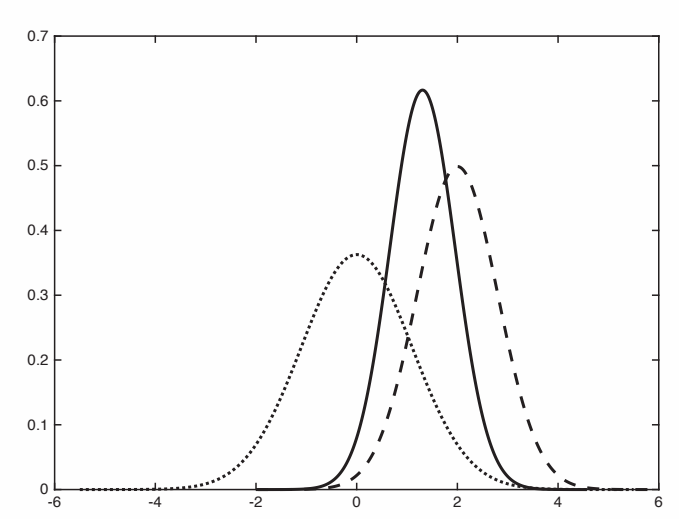

Figure 3.3. *A Gaussian product example for forecasting temperature: prior (dotted), instrument (dashed), and posterior (solid) distributions.*

The code for this calculation can be found in `gaussian_product.m` [DART toolbox, 2013].

3.3.3 ▪ Estimating the parameters of a pendulum

We present an example of a simple mechanical system and seek an estimation of its parameters from noisy measurements. Consider a model for the angular displacement, x_t, of an ideal pendulum (no friction, no drag),

$$x_t = \sin(\theta t) + \epsilon_t,$$

where ϵ_t is a Gaussian noise with zero mean and variance σ^2, the pendulum parameter is denoted by θ, and t is time. From these noisy measurements (suppose that the instrument is not very accurate) of x_t we want to estimate θ, which represents the physical properties of the pendulum—in fact $\theta = \sqrt{g/L}$, where g is the gravitational constant and L is the pendulum's length. Using this physical model, can we estimate (or infer) the unknown physical parameters of the pendulum?

If the measurements are independent, then the likelihood of a set of T observations $x_1, \ldots, x_T$ is given by the product

$$p(x_1, \ldots, x_T|\theta) = \prod_{t=1}^{T} p(x_t|\theta).$$

In addition, suppose that we have some prior estimation (before obtaining the measurements) of the probabilities of a set of possible values of θ. Then the posterior distribution of θ, given the measurements, can be calculated from Bayes' law, as seen above,

$$p(\theta|x_1, \ldots, x_T) \propto p(\theta) \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_t - \sin(\theta t))^2},$$

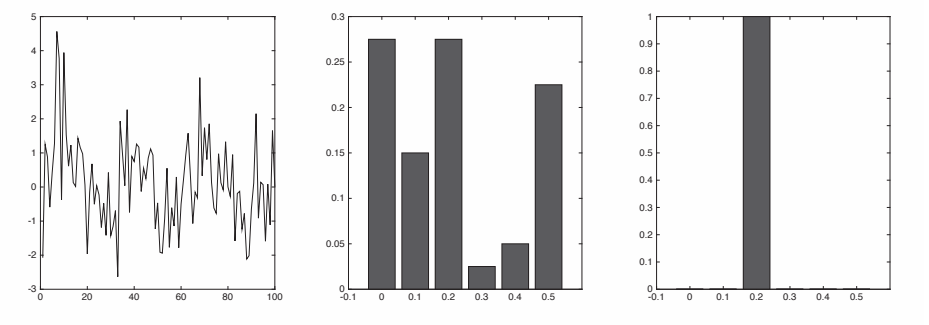

Figure 3.4. *Bayesian estimation of noisy pendulum parameter, $\theta = 0.2$. Observations of 100 noisy positions (left). Prior distribution of parameter values (center). Posterior distribution for θ (right).*

where we have omitted the denominator. We are given the following table of priors:

$[\theta_{\min}, \theta_{\max}]$	$p(\theta_{\min} < \theta < \theta_{\max})$
$[0, 0.05]$	0.275
$[0.05, 0.15]$	0.15
$[0.15, 0.25]$	0.275
$[0.25, 0.35]$	0.025
$[0.35, 0.45]$	0.05
$[0.45, 0.55]$	0.225

After performing numerical simulations, we observe (see Figure 3.4) that the posterior for θ develops a prominent peak for a large number ($T = 100$) of measurements, centered around the real value $\theta = 0.2$ (which was used to generate the time series, x_t).

3.3.4 ▪ Vector/multivariate Gaussian distribution example

As a final case that will lead us naturally to the following section, let us consider the vector/multivariate extension of the example in Section 3.3.1. We will now study a vector process, $\mathbf{x}$, with n components and a prior distribution

$$\mathbf{x} \sim \mathcal{N}(\mu, \mathbf{B}),$$

where the mean vector μ and the covariance matrix $\mathbf{B}$ are assumed to be known (as usual from historical data, model forecasts, etc.). The observation now takes the form of a data vector, $\mathbf{y}$, of dimension p and has the conditional distribution/model:

$$\mathbf{y} \mid \mathbf{x} \sim \mathcal{N}(\mathbf{Hx}, \mathbf{R}),$$

where the $(p \times n)$ observation matrix $\mathbf{H}$ maps the process to the measurements and the error covariance matrix $\mathbf{R}$ is known. These are exactly the same matrices that we have already encountered in the variational approach—see Chapters 1 and 2. The difference is that now our modeling is placed in a richer, Bayesian framework.

As before, we would like to calculate the posterior conditional distribution of $\mathbf{x} \mid \mathbf{y}$, given by

$$p(\mathbf{x} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \mathbf{x}) p(\mathbf{x}).$$

Just as with the scalar/univariate case, the product of two Gaussians is Gaussian, and the posterior law is the multidimensional analogue of (3.6) and can be shown to take the form

$$\mathbf{x} \mid \mathbf{y} \sim \mathcal{N}\left(\mu_{x|\mathbf{y}}, \Sigma_{x|\mathbf{y}}\right),$$

where

$$\mu_{x|\mathbf{y}} = \left(\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H} + \mathbf{B}^{-1}\right)^{-1}\left(\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}y + \mathbf{B}^{-1}\mu\right)$$

and

$$\Sigma_{x|\mathbf{y}} = \left(\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H} + \mathbf{B}^{-1}\right)^{-1}.$$

As above, we will now rewrite the posterior mean and variance in a special form. The posterior conditional mean becomes

$$\begin{aligned} \mathrm{E}(\mathbf{x} \mid \mathbf{y}) &= \left(\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H} + \mathbf{B}^{-1}\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}y + \left(\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H} + \mathbf{B}^{-1}\right)^{-1}\mathbf{B}^{-1}\mu \\ &= \mu + \mathbf{K}(y - \mathbf{H}\mu), \end{aligned} \tag{3.9}$$

where the gain matrix is now

$$\mathbf{K} = \mathbf{B}\mathbf{H}^{\mathrm{T}}(\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}})^{-1}.$$

In the same manner, the posterior conditional covariance matrix can be reformulated as

$$\begin{aligned} \Sigma(\mathbf{x} \mid \mathbf{y}) &= \left(\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H} + \mathbf{B}^{-1}\right)^{-1} \\ &= (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{B}, \end{aligned} \tag{3.10}$$

with the same gain matrix $\mathbf{K}$ as for the posterior mean. As before, these last two equations, (3.9) and (3.10), are fundamental for a good understanding of DA, since they clearly express the interplay between prior and data and the effect that each has on the posterior. They are, in fact, the veritable foundation of DA.

3.3.5 ▪ Connections with variational and sequential approaches

As was already indicated in the first two chapters of this book, the link between variational approaches and optimal BLUE is well established. The BLUE approach is also known as kriging in spatial geostatistics, or optimal interpolation (OI) in oceanography and atmospheric science. In the special, but quite widespread, case of a multivariate Gaussian model (for data and priors), the posterior mode (which is equivalent to the mean in this case) can equally be obtained by minimizing the quadratic objective function (2.31),

$$J(\mathbf{x}) = \frac{1}{2}\left(\mathbf{x} - \mathbf{x}^{\mathrm{b}}\right)^{\mathrm{T}}\mathbf{B}^{-1}\left(\mathbf{x} - \mathbf{x}^{\mathrm{b}}\right) + \frac{1}{2}(\mathbf{H}\mathbf{x} - \mathbf{y})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{H}\mathbf{x} - \mathbf{y}).$$

This is the fundamental link between variational and statistical approaches. Though strictly equivalent to the Bayes formulation, the variational approach has, until now, been privileged for operational, high-dimensional DA problems—though this is changing with the arrival of new hardware and software capabilities for treating "big data and extreme computing" challenges [Reed and Dongarra, 2015].

Since many physical systems are dynamic and evolve in time, we could improve our estimations considerably if, as new measurements became available, we could simply update the previous optimal estimate of the state process without having to redo all computations. The perfect framework for this sequential updating is the KF, which we will now present in detail.

3.4 ▪ Sequential DA and Kalman filters

We have seen that DA, from the statistical/Bayesian point of view, strives to have as complete a knowledge as possible of the a posteriori probability law, that is, the conditional law of the state given the observations. But it is virtually impossible to determine the complete distribution, so we seek instead an estimate of its statistical parameters, such as its mean and/or its variance. Numerous proven statistical methods can lead to best or optimal estimates [Anderson and Moore, 1979; Garthwaite et al., 2002; Hogg et al., 2013; Ross, 2014]; for example, the minimum variance (MV) estimator is the conditional mean of the state given the observations, and the maximum a posteriori (MAP) estimator produces the mode of the conditional distribution. As seen above, assuming Gaussian distributions for the measurements and the process, we can determine the complete a posteriori law, and, in this case, it is clear that the MV and MAP estimators coincide. In fact, the MV estimator produces the optimal interpolation (OI) or kriging equations, whereas the MAP estimator leads to 3D-Var. In conclusion, for the case of a linear observation operator together with Gaussian error statistics, 3D-Var and OI are strictly equivalent.

So far we have been looking at the spatial estimation problem, where all observations are distributed in space but at a single instant in time. For stationary stochastic processes, the mean and covariance are constant in time [Parzen, 1999; Ross, 1997], so such a DA scheme could be used at different times based on the invariant statistics. This is not so rare: in practice, for global NWP, the errors have been considered stationary over a one-month time scale.[30] However, for general environmental applications, the governing equations vary with time and we must take into account nonstationary processes.

Within the significant box of mathematical tools that can be used for statistical estimation from noisy sensor measurements over time, one of the most well known and often used tools is the Kalman filter (KF). The KF is named after Rudolph E. Kalman, who in 1960 published his famous paper describing a recursive solution to the discrete data linear filtering problem [Kalman, 1960]. There exits a vast literature on the KF, and a very "friendly" introduction to the general idea of the KF can be found in Chapter 1 of Maybeck [1979]. As just stated above, it would be ideal and very efficient if, as new data or measurements became available, we could easily update the previous optimal estimates without having to recompute everything. The KF provides exactly this solution.

To this end, we will now consider a dynamical system that evolves in time, and we will seek to estimate a series of *true* states, $\mathbf{x}_k^t$ (a sequence of random vectors), where discrete time is indexed by the letter k. These times are those when the observations or measurements are taken, as shown in Figure 3.5. The assimilation starts with an unconstrained model trajectory from $t_0, t_1, \ldots, t_{k-1}, t_k, \ldots, t_n$ and aims to provide an optimal fit to the available observations/measurements given their uncertainties (error bars). For example, in current synoptic scale weather forecasts, $t_k - t_{k-1} = 6$ hours; the time step is less for the convective scale.

3.4.1 ▪ Bayesian modeling

Let us recall the principles of Bayesian modeling from Section 3.2 on statistical estimation and rewrite them in the terminology of the DA problem. We have a vector, $\mathbf{x}$, of (unknown) unobserved quantities of interest (temperature, pressure, wind, etc.) and

[30] This assumption is no longer employed at MétéoFrance or ECWMF, for example.

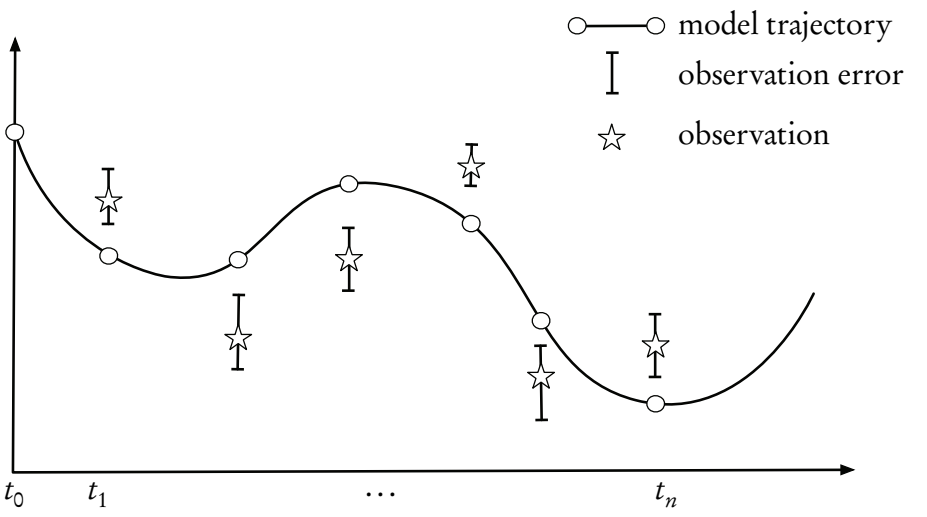

Figure 3.5. *Sequential assimilation: a computed model trajectory, observations, and their error bars.*

a vector, $\mathbf{y}$, of (known) observed data (at various locations, and at various times). The full joint probability model can always be factored into two components,

$$\begin{aligned} p(\mathbf{x},\mathbf{y}) &= p(\mathbf{y}\,|\,\mathbf{x})p(\mathbf{x}) \\ &= p(\mathbf{x}\,|\,\mathbf{y})p(\mathbf{y}), \end{aligned}$$

and thus

$$p(\mathbf{x}\,|\,\mathbf{y}) = \frac{p(\mathbf{y}\,|\,\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})},$$

provided that $p(\mathbf{y}) \neq 0$.

The KF can be rigorously derived from this Bayesian perspective following the presentation above in Section 3.3.4.

3.4.2 ▪ Stochastic model of the system

We seek to estimate the state $\mathbf{x} \in \mathbb{R}^n$ of a discrete-time dynamic process that is governed by the linear stochastic difference equation

$$\mathbf{x}_{k+1} = \mathbf{M}_{k+1}\mathbf{x}_k + \mathbf{w}_k, \tag{3.11}$$

with a measurement/observation $\mathbf{y} \in \mathbb{R}^m$:

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{x}_k + \mathbf{v}_k. \tag{3.12}$$

Note that $\mathbf{M}_{k+1}$ and $\mathbf{H}_k$ are considered linear here. The random vectors $\mathbf{w}_k$ and $\mathbf{v}_k$ represent the process/modeling and measurement/observation errors, respectively. They are assumed to be independent, white-noise processes with Gaussian/normal probability distributions

$$\begin{aligned} \mathbf{w}_k &\sim \mathcal{N}(0, \mathbf{Q}_k), \\ \mathbf{v}_k &\sim \mathcal{N}(0, \mathbf{R}_k), \end{aligned}$$

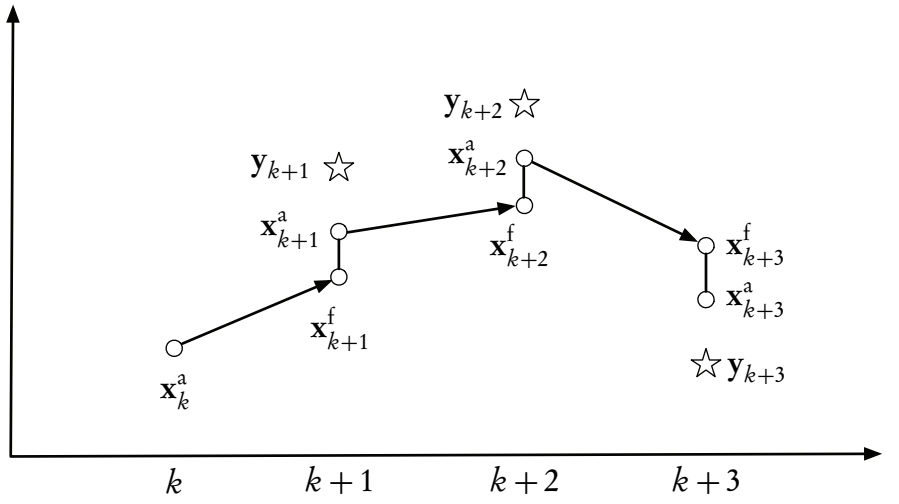

Figure 3.6. *Sequential assimilation scheme for the KF. The x-axis denotes time; the y-axis denotes the values of the state and observation vectors.*

where $\mathbf{Q}$ and $\mathbf{R}$ are the covariance matrices (assumed known) of the modeling and observation errors, respectively. All these assumptions about unbiased and uncorrelated errors (in time and between each other) are not limiting, since extensions of the standard KF can be developed should any of these not be valid—see below and Chapters 5, 6, and 7.

We note that for a broader mathematical view on the above system, we could formulate everything in terms of stochastic differential equations (SDEs). Then the theory of Itô can provide a detailed solution of the problem of optimal filtering as well as existence and uniqueness results—see Oksendal [2003], where one can find such a precise mathematical formulation.

3.4.3 ▪ Sequential assimilation scheme

The typical assimilation scheme is made up of two major steps: a prediction/forecast step and a correction/analysis step. At time t_k we have the result of a previous forecast, $\mathbf{x}^{\mathrm{f}}_k$ (the analogue of the background state $\mathbf{x}^{\mathrm{b}}_k$), and the result of an ensemble of observations in $\mathbf{y}_k$. Based on these two vectors, we perform an analysis that produces $\mathbf{x}^{\mathrm{a}}_k$. We then use the evolution model to obtain a prediction of the state at time t_{k+1}. The result of the forecast is denoted $\mathbf{x}^{\mathrm{f}}_{k+1}$ and becomes the background, or initial guess, for the next time step. This process is summarized in Figure 3.6. The KF problem can be summarized as follows: given a prior/background estimate, $\mathbf{x}^{\mathrm{f}}$, of the system state at time t_k, what is the best update/analysis, $\mathbf{x}^{\mathrm{a}}_k$, based on the currently available measurements, $\mathbf{y}_k$?

We can now define forecast (a priori) and analysis (a posteriori) estimate errors as

$$\mathbf{e}^{\mathrm{f}}_k = \mathbf{x}^{\mathrm{f}}_k - \mathbf{x}^{\mathrm{t}}_k,$$
$$\mathbf{e}^{\mathrm{a}}_k = \mathbf{x}^{\mathrm{a}}_k - \mathbf{x}^{\mathrm{t}}_k,$$

where $\mathbf{x}_k^{\mathrm{t}}$ is the (unknown) true state. Their respective error covariance matrices are

$$\begin{aligned}\mathbf{P}_k^{\mathrm{f}} &= \mathrm{cov}(\mathbf{e}_k^{\mathrm{f}}) = E\left[\mathbf{e}_k^{\mathrm{f}}(\mathbf{e}_k^{\mathrm{f}})^{\mathrm{T}}\right],\\ \mathbf{P}_k^{\mathrm{a}} &= \mathrm{cov}(\mathbf{e}_k^{\mathrm{a}}) = E\left[\mathbf{e}_k^{\mathrm{a}}(\mathbf{e}_k^{\mathrm{a}})^{\mathrm{T}}\right].\end{aligned} \tag{3.13}$$

The goal of the KF is to compute an optimal a posteriori estimate, $\mathbf{x}_k^{\mathrm{a}}$, that is a linear combination of an a priori estimate, $\mathbf{x}_k^{\mathrm{f}}$, and a weighted difference between the actual measurement, $\mathbf{y}_k$, and the measurement prediction, $\mathbf{H}_k\mathbf{x}_k^{\mathrm{f}}$. This is none other than the BLUE that we have seen above. The filter is thus of the linear, recursive form

$$\mathbf{x}_k^{\mathrm{a}} = \mathbf{x}_k^{\mathrm{f}} + \mathbf{K}_k\left(\mathbf{y}_k - \mathbf{H}_k\mathbf{x}_k^{\mathrm{f}}\right). \tag{3.14}$$

The difference $\mathbf{d}_k = \mathbf{y}_k - \mathbf{H}_k\mathbf{x}_k^{\mathrm{f}}$ is called the *innovation* and reflects the discrepancy between the actual and the predicted measurements at time t_k. Note that for generality, the matrices are shown with a time dependence. When this is not the case, the subscripts k can be dropped. The *Kalman gain* matrix, $\mathbf{K}$, is chosen to minimize the a posteriori error covariance equation (3.13).

To compute this *optimal gain* requires a careful derivation. Begin by substituting the observation equation (3.12) into the linear filter equation (3.14):

$$\begin{aligned}\mathbf{x}_k^{\mathrm{a}} &= \mathbf{x}_k^{\mathrm{f}} + \mathbf{K}_k\left(\mathbf{H}_k\mathbf{x}_k^{\mathrm{t}} + \mathbf{v}_k - \mathbf{H}_k\mathbf{x}_k^{\mathrm{f}}\right)\\ &= \mathbf{x}_k^{\mathrm{f}} + \mathbf{K}_k\left(\mathbf{H}_k\left(\mathbf{x}_k^{\mathrm{t}} - \mathbf{x}_k^{\mathrm{f}}\right) + \mathbf{v}_k\right).\end{aligned}$$

Now place this last expression into the definition of $\mathbf{e}_k^{\mathrm{a}}$:

$$\begin{aligned}\mathbf{e}_k^{\mathrm{a}} &= \mathbf{x}_k^{\mathrm{a}} - \mathbf{x}_k^{\mathrm{t}}\\ &= \mathbf{x}_k^{\mathrm{f}} + \mathbf{K}_k\left(\mathbf{H}_k\left(\mathbf{x}_k^{\mathrm{t}} - \mathbf{x}_k^{\mathrm{f}}\right) + \mathbf{v}_k\right) - \mathbf{x}_k^{\mathrm{t}}\\ &= \mathbf{K}_k\left(-\mathbf{H}_k\left(\mathbf{x}_k^{\mathrm{f}} - \mathbf{x}_k^{\mathrm{t}}\right) + \mathbf{v}_k\right) + \left(\mathbf{x}_k^{\mathrm{f}} - \mathbf{x}_k^{\mathrm{t}}\right).\end{aligned}$$

Then substitute in the error covariance equation (3.13):

$$\begin{aligned}\mathbf{P}_k^{\mathrm{a}} &= \mathrm{E}\left[\mathbf{e}_k^{\mathrm{a}}(\mathbf{e}_k^{\mathrm{a}})^{\mathrm{T}}\right]\\ &= \mathrm{E}\left[\left(\mathbf{K}_k\left(\mathbf{v}_k - \mathbf{H}_k\left(\mathbf{x}_k^{\mathrm{f}} - \mathbf{x}_k^{\mathrm{t}}\right)\right) + \left(\mathbf{x}_k^{\mathrm{f}} - \mathbf{x}_k^{\mathrm{t}}\right)\right)\left(\mathbf{K}_k\left(\mathbf{v}_k - \mathbf{H}_k\left(\mathbf{x}_k^{\mathrm{f}} - \mathbf{x}_k^{\mathrm{t}}\right)\right) + \left(\mathbf{x}_k^{\mathrm{f}} - \mathbf{x}_k^{\mathrm{t}}\right)\right)^{\mathrm{T}}\right].\end{aligned}$$

Now perform the indicated expectations over the r.v.'s, noting that $\left(\mathbf{x}_k^{\mathrm{f}} - \mathbf{x}_k^{\mathrm{t}}\right) = \mathbf{e}_k^{\mathrm{f}}$ is the a priori estimation error, that this error is uncorrelated with the observation error $\mathbf{v}_k$, that by definition $\mathbf{P}_k^{\mathrm{f}} = \mathrm{E}\left[\mathbf{e}_k^{\mathrm{f}}(\mathbf{e}_k^{\mathrm{f}})^T\right]$ and that $\mathbf{R}_k = \mathrm{E}\left[\mathbf{v}_k\mathbf{v}_k^{\mathrm{T}}\right]$. We thus get

$$\begin{aligned}\mathbf{P}_k^{\mathrm{a}} &= \mathrm{E}\left[\left(\mathbf{K}_k\left(\mathbf{v}_k - \mathbf{H}_k\mathbf{e}_k^{\mathrm{f}}\right) + \mathbf{e}_k^{\mathrm{f}}\right)\left(\mathbf{K}_k\left(\mathbf{v}_k - \mathbf{H}_k\mathbf{e}_k^{\mathrm{f}}\right) + \mathbf{e}_k^{\mathrm{f}}\right)^{\mathrm{T}}\right]\\ &= \mathrm{E}\left[\left(\mathbf{e}_k^{\mathrm{f}}\right)\left(\mathbf{e}_k^{\mathrm{f}}\right)^{\mathrm{T}} - \left(\mathbf{K}_k\mathbf{H}_k\mathbf{e}_k^{\mathrm{f}}\right)\left(\mathbf{K}_k\mathbf{H}_k\mathbf{e}_k^{\mathrm{f}}\right)^{\mathrm{T}} + (\mathbf{K}_k\mathbf{v}_k)(\mathbf{K}_k\mathbf{v}_k)^{\mathrm{T}}\right]\\ &= (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_k^{\mathrm{f}}(\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)^{\mathrm{T}} + \mathbf{K}_k\mathbf{R}_k\mathbf{K}_k^{\mathrm{T}}.\end{aligned} \tag{3.15}$$

Note that this is a completely general formula for the updated covariance matrix and that it is valid for any gain $\mathbf{K}_k$, not necessarily optimal.

Now we still need to compute the optimal gain that minimizes the matrix entries along the principal diagonal of $\mathbf{P}_k^{\mathrm{a}}$, since these terms are the ones that represent the estimation error variances for the entries of the state vector itself. We will use the classical approach of variational calculus, by taking the derivative of the trace of the result with respect to $\mathbf{K}$ and then setting the resulting derivative expression equal to zero. But for this, we require two results from matrix differential calculus [Petersen and Pedersen, 2012]. These are

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{A}}\mathrm{Tr}(\mathbf{AB}) = \mathbf{B}^{\mathrm{T}},$$
$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{A}}\mathrm{Tr}\left(\mathbf{ACA}^{\mathrm{T}}\right) = 2\mathbf{AC},$$

where Tr denotes the matrix trace operator and we assume that $\mathbf{AB}$ is square and that $\mathbf{C}$ is a symmetric matrix. The derivative of a scalar quantity with respect to a matrix is defined as the matrix of derivatives of the scalar with respect to each element of the matrix. Before differentiating, we expand (3.15) to obtain

$$\mathbf{P}_k^{\mathrm{a}} = \mathbf{P}_k^{\mathrm{f}} - \mathbf{K}_k\mathbf{H}_k\mathbf{P}_k^{\mathrm{f}} - \mathbf{P}_k^{\mathrm{f}}\mathbf{H}_k^{\mathrm{T}}\mathbf{K}_k^{\mathrm{T}} + \mathbf{K}_k\left(\mathbf{H}_k\mathbf{K}_k\mathbf{H}_k^{\mathrm{T}} + \mathbf{R}_k\right)\mathbf{K}_k^{\mathrm{T}}.$$

There are two linear terms and one quadratic term in $\mathbf{K}_k$. To minimize the trace of $\mathbf{P}_k^{\mathrm{a}}$, we can now apply the above matrix differentiation formulas (supposing that the individual squared errors are also minimized when their sum is minimized) to obtain

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{K}_k}\mathrm{Tr}\,\mathbf{P}_k^{\mathrm{a}} = -2\left(\mathbf{H}_k\mathbf{P}_k^{\mathrm{f}}\right)^{\mathrm{T}} + 2\mathbf{K}_k\left(\mathbf{H}_k\mathbf{K}_k\mathbf{H}_k^{\mathrm{T}} + \mathbf{R}_k\right).$$

Setting this last result equal to zero, we can finally solve for the optimal gain. The resulting $\mathbf{K}$ that minimizes equation (3.13) is given by

$$\mathbf{K}_k = \mathbf{P}_k^{\mathrm{f}}\mathbf{H}_k^{\mathrm{T}}\left(\mathbf{H}_k\mathbf{P}_k^{\mathrm{f}}\mathbf{H}_k^{\mathrm{T}} + \mathbf{R}_k\right)^{-1}, \tag{3.16}$$

where we remark that $\mathbf{H}\mathbf{P}_k^{\mathrm{f}}\mathbf{H}_k^{\mathrm{T}} + \mathbf{R}_k = \mathrm{E}\left[\mathbf{d}_k\mathbf{d}_k^{\mathrm{T}}\right]$ is the covariance of the innovation. Looking at this expression for $\mathbf{K}_k$, we see that when the measurement error covariance, $\mathbf{R}_k$, approaches zero, the gain, $\mathbf{K}_k$, weights the innovation more heavily, since

$$\lim_{\mathbf{R}\to 0}\mathbf{K}_k = \mathbf{H}_k^{-1}.$$

On the other hand, as the a priori error estimate covariance, $\mathbf{P}_k^{\mathrm{f}}$, approaches zero, the gain, $\mathbf{K}_k$, weights the innovation less heavily, and

$$\lim_{\mathbf{P}_k^{\mathrm{f}}\to 0}\mathbf{K}_k = 0.$$

Another way of thinking about the weighting of $\mathbf{K}$ is that as the measurement error covariance, $\mathbf{R}$, approaches zero, the actual measurement, $\mathbf{y}_k$, is "trusted" more and more, while the predicted measurement, $\mathbf{H}_k\mathbf{x}_k^{\mathrm{f}}$, is trusted less and less. On the other hand, as the a priori error estimate covariance, $\mathbf{P}_k^{\mathrm{f}}$, approaches zero, the actual measurement, $\mathbf{y}_k$, is trusted less and less, while the predicted measurement, $\mathbf{H}_k\mathbf{x}_k^{\mathrm{f}}$, is trusted more and more—see the computational example below.

The covariance matrix associated with the optimal gain can now be computed from (3.15). We already have

$$\begin{aligned}\mathbf{P}_k^{\mathrm{a}} &= (\mathbf{I}-\mathbf{K}_k\mathbf{H}_k)\,\mathbf{P}_k^{\mathrm{f}}\,(\mathbf{I}-\mathbf{K}_k\mathbf{H}_k)^{\mathrm{T}} + \mathbf{K}_k\mathbf{R}_k\mathbf{K}_k^{\mathrm{T}} \\ &= \mathbf{P}_k^{\mathrm{f}} - \mathbf{K}_k\mathbf{H}_k\mathbf{P}_k^{\mathrm{f}} - \mathbf{P}_k^{\mathrm{f}}\mathbf{H}_k^{\mathrm{T}}\mathbf{K}_k^{\mathrm{T}} + \mathbf{K}_k\left(\mathbf{H}_k\mathbf{K}_k\mathbf{H}_k^{\mathrm{T}} + \mathbf{R}_k\right)\mathbf{K}_k^{\mathrm{T}},\end{aligned}$$

and, substituting the optimal gain (3.16), we can derive three more alternative expressions:

$$\mathbf{P}_k^{\mathrm{a}} = \mathbf{P}_k^{\mathrm{f}} - \mathbf{P}_k^{\mathrm{f}}\mathbf{H}_k^{\mathrm{T}}\left(\mathbf{H}_k\mathbf{P}_k^{\mathrm{f}}\mathbf{H}_k^{\mathrm{T}} + \mathbf{R}_k\right)^{-1}\mathbf{H}_k\mathbf{P}_k^{\mathrm{f}},$$

$$\mathbf{P}_k^{\mathrm{a}} = \mathbf{P}_k^{\mathrm{f}} - \mathbf{K}_k\left(\mathbf{H}_k\mathbf{P}_k^{\mathrm{f}}\mathbf{H}_k^{\mathrm{T}} + \mathbf{R}_k\right)\mathbf{K}_k^{\mathrm{T}},$$

and

$$\mathbf{P}_k^{\mathrm{a}} = (\mathbf{I}-\mathbf{K}_k\mathbf{H}_k)\,\mathbf{P}_k^{\mathrm{f}}. \tag{3.17}$$

Each of these four expressions for $\mathbf{P}_k^{\mathrm{a}}$ would give the same results with perfectly precise arithmetic, but in real-world applications some may perform better numerically. In what follows, we will use the simplest form (3.17), but this is by no means restrictive, and any one of the others could be substituted.

3.4.3.1 ▪ Predictor/forecast step

We start from a previous analyzed state, $\mathbf{x}_k^{\mathrm{a}}$, or from the initial state if $k = 0$, characterized by the Gaussian PDF $p(\mathbf{x}_k^{\mathrm{a}} \mid \mathbf{y}_{1:k}^{\mathrm{o}})$ of mean $\mathbf{x}_k^{\mathrm{a}}$ and covariance matrix $\mathbf{P}_k^{\mathrm{a}}$. We use here the classical notation $\mathbf{y}_{i:j} = (\mathbf{y}_i, \mathbf{y}_{i+1}, \ldots, \mathbf{y}_j)$ for $i \le j$ that denotes conditioning on all the observations in the interval. An estimate of $\mathbf{x}_{k+1}^{\mathrm{t}}$ is given by the dynamical model, which defines the forecast as

$$\mathbf{x}_{k+1}^{\mathrm{f}} = \mathbf{M}_{k+1}\mathbf{x}_k^{\mathrm{a}}, \tag{3.18}$$

$$\mathbf{P}_{k+1}^{\mathrm{f}} = \mathbf{M}_{k+1}\mathbf{P}_k^{\mathrm{a}}\mathbf{M}_{k+1}^{T} + \mathbf{Q}_{k+1}, \tag{3.19}$$

where the expression for $\mathbf{P}_{k+1}^{\mathrm{f}}$ is obtained from the dynamics equation and the definition of the model noise covariance, $\mathbf{Q}$.

3.4.3.2 ▪ Corrector/analysis step

At time t_{k+1}, the PDF $p(\mathbf{x}_{k+1}^{\mathrm{f}} \mid \mathbf{y}_{1:k}^{\mathrm{o}})$ is known, thanks to the mean, $\mathbf{x}_{k+1}^{\mathrm{f}}$, and covariance matrix, $\mathbf{P}_{k+1}^{\mathrm{f}}$, just calculated, as well as the assumption of a Gaussian distribution. The analysis step then consists of correcting this PDF using the observation available at time t_{k+1} to compute $p(\mathbf{x}_{k+1}^{\mathrm{a}} \mid \mathbf{y}_{k+1:1}^{\mathrm{o}})$. This comes from the BLUE in the dynamical context and gives

$$\mathbf{K}_{k+1} = \mathbf{P}_{k+1}^{\mathrm{f}}\mathbf{H}^{T}\left(\mathbf{H}\mathbf{P}_{k+1}^{\mathrm{f}}\mathbf{H}^{T} + \mathbf{R}_{k+1}\right)^{-1}, \tag{3.20}$$

$$\mathbf{x}_{k+1}^{\mathrm{a}} = \mathbf{x}_{k+1}^{\mathrm{f}} + \mathbf{K}_{k+1}\left(\mathbf{y}_{k+1} - \mathbf{H}\mathbf{x}_{k+1}^{\mathrm{f}}\right), \tag{3.21}$$

$$\mathbf{P}_{k+1}^{\mathrm{a}} = \left(\mathbf{I}-\mathbf{K}_{k+1}\mathbf{H}\right)\mathbf{P}_{k+1}^{\mathrm{f}}. \tag{3.22}$$

The predictor–corrector loop is illustrated in Figure 3.7 and can be immediately transposed into an operational algorithm.

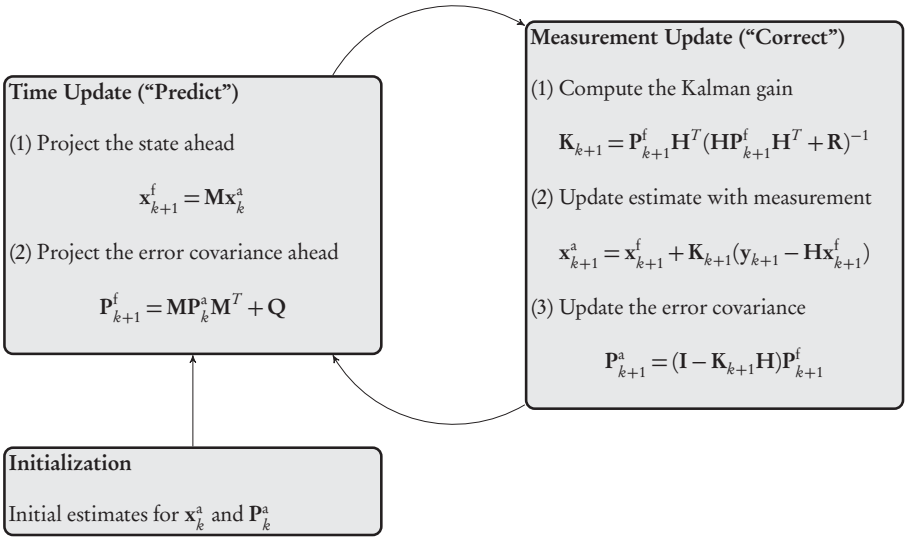

Figure 3.7. *Kalman filter loop.*

3.4.4 ▪ Note on the relation between Bayes and BLUE

If we know that the a priori and the observation data are both Gaussian, Bayes' rule can be applied to compute the a posteriori PDF. The a posteriori PDF is then Gaussian, and its parameters are given by the BLUE equations. Hence, with Gaussian PDFs and a linear observation operator, there is no need to use Bayes' rule. The BLUE equations can be used instead to compute the parameters of the resulting PDF. Since the BLUE provides the same result as Bayes' rule, it is the best estimator of all.

In addition (see the previous chapter), one can recognize the 3D-Var cost function. By minimizing this cost function, 3D-Var finds the MAP estimate of the Gaussian PDF, which is equivalent to the MV estimate found by the BLUE.

3.5 ▪ Implementation of the Kalman Filter

We now describe some important implementation issues and discuss ways to overcome the difficulties that they give rise to.

3.5.1 ▪ Stability and convergence

Stability is a concern for any dynamic system. The KF will be uniformly asymptotically stable if the system model itself is controllable and observable. The reader is referred to Friedland [1986], Gelb [1974], and Maybeck [1979] for detailed explanations of these concepts.

If the model is linear and time invariant (i.e., system matrices do not vary with time), the autocovariances will converge toward steady-state values. Consequently, the KF gain will converge toward a steady-state KF gain value that can be precalculated by solving an algebraic Ricatti equation. It is quite common to use only the steady-state gain in applications. For a nonlinear system, the gain may vary with the operating point (if the system matrix of the linearized model varies with the operating point).

In practical applications, the gain may be recalculated as the operating point changes.

In practical situations, there are a vast number of different sources for nonconvergence. In Grewal and Andrews [2001], the reader can find a very well explained presentation of all these (and many more). In particular, as we will point out, there are various remedies for

- convergence, divergence, and failure to converge;
- testing for unpredictable behavior;
- effects due to incorrect modeling;
- reduced-order and suboptimal filtering (see Chapter 5);
- reduction of round-off errors and computational expenses;
- analysis and repair of covariance matrices (see next subsection).

3.5.2 ▪ Filter divergence and covariance matrices

If the a priori statistical information is not well specified, the filter might underestimate the variances of the state errors, $\mathbf{e}_k^a$. Too much confidence is put in the state estimation and too little confidence is put in the information contained in the observations. The effect of the analysis is minimized, and the gain becomes too small. In the most extreme case, observations are simply rejected. This is known as *filter divergence*, where the filter seems to behave well, with low predicted analysis error variance, but where the analysis is in fact drifting away from the reality.

Very often filter divergence is easy to diagnose:

- state error variances are small,
- the time sequence of innovations is biased, and
- the Kalman gains tend to zero as time increases.

It is thus important to monitor the innovation sequence and check that it is "white," i.e., unbiased and normally distributed. If this is not the case, then some of your assumptions are not valid.

There are a few rules to follow to avoid divergence:

- Do not underestimate model errors; rather, overestimate them.
- If possible, it is better to use an adaptive scheme to tune model errors by estimating them on the fly using the innovations.
- Give more weight to recent data, thus reducing the filter's memory of older data and forcing the data into the KF.
- Place some empirical, relevant lower bound on the Kalman gains.

3.5.3 ▪ Problem size and optimal interpolation

The straightforward application of the KF implies the "propagation" of an $n \times n$ covariance matrix at each time step. This can result in a very large problem in terms of computations and storage. If the state has a spatial dimension of 10^7 (which is not uncommon in large-scale geophysical and other simulations), then the covariance matrices will be of order 10^{14}, which will exceed the resources of most available computer installations. To overcome this, we must resort to various suboptimal schemes (an example of which is detailed below) or switch to ensemble approaches (see Chapters 6 and 7).

If the computational cost of propagating $\mathbf{P}^{\mathrm{a}}_{k+1}$ is an issue, we can use a *frozen covariance matrix*,

$$\mathbf{P}^{\mathrm{a}}_{k} = \mathbf{P}^{\mathrm{b}}, \quad k = 1, \ldots, n.$$

This defines the OI class of methods. Under this simplifying hypothesis, the two-step assimilation cycle defined above becomes the following:

1. **Forecast:**

$$\mathbf{x}^{\mathrm{f}}_{k+1} = \mathbf{M}_{k+1}\mathbf{x}^{\mathrm{a}}_{k},$$
$$\mathbf{P}^{\mathrm{f}}_{k+1} = \mathbf{P}^{\mathrm{b}}.$$

2. **Analysis:**

$$\mathbf{K}_{k+1} = \mathbf{P}^{\mathrm{b}}\mathbf{H}^{T}\left(\mathbf{H}\mathbf{P}^{\mathrm{b}}\mathbf{H}^{T} + \mathbf{R}_{k+1}\right)^{-1},$$
$$\mathbf{x}^{\mathrm{a}}_{k+1} = \mathbf{x}^{\mathrm{f}}_{k+1} + \mathbf{K}_{k+1}\left(\mathbf{y}_{k+1} - \mathbf{H}\mathbf{x}^{\mathrm{f}}_{k+1}\right),$$
$$\mathbf{P}^{\mathrm{a}}_{k+1} = \mathbf{P}^{\mathrm{b}}.$$

There are at least two ways to compute the static covariance matrix $\mathbf{P}^{\mathrm{b}}$. The first is an analytical formulation,

$$\mathbf{P}^{\mathrm{b}} = \mathbf{D}^{1/2}\mathbf{C}\mathbf{D}^{1/2},$$

where $\mathbf{D}$ is a diagonal matrix of variances and $\mathbf{C}$ is a correlation matrix that can be defined, for example, as

$$C_{ij} = \left(1 + ah + \frac{1}{3}a^2h^2\right)e^{-ah},$$

where a is a tuneable parameter, h is the grid size, and the exponential function provides a local spatial dependence effect that often corresponds well to the physics. The second approach uses an ensemble of N_{e} snapshots of the state vector taken from a model free run, from which we compute the first and second statistical moments as follows:

$$\mathbf{x}^{\mathrm{b}} = \frac{1}{N_{\mathrm{e}}}\sum_{l=1}^{N_{\mathrm{e}}}\mathbf{x}_{l},$$

$$\mathbf{P}^{\mathrm{b}} = \frac{1}{N_{\mathrm{e}} - 1}\sum_{l=1}^{N_{\mathrm{e}}}\left(\mathbf{x}_{l} - \mathbf{x}^{\mathrm{b}}\right)\left(\mathbf{x}_{l} - \mathbf{x}^{\mathrm{b}}\right)^{\mathrm{T}}.$$

The static approach is more suited to successive assimilation cycles that are separated by a long enough time delay so that the corresponding dynamical states are sufficiently decorrelated.

Other methods are detailed in the sections on reduced methods—see Chapter 5.

3.5.4 ▪ Evolution of the state error covariance matrix

In principle, equation (3.19) generates a symmetric matrix. In practice, this may not be the case, and numerical truncation errors may lead to an asymmetric covariance matrix and a subsequent collapse of the filter. A remedy is to add an extra step to enforce symmetry, such as

$$\mathbf{P}^{\mathrm{f}}_{k+1} = \frac{1}{2}\left(\mathbf{P}^{\mathrm{f}}_{k+1} + (\mathbf{P}^{\mathrm{f}}_{k+1})^{\mathrm{T}}\right),$$

or a square root decomposition—see Chapter 5.

3.6 ▪ Nonlinearities and extensions of the KF

In real-life problems, we are most often confronted with a nonlinear process and/or a nonlinear measurement operator. Our dynamic system now takes the more general form

$$\begin{aligned}\mathbf{x}_{k+1} &= M_{k+1}(\mathbf{x}_k) + \mathbf{w}_k,\\ \mathbf{y}_k &= H_k(\mathbf{x}_k) + \mathbf{v}_k,\end{aligned}$$

where M_k now represents a nonlinear function of the state at time step k and H_k represents the nonlinear observation operator.

To deal with these nonlinearities, one approach is to linearize about the current mean and covariance, which is called the *extended Kalman filter* (EKF). This approach and its variants are presented in Chapter 6.

As previously mentioned, the KF is only optimal in the case of Gaussian statistics and linear operators, in which case the first two moments (the mean and the covariances) suffice to describe the PDF entering the estimation problem. Practitioners report that the linearized extension to nonlinear problems, the EKF, only works for moderate deviations from linearity and Gaussianity. The ensemble Kalman filter (EnKF) [Evensen, 2009] is a method that has been designed to deal with nonlinearities and non-Gaussian statistics, whereby the PDF is described by an ensemble of N_{e} time-dependent states $\mathbf{x}_{k,e}$. This method is presented in detail in Chapter 6. The appeal of this approach is its conceptual simplicity, the fact that it does not require any TLM or adjoint model (see Chapter 2), and the fact that it is extremely well suited to parallel programming paradigms, such as MPI [Gropp et al., 2014].

What happens if both the models are nonlinear and the PDFs are non-Gaussian? The KF and its extensions are no longer optimal and, more important, can easily fail the estimation process. Another approach must be used. A promising candidate is the *particle filter,* which is described below. The particle filter (see [Doucet and Johansen, 2011] and references therein) works sequentially in the spirit of the KF, but unlike the latter, it handles an ensemble of states (the particles) whose distribution approximates the PDF of the true state. Bayes' rule (3.5) and the marginalization formula (3.1) are explicitly used in the estimation process. The linear and Gaussian hypotheses can then be ruled out, in theory. In practice, though, the particle filter cannot yet be applied to very high dimensional systems (this is often referred to as "the curse of dimensionality").

Finally, there is a new class of hybrid methods, called *ensemble variational methods,* that attempt to combine variational and ensemble approaches—see Chapter 7 for a detailed presentation. The aim is to seek compromises to exploit the best aspects of (4D) variational and ensemble DA algorithms.

For further details of all these extensions, the reader should consult the advanced methods section (Part II) and the above references.

3.7 ▪ Particle filters for geophysical applications

Can we actually design a filtering numerical algorithm that converges to the Bayesian solution? Such a numerical approach would typically belong to the class of *sequential Monte Carlo* methods. That is to say, a PDF is represented by a discrete sample of the targeted PDF. Rather than trying to compute the exact solution of the Bayesian filtering equations, the transformations of such filtering (Bayes' rule for the analysis; model propagation for the forecast) are applied to the members of the sample. The statistical properties of the sample, such as the moments, are meant to be those of the targeted PDF. Obviously this sampling strategy can only be exact in the asymptotic limit, that is, in the limit where the number of members (or particles) goes to infinity.

This is the focus of a large body of applied mathematics that led to the design of many very successful Monte Carlo type methods [see, for instance, Doucet et al., 2001]. However, they have mostly been applied to very low dimensional systems (only a few dimensions). Their efficiency for high-dimensional models has been studied more recently, in particular thanks to a strong interest in these approaches in the geosciences. In the following, we give a brief biased overview of the subject as seen by the geosciences DA community.

3.7.1 ▪ Sequential Monte Carlo

The most popular and simple algorithm of Monte Carlo type that solves the Bayesian filtering equations is called the *bootstrap particle filter* [Gordon et al., 1993]. Its description follows.

3.7.1.1 ▪ Sampling

Let us consider a sample of particles $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$. The related PDF at time t_k is $p_k(\mathbf{x})$, where $p_k(\mathbf{x}) \simeq \sum_{i=1}^{m} \omega_i^k \delta(\mathbf{x} - \mathbf{x}_k^i)$, δ is the Dirac mass, and the sum is meant to be an approximation of the exact density that the sample emulates. A positive scalar, ω_k^i, weights the importance of particle i within the ensemble. At this stage, we assume that the weights, ω_i^k, are uniform and $\omega_i^k = 1/m$.

3.7.1.2 ▪ Forecast

At the forecast step, the particles are propagated by the model without approximation, $p_{k+1}(\mathbf{x}) \simeq \sum_{i=1}^{m} \omega_k^i \delta(\mathbf{x} - \mathbf{x}_{k+1}^i)$, with $\mathbf{x}_{k+1}^i = \mathcal{M}_{k+1}(\mathbf{x}_k)$. A stochastic noise can be optionally added to the dynamics of each particle (see below).

3.7.1.3 ▪ Analysis

The analysis step of the particle filter is extremely simple and elegant. The rigorous implementation of Bayes' rule ascribes to each particle a statistical weight that corresponds to the likelihood of the particle given the data. The weight of each particle is updated according to (see Figure 3.8)

$$\omega_{k+1}^{\mathrm{a},i} \propto \omega_{k+1}^{\mathrm{f},i} p(\mathbf{y}_{k+1} | \mathbf{x}_{k+1}^i). \tag{3.23}$$

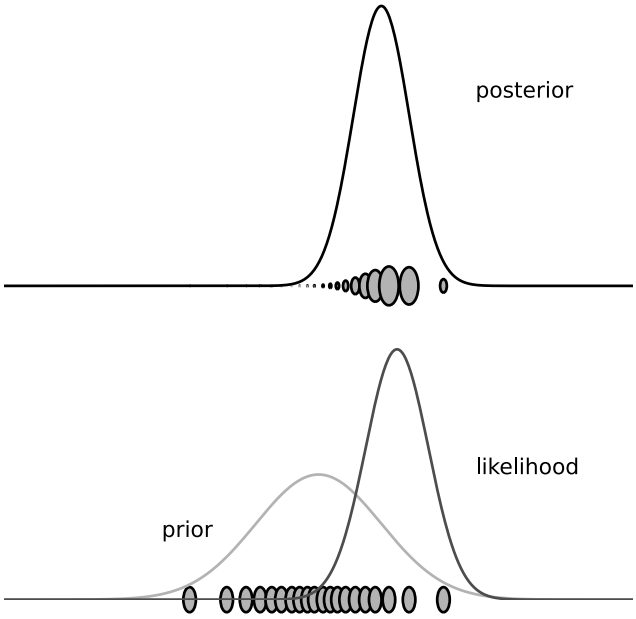

Figure 3.8. *Analysis of the particle filter. The initial ensemble of particles is sampled from a normal prior, with equal weights (bottom). Given an observation with Gaussian noise and the relative state likelihood (bottom), the particle filter analysis ascribes a weight to each particle, which is proportional to the likelihood of the particle given the observation (top). The major axis of the ellipses, representing the particles, is proportional to the particle weight.*

It is remarkable that the analysis is carried out with only a few multiplications. It does not involve inverting any system or matrix, as opposed, for instance, to the KF.

3.7.1.4 ▪ Resampling

Unfortunately, these normalized statistical weights have a potentially large amplitude of fluctuation. Even worse, as sequential filtering progresses, one particle (one trajectory of the model) will stand out from the others. Its weight will largely dominate the others ($\omega_i \lesssim 1$), while the other weights will vanish. Then the particle filter becomes very inefficient as an estimating tool since it has lost its variability. This phenomenon is called *degeneracy* of the particle filter [Kong et al., 1994]. An example of such degeneracy is given in Figure 3.9, where the statistical properties of the biggest weight are studied on an extensive meteorological toy model of 40 and 80 variables. In a degenerate case, the maximum weight will often reach 1 or close to 1, whereas in a balanced case, values very close to 1 will be less frequent.

One way to mitigate this phenomenon is to resample the particles by redrawing a sample with uniform weights from the degenerate distribution. After resampling, all particles have the same weight: $\omega_k^i = 1/m$.

The particle filter is very efficient for highly nonlinear models but with low dimensionality. Unfortunately, it is not suited for DA systems with models of high dimension, as soon as the dimension exceeds, say, about 10. Avoiding degeneracy requires a great number of particles. This number typically increases exponentially with

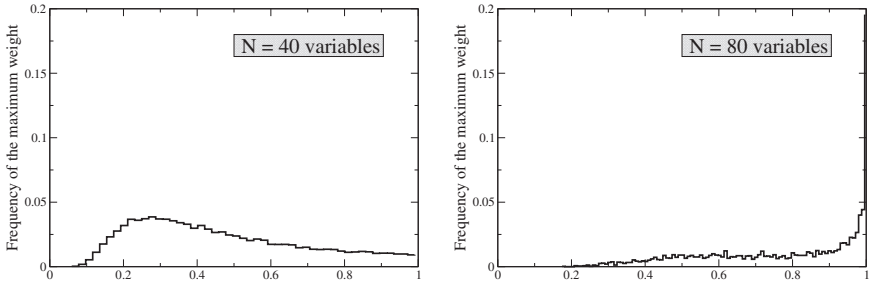

Figure 3.9. *On the left: statistical distribution of the maximal weight of the particle bootstrap filter in a balanced case. The physical system is a Lorenz-*95 *model with* 40 *variables [Lorenz and Emmanuel,* 1998*]. On the right: the same particle filter is applied to a Lorenz-*95 *low-order model, but with* 80 *variables. The maximal weight clearly degenerates with a peak close to* 1.

the system state space dimension. This is because the support of the prior PDF overlaps exponentially less with the support of the likelihood as the dimension of the state space of the systems increases. This is known as the *curse of dimensionality*.

For the forecast step, it could also be crucial to introduce stochastic perturbations of the states. Indeed, the ensemble will become impoverished with the many resamplings that it has to undergo. To enrich the sample, it is necessary to stochastically perturb the states of the system.

3.7.2 ▪ Application in the geosciences

The applicability of particle filters to high-dimensional models has been investigated in the geosciences [van Leeuwen, 2009; Bocquet et al., 2010]. The impact of the curse of dimensionality has been quantitatively studied in Snyder et al. [2008]. It has been shown, on a heuristic basis, that the number of particles m required to efficiently track the system must scale like the variance of the log-likelihood,

$$\ln(m) \propto \mathrm{Var}[\ln(p(\mathbf{y}|\mathbf{x}))], \tag{3.24}$$

which usually scales like the size of the system for typical geophysical problems. It is known [see, for instance, MacKay, 2003] that using an importance proposal to guide the particles toward regions of high probability will not change this trend, albeit with a smaller proportionality factor in (3.24). Snyder et al. [2015] confirmed this and gave bounds to the optimal proposal for particle filters that use an importance proposal leading to a minimal variance in the weights. They conclude again on the exponential dependence of the effective ensemble size with the problem dimension.

When smoothing is combined with a particle filter (which becomes a particle smoother) over a DA window, alternative and more efficient particle filters can be designed, such as the implicit particle filter [Morzfeld et al., 2012].

Particle filters can nevertheless be useful for high-dimensional models if the significant degree of nonlinearity is confined to a small subspace of the state space. For instance, in Lagrangian DA, the errors on the location of moving observation platforms have significantly non-Gaussian statistics. In this case, these degrees of freedom can be addressed with a particle filter, while the rest is controlled by an EnKF, which is practical for high-dimensional models [Slivinski et al., 2015].

If we drop the assumption that a particle filter should have the proper Bayesian asymptotic limit, it becomes possible to design nonlinear filters for DA with

high-dimensional models such as the equal-weight particle filter (see [Ades and van Leeuwen, 2015] and references therein).

Finally, if the system cannot be split, then a solution to implement a particle filter in high dimension could come from localization, just as with the EnKF (Chapter 6). This was proven to be more difficult because locally updated particles cannot easily be glued together into global particles. However, an ensemble transform representation that has been built for the EnKF [Bishop et al., 2001] is better suited to ensure a smoother gluing of the local updates [Reich, 2013]. An astute merging of the particles has been shown to yield a local particle filter that could outperform the EnKF in specific regimes with a moderate number of particles [Poterjoy, 2016].

3.8 ▪ Examples

In this section we present a number of examples of special cases of the KF—both analytical and numerical. Though they may seem overly simple, the intention is that you, the user, gain the best possible feeling and intuition regarding the actual operation of the filter. This understanding is essential for more complex cases, such as those presented in the advanced methods and applications chapters.

Example 3.31. *Case without observations.* Here, the observation matrix $\mathbf{H}_k = 0$ and thus $\mathbf{K}_k = 0$ as well. Hence the KF equations (3.18)–(3.22) reduce to

$$\begin{aligned}\mathbf{x}^{\mathrm{f}}_{k+1} &= \mathbf{M}_{k+1}\mathbf{x}^{\mathrm{a}}_k,\\ \mathbf{P}^{\mathrm{f}}_{k+1} &= \mathbf{M}_{k+1}\mathbf{P}^{\mathrm{a}}_k\mathbf{M}^T_{k+1} + \mathbf{Q}_{k+1},\end{aligned}$$

and

$$\begin{aligned}\mathbf{K}_{k+1} &= 0,\\ \mathbf{x}^{\mathrm{a}}_{k+1} &= \mathbf{x}^{\mathrm{f}}_{k+1},\\ \mathbf{P}^{\mathrm{a}}_{k+1} &= \mathbf{P}^{\mathrm{f}}_{k+1}.\end{aligned}$$

Thus, we can completely eliminate the analysis stage of the algorithm to obtain

$$\begin{aligned}\mathbf{x}^{\mathrm{f}}_{k+1} &= \mathbf{M}_{k+1}\mathbf{x}^{\mathrm{f}}_k,\\ \mathbf{P}^{\mathrm{f}}_{k+1} &= \mathbf{M}_{k+1}\mathbf{P}^{\mathrm{f}}_k\mathbf{M}^T_{k+1} + \mathbf{Q}_{k+1},\end{aligned}$$

initialized by

$$\begin{aligned}\mathbf{x}^{\mathrm{f}}_0 &= \mathbf{x}_0,\\ \mathbf{P}^{\mathrm{f}}_0 &= \mathbf{P}_0.\end{aligned}$$

The model then runs without any input of data, and if the dynamics are neutral or unstable, the forecast error will grow without limit. For example, in a typical NWP assimilation cycle, where observations are obtained every 6 hours, the model runs for 6 hours without data. During this period, the forecast error grows and is damped only when the data arrives, thus giving rise to the characteristic "sawtooth" pattern of error variance evolution. ■

Example 3.32. *Perfect observations at all grid points.* In the case of perfect observations, the observation error covariance matrix $\mathbf{R}_k = 0$ and the observation operator $\mathbf{H}$ is the identity. Hence the KF equations (3.18)–(3.22) reduce to

$$\mathbf{x}^{\mathrm{f}}_{k+1} = \mathbf{M}_{k+1}\mathbf{x}^{\mathrm{a}}_{k},$$
$$\mathbf{P}^{\mathrm{f}}_{k+1} = \mathbf{M}_{k+1}\mathbf{P}^{\mathrm{a}}_{k}\mathbf{M}^{T}_{k+1} + \mathbf{Q}_{k+1},$$

and

$$\mathbf{K}_{k+1} = \mathbf{P}^{\mathrm{f}}_{k+1}\mathbf{H}^{T}\left(\mathbf{H}\mathbf{P}^{\mathrm{f}}_{k+1}\mathbf{H}^{T}\right)^{-1} = \mathbf{I},$$
$$\mathbf{x}^{\mathrm{a}}_{k+1} = \mathbf{x}^{\mathrm{f}}_{k+1} + \left(\mathbf{y}_{k+1} - \mathbf{x}^{\mathrm{f}}_{k+1}\right),$$
$$\mathbf{P}^{\mathrm{a}}_{k+1} = \left(\mathbf{I} - \mathbf{K}_{k+1}\mathbf{H}\right)\mathbf{P}^{\mathrm{f}}_{k+1} = 0.$$

This is obviously another case of ideal observations, and we can once again completely eliminate the analysis stage to obtain

$$\mathbf{x}^{\mathrm{f}}_{k+1} = \mathbf{M}_{k+1}\mathbf{x}^{\mathrm{f}}_{k},$$
$$\mathbf{P}^{\mathrm{f}}_{k+1} = \mathbf{Q}_{k+1},$$

with initial conditions

$$\mathbf{x}^{\mathrm{f}}_{0} = \mathbf{y}_{0},$$
$$\mathbf{P}^{\mathrm{f}}_{0} = 0.$$

Since $\mathbf{R}$ is in fact the sum of measurement and representation errors, $\mathbf{R} = 0$ implies that the only scales that are observed are those resolved by the model. The forecast is thus an integration of the observed state, and the forecast error reduces to the model error. ∎

Example 3.33. *Scalar case.* As in Section 2.4.5, let us consider the same scalar example, but this time apply the KF to it. We take the simplest linear forecast model,

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\alpha x,$$

with α a known positive constant. We assume the same discrete dynamics considered in (2.49) with a single observation at time step 3.

The stochastic system (3.11)–(3.12) is

$$x^{\mathrm{t}}_{k+1} = M(x^{\mathrm{t}}_{k}) + w_k,$$
$$y_{k+1} = x^{\mathrm{t}}_{k} + v_k,$$

where $w_k \sim \mathcal{N}(0, \sigma^2_Q)$, $v_k \sim \mathcal{N}(0, \sigma^2_R)$, and $x^{\mathrm{t}}_0 - x^{\mathrm{b}}_0 \sim \mathcal{N}(0, \sigma^2_B)$. The KF steps are as follows:

Forecast:

$$x^{\mathrm{f}}_{k+1} = M(x^{\mathrm{a}}_{k}) = \gamma x_k,$$
$$P^{\mathrm{f}}_{k+1} = \gamma^2 P^{\mathrm{a}}_{k} + \sigma^2_Q.$$

Analysis:

$$K_{k+1} = P^{\mathrm{f}}_{k+1} H \left(H^2 P^{\mathrm{f}}_{k+1} + \sigma_R^2 \right)^{-1},$$
$$x^{\mathrm{a}}_{k+1} = x^{\mathrm{f}}_{k+1} + K_{k+1}(x^{\mathrm{obs}}_{k+1} - H x^{\mathrm{f}}_{k+1}),$$
$$P^{\mathrm{a}}_{k+1} = (1 - K_{k+1} H) P^{\mathrm{f}}_{k+1} = \left(\frac{1}{P^{\mathrm{f}}_{k+1}} + \frac{1}{\sigma_R^2} \right)^{-1}, \quad H = 1.$$

Initialization:

$$x^{\mathrm{a}}_0 = x^{\mathrm{b}}_0,$$
$$P^{\mathrm{a}}_0 = \sigma_B^2.$$

We start with the initial state, at time step $k = 0$. The initial conditions are as above. The forecast is

$$x^{\mathrm{f}}_1 = M(x^{\mathrm{a}}_0) = \gamma x^{\mathrm{b}}_0,$$
$$P^{\mathrm{f}}_1 = \gamma^2 \sigma_B^2 + \sigma_Q^2.$$

Since there is no observation available, $H = 0$, and the analysis gives

$$K_1 = 0,$$
$$x^{\mathrm{a}}_1 = x^{\mathrm{f}}_1 = \gamma x^{\mathrm{b}}_0,$$
$$P^{\mathrm{a}}_1 = P^{\mathrm{f}}_1 = \gamma^2 \sigma_B^2 + \sigma_Q^2.$$

At the next time step, $k = 1$, and the forecast gives

$$x^{\mathrm{f}}_2 = M(x^{\mathrm{a}}_1) = \gamma^2 x^{\mathrm{b}}_0,$$
$$P^{\mathrm{f}}_2 = \gamma^2 P^{\mathrm{a}}_1 + \sigma_Q^2 = \gamma^4 \sigma_B^2 + (\gamma^2 + 1)\sigma_Q^2.$$

Once again there is no observation available, $H = 0$, and the analysis yields

$$K_2 = 0,$$
$$x^{\mathrm{a}}_2 = x^{\mathrm{f}}_2 = \gamma^2 x^{\mathrm{b}}_0,$$
$$P^{\mathrm{a}}_2 = P^{\mathrm{f}}_2 = \gamma^4 \sigma_B^2 + (\gamma^2 + 1)\sigma_Q^2.$$

Moving on to $k = 2$, we have the new forecast:

$$x^{\mathrm{f}}_3 = M(x^{\mathrm{a}}_2) = \gamma^3 x^{\mathrm{b}}_0,$$
$$P^{\mathrm{f}}_3 = \gamma^2 P^{\mathrm{a}}_2 + \sigma_Q^2 = \gamma^6 \sigma_B^2 + (\gamma^4 + \gamma^2 + 1)\sigma_Q^2.$$

Now there is an observation, x^{o}_3, available, so $H = 1$, and the analysis is

$$K_3 = P^{\mathrm{f}}_3 \left(P^{\mathrm{f}}_3 + \sigma_R^2 \right)^{-1},$$
$$x^{\mathrm{a}}_3 = x^{\mathrm{f}}_3 + K_3(x^{\mathrm{o}}_3 - x^{\mathrm{f}}_3),$$
$$P^{\mathrm{a}}_3 = (1 - K_3) P^{\mathrm{f}}_3.$$

Substituting and simplifying, we find

$$x_3^{\mathrm{a}} = \gamma^3 x_0^{\mathrm{b}} + \frac{\gamma^6\sigma_B^2 + (\gamma^4+\gamma^2+1)\sigma_Q^2}{\sigma_R^2 + \gamma^6\sigma_B^2 + (\gamma^4+\gamma^2+1)\sigma_Q^2}\left(x_3^{\mathrm{o}} - \gamma^3 x_0^{\mathrm{b}}\right). \tag{3.25}$$

Case 1: Assume we have a perfect model. Then $\sigma_Q^2 = 0$ and the KF state (3.25) becomes

$$x_3^{\mathrm{a}} = \gamma^3 x_0^{\mathrm{b}} + \frac{\gamma^6\sigma_B^2}{\sigma_R^2 + \gamma^6\sigma_B^2}\left(x_3^{\mathrm{o}} - \gamma^3 x_0^{\mathrm{b}}\right),$$

which is precisely the 4D-Var expression (2.51) obtained before.

Case 2: When the parameter α tends to zero, then γ tends to one, the model is stationary, and the KF state (3.25) becomes

$$x_3^{\mathrm{a}} = x_0^{\mathrm{b}} + \frac{\sigma_B^2 + 3\sigma_Q^2}{\sigma_R^2 + \sigma_B^2 + 3\sigma_Q^2}\left(x_3^{\mathrm{o}} - x_0^{\mathrm{b}}\right),$$

which, when $\sigma_Q^2 = 0$, reduces to the 3D-Var solution,

$$x_3^{\mathrm{a}} = x_0^{\mathrm{b}} + \frac{\sigma_B^2}{\sigma_R^2 + \sigma_B^2}\left(x_3^{\mathrm{o}} - x_0^{\mathrm{b}}\right),$$

that was obtained before in (2.52).

Case 3: When α tends to infinity, then γ goes to zero, and we are in the case where there is no longer any memory with

$$x_3^{\mathrm{a}} = \frac{\sigma_Q^2}{\sigma_R^2 + \sigma_Q^2} x_3^{\mathrm{o}}.$$

Then, if the model is perfect, $\sigma_Q^2 = 0$ and $x_3^{\mathrm{a}} = 0$. If the observation is perfect, $\sigma_R^2 = 0$ and $x_3^{\mathrm{a}} = x_3^{\mathrm{o}}$.

This example shows the complete chain, from the KF solution through the 4D-Var and finally reaching the 3D-Var solution. We hope that this clarifies the relationship between the three and demonstrates why the KF provides the most general solution possible. ∎

Example 3.34. *Brownian motion.* Here we compute a numerical application of the scalar case seen above in Example 3.33. We have the following state and measurement equations:

$$x_{k+1} = x_k + w_k,$$
$$y_{k+1} = x_k + v_k,$$

where the dynamic transition matrix $M_k = 1$ and the observation operator $H = 1$. Let us suppose constant error variances of $Q_k = 1$ and $R_k = 0.25$ for the process and measurement errors, respectively. Here the KF equations (3.18)–(3.22) reduce to

$$x_{k+1}^{\mathrm{f}} = x_k^{\mathrm{a}},$$
$$P_{k+1}^{\mathrm{f}} = P_k^{\mathrm{a}} + 1,$$

and

$$\begin{aligned} K_{k+1} &= P^{\mathrm{f}}_{k+1}\left(P^{\mathrm{f}}_{k+1}+0.25\right)^{-1}, \\ x^{\mathrm{a}}_{k+1} &= x^{\mathrm{f}}_{k+1}+K_{k+1}\left(y_{k+1}-x^{\mathrm{f}}_{k+1}\right), \\ P^{\mathrm{a}}_{k+1} &= \left(I-K_{k+1}\right)P^{\mathrm{f}}_{k+1}. \end{aligned}$$

By substituting for P^{f}_{k+1} from the forecast equation, we can rewrite the Kalman gain in terms of P^{a}_{k} as

$$K_{k+1} = \frac{P^{\mathrm{a}}_{k}+1}{P^{\mathrm{a}}_{k}+1.25},$$

and we obtain the update for the error variance:

$$P^{\mathrm{a}}_{k+1} = \frac{P^{\mathrm{a}}_{k}+1}{4P^{\mathrm{a}}_{k}+5}.$$

Plugging into the analysis equation, we now have the complete update:

$$\begin{aligned} x^{\mathrm{a}}_{k+1} &= x^{\mathrm{a}}_{k}+\frac{P^{\mathrm{a}}_{k}+1}{P^{\mathrm{a}}_{k}+1.25}\left(y_{k+1}-x^{\mathrm{a}}_{k}\right), \\ P^{\mathrm{a}}_{k+1} &= \frac{P^{\mathrm{a}}_{k}+1}{4P^{\mathrm{a}}_{k}+5}. \end{aligned}$$

Let us now, manually, perform a couple of iterations. Taking as initial conditions

$$x^{\mathrm{a}}_{0}=0, \quad P^{\mathrm{a}}_{0}=0,$$

we readily compute, for $k=0$,

$$\begin{aligned} K_1 &= \frac{1}{1.25}=0.8, \\ x^{\mathrm{a}}_{1} &= 0+K_1(y_1-0)=0.8y_1, \\ P^{\mathrm{a}}_{1} &= \frac{1}{5}=0.2. \end{aligned}$$

Then for $k=1$,

$$\begin{aligned} K_2 &= \frac{0.2+1}{0.2+1.25}\approx 0.8276, \\ x^{\mathrm{a}}_{2} &= 0.8y_1+K_2(y_2-0.8y_1)\approx 0.138y_1+0.828y_2, \\ P^{\mathrm{a}}_{2} &= \frac{0.2+1}{0.8+5}=\frac{6}{29}\approx 0.207. \end{aligned}$$

One more step for $k=2$ gives

$$\begin{aligned} K_3 &= \frac{6/29+1}{6/29+1.25}\approx 0.8284, \\ x^{\mathrm{a}}_{3} &= 0.138y_1+0.828y_2+K_3(y_3-0.138y_1-0.828y_2)\approx 0.024y_1+0.142y_2+0.828y_3, \\ P^{\mathrm{a}}_{3} &= \frac{6/29+1}{24/29+5}\approx 0.207. \end{aligned}$$

Let us see what happens in the limit, $k \to \infty$. We observe that $P_{k+1} \approx P_k$; thus

$$P_\infty^{\mathrm{a}} = \frac{P_\infty^{\mathrm{a}} + 1}{4P_\infty^{\mathrm{a}} + 5},$$

which is a quadratic equation for P_∞^{a}, whose solutions are

$$P_\infty^{\mathrm{a}} = \frac{1}{2}\left(-1 \pm \sqrt{2}\right).$$

The positive definite solution is

$$P_\infty^{\mathrm{a}} = \frac{1}{2}\left(-1 + \sqrt{2}\right) \approx 0.2071,$$

and hence

$$K_\infty = \frac{2+2\sqrt{2}}{3+2\sqrt{2}} \approx 0.8284.$$

We observe in this case that the KF tends toward a steady-state filter after only two steps. The reasons for this rapid convergence are that the dynamics are neutral and that the observation error covariance is relatively small when compared to the process error, $R \ll Q$, which means that the observations are relatively precise compared to the model error. In addition, the state (being scalar) is completely observed whenever an observation is available.

In conclusion, the combination of dense, precise observations with steady, linear dynamics will always lead to a stable filter. ∎

Example 3.35. *Estimation of a random constant.* In this simple numerical example, let us attempt to estimate a scalar random constant, for example, a voltage. Let us assume that we have the ability to take measurements of the constant, but that the measurements are corrupted by a 0.1 volt root mean square (rms) white measurement noise (e.g., our analog-to-digital converter is not very accurate). In this example, our process is governed by the state equation

$$x_k = Mx_{k-1} + w_k = x_{k-1} + w_k$$

and the measurement equation

$$y_k = Hx_k + v_k = x_k + v_k.$$

The state, being constant, does not change from step to step, so $M = I$. Our noisy measurement is of the state directly, so $H = 1$. We are in fact in the same Brownian motion context as the previous example.

The time update (forecast) equations are

$$\begin{aligned} x_{k+1}^{\mathrm{f}} &= x_k^{\mathrm{a}}, \\ P_{k+1}^{\mathrm{f}} &= P_k^{\mathrm{a}} + Q, \end{aligned}$$

and the measurement update (analysis) equations are

$$\begin{aligned} K_{k+1} &= P_{k+1}^{\mathrm{f}}(P_{k+1}^{\mathrm{f}} + R)^{-1}, \\ x_{k+1}^{\mathrm{a}} &= x_{k+1}^{\mathrm{f}} + K_{k+1}(y_{k+1} - x_{k+1}^{\mathrm{f}}), \\ P_{k+1}^{\mathrm{a}} &= (1 - K_{k+1})P_{k+1}^{\mathrm{f}}. \end{aligned}$$

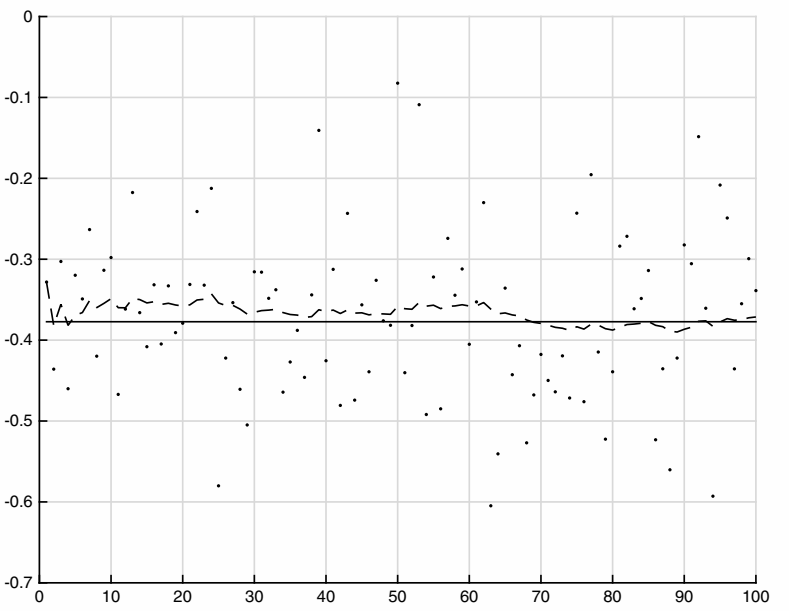

Figure 3.10. *Estimating a constant—simulation with $R = 0.01$. True value (solid), measurements (dots), KF estimation (dashed). The x-axis denotes time; the y-axis denotes the state variable.*

Initialization. Presuming a very small process variance, we let $Q = 1.e-5$. We could certainly let $Q = 0$, but assuming a small but nonzero value gives us more flexibility in "tuning" the filter, as we will demonstrate below. Let's assume that from experience we know that the true value of the random constant has a standard Gaussian probability distribution, so we will "seed" our filter with the guess that the constant is 0. In other words, before starting, we let $x_0 = 0$. Similarly, we need to choose an initial value for P_k^a; call it P_0. If we were absolutely certain that our initial state estimate was correct, we would let $P_0 = 0$. However, given the uncertainty in our initial estimate, x_0, choosing $P_0 = 0$ would cause the filter to initially and always believe that $x_k^a = 0$. As it turns out, the alternative choice is not critical. We could choose almost any $P_0 \neq 0$ and the filter would eventually converge. We will start our filter with $P_0 = 1$.

Simulations. To begin with, we randomly chose a scalar constant $y = -0.37727$. We then simulated 100 distinct measurements that had an error normally distributed around zero with a standard deviation of 0.1 (remember we presumed that the measurements are corrupted by a 0.1 volt rms white measurement noise).

In the first simulation we fixed the measurement variance at $R = (0.1)^2 = 0.01$. Because this is the "true" measurement error variance, we would expect the "best" performance in terms of balancing responsiveness and estimate variance. This will become more evident in the second and third simulations. Figure 3.10 depicts the results of this first simulation. The true value of the random constant, $x = -0.37727$, is given by the solid line, the noisy measurements by the dots, and the filter estimate by the remaining dashed curve.

In Figures 3.11 and 3.12 we can see what happens when the measurement error variance, R, is increased or decreased by a factor of 100. In Figure 3.11, the filter was told

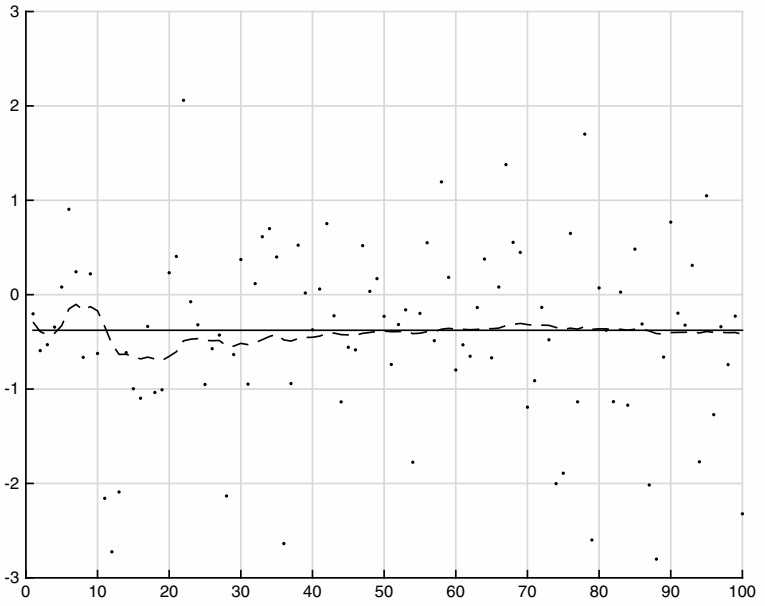

Figure 3.11. *Estimating a constant—simulation with $R = 1$. True value (solid), measurements (dots), KF estimation (dashed). The x-axis denotes time; the y-axis denotes the state variable.*

that the measurement variance was 100 times as great (i.e., $R = 1$), so it was "slower" to believe the measurements. In Figure 3.12, the filter was told that the measurement variance was 1/100th the size (i.e., $R = 0.0001$), so it was very "quick" to believe the noisy measurements.

While the estimation of a constant is relatively straightforward, this example clearly demonstrates the workings of the KF. In Figure 3.11 in particular the Kalman "filtering" is evident, as the estimate appears considerably smoother than the noisy measurements. We observe the speed of convergence of the variance in Figure 3.13.

Here is the MATLAB code used to perform the simulations.

```
% SCALAR EXAMPLE (estimate a constant):
%
% Define the system as a constant of  -0.37727 volts:
clear s
s.x =  -0.37727;
s.A = 1;
% Define a process noise:
s.Q = 0.00001; % variance
% Define the voltmeter to measure the voltage itself:
s.H = 1;
% Define a measurement error:
s.R   = 0.01^2; % variance, hence stdev^2
Rstd  = sqrt(s.R); % random measurement noise stdev
% Specify an initial state:
s.x =  -0.37727;
s.P = 1;
% Generate random voltages and perform the filter operation.
tru=[]; % true voltage
for t=1:100
```

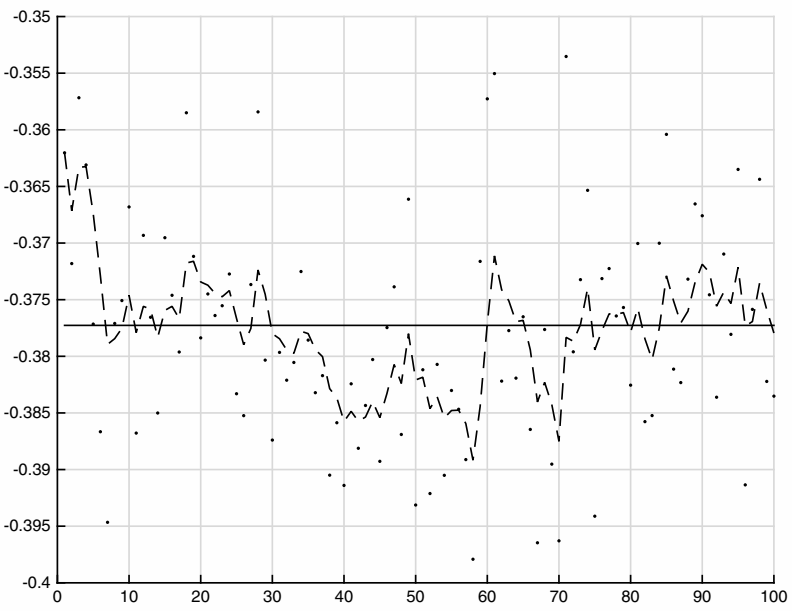

Figure 3.12. *Estimating a constant—simulation with $R = 0.0001$. True value (solid), measurements (dots), KF estimation (dashed). The x-axis denotes time; the y-axis denotes the state variable.*

```
    tru(end+1) =  -0.37727;
    s(end).y = tru(end) + Rstd*randn; % create a measurement
    s(end+1)=kalmanf(s(end)); % perform a Kalman filter iteration
end
% plot measurement data:
figure, hold on, grid on
hy=plot([s(1:end-1).y],'r.');
% plot a-posteriori state estimates:
hk=plot([s(2:end).x],'b--');
% plot true data
ht=plot(tru,'g-');
%legend([hy hk ht],'observations','Kalman output','true voltage',0)
%title('Estimating a constant')
hold off
% KALMANF - updates a system state vector estimate based upon an
%           observation, using a discrete Kalman filter.
%
% Version 1.1, August 13, 2015
%
% This function is based on the original of Michael C. Kleder
%
% INTRODUCTION
%
% Applying the filter to a basic linear system is actually very
% easy.
% This MATLAB file demonstrates that.
%
% An excellent paper on Kalman filtering at the introductory level,
% without detailing the mathematical underpinnings, is
% "An Introduction to the Kalman Filter"
```

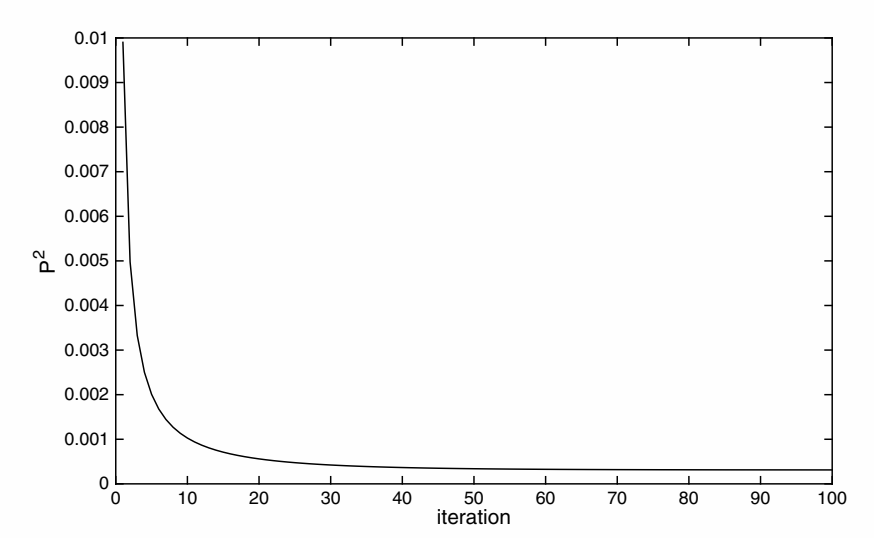

Figure 3.13. *Estimating a constant—convergence of the variance with* $R = 0.01$.

```
% Greg Welch and Gary Bishop, University of North Carolina
% http://www.cs.unc.edu/~welch/kalman/kalmanIntro.html
%
% PURPOSE:
%
% The purpose of each iteration of a Kalman filter is to update
% the estimate of the state vector of a system (and the covariance
% of that vector) based upon the information in a new observation.
% The version of the Kalman filter in this function assumes that
% observations occur at fixed discrete time intervals. Also, this
% function assumes a linear system, meaning that the time evolution
% of the state vector can be calculated by means of a state
% transition matrix.
%
% USAGE:
%
% s = kalmanf(s)
%
% "s" is a "system" struct containing various fields used as input
% and output. The state estimate "x" and its covariance "P" are
% updated by the function. The other fields describe the mechanics
% of the system and are left unchanged. A calling routine may change
% these other fields as needed if state dynamics are time dependent;
% otherwise, they should be left alone after initial values are set.
% The exceptions are the observation vector "y"
%
% SYSTEM DYNAMICS:
%
% The system evolves according to the following difference
% equations, where quantities are further defined below:
%
% x = Ax + w       meaning the state vector x evolves during one
%                  time step by premultiplying by the "state
%                  transition matrix" A. There is also Gaussian
%                  process noise w.
% y = Hx + v       meaning the observation vector y is a linear
```

```
%                   function of the state vector, and this linear
%                   relationship is represented by premultiplication
%                   by "observation matrix" H. There is also
%                   Gaussian measurement noise v.
% where w ~ N(0,Q) meaning w is Gaussian noise with covariance Q
%       v ~ N(0,R) meaning v is Gaussian noise with covariance R
%
% VECTOR VARIABLES:
%
% s.x = state vector estimate. In the input struct, this is the
%       "a priori" state estimate (prior to the addition of the
%       information from the new observation). In the output struct,
%       this is the "a posteriori" state estimate (after the new
%       measurement information is included).
% s.y = observation vector
%
% MATRIX VARIABLES:
%
% s.A = state transition matrix (defaults to identity).
% s.P = covariance of the state vector estimate. In the input
%       struct, this is "a priori," and in the output it is
%       "a posteriori" (required unless autoinitializing as
%       described below).
% s.Q = process noise covariance (defaults to zero).
% s.R = measurement noise covariance (required).
% s.H = observation matrix (defaults to identity).
%
% NORMAL OPERATION:
%
% (1) define all state definition fields: A,H,Q,R
% (2) define initial state estimate: x,P
% (3) obtain observation vector: y
% (4) call the filter to obtain updated state estimate: x,P
% (5) return to step (3) and repeat
%
% INITIALIZATION:
%
% If an initial state estimate is unavailable, it can be obtained
% from the first observation as follows, provided that there are
% the same number of observable variables as state variables.
% This "auto-initialization" is done automatically if s.x is
% absent or NaN.
%
% x = inv(H)*y
% P = inv(H)*R*inv(H')
%
% This is mathematically equivalent to setting the initial state
% estimate covariance to infinity.
%

function s = kalmanf(s)

% set defaults for absent fields:
if ~isfield(s,'y'); error('Observation␣vector␣missing'); end
if ~isfield(s,'x'); s.x=nan*y; end
if ~isfield(s,'P'); s.P=nan; end
if ~isfield(s,'A'); s.A=eye(length(x)); end
if ~isfield(s,'Q'); s.Q=zeros(length(x)); end
if ~isfield(s,'R'); error('Observation␣covariance␣missing'); end
if ~isfield(s,'H'); s.H=eye(length(x)); end

if isnan(s.x)
   % initialize state estimate from first observation
```

```
    if diff(size(s.H))
       error('Observation matrix must be square and invertible ' ...
             'for state auto-initialization')
    end
    s.x = inv(s.H)*s.y;
    s.P = inv(s.H)*s.R*inv(s.H');
else

    % This is the code which implements the discrete Kalman filter:

    % Prediction for state vector and covariance:
    s.x = s.A*s.x;
    s.P = s.A * s.P * s.A' + s.Q;

    % Compute Kalman gain factor:
    K = (s.P)*(s.H')*inv(s.H*s.P*s.H'+s.R);

    % Correction based on observation:
    s.x = s.x + K*(s.y-s.H*s.x);
    s.P = s.P - K*s.H*s.P;

    % Note that the desired result, which is an improved estimate
    % of the system state vector x and its covariance P, was obtained
    % in only five lines of code, once the system was defined. (That's
    % how simple the discrete Kalman filter is to use.)
end
return
```

■

Example 3.36. *Estimation of a linear trajectory.* We now go one step further and introduce some simple linear dynamics into the problem. We consider a train moving at an unknown, constant velocity, and we would like to predict both its position and its velocity from noisy observations of the position only. Note that this example has many concrete applications in automatic pilots, robotics, and other inertial navigation instruments, where KFs are extensively employed. Let us write down the equations.

The equation of motion for the actual position, x, is

$$x(t) = x_0 + st,$$

where x_0 is the initial position and s is the constant speed of the train. For state space notation, we introduce the state vector,

$$\mathbf{x} = \begin{bmatrix} x \\ \dot{x} \end{bmatrix},$$

where x is the position and $\dot{x}$ is the velocity. The state equation can then be written as

$$\mathbf{x}_{k+1} = \mathbf{M}\mathbf{x}_k + \mathbf{w}_k,$$

where we have assumed that the system is perturbed by a white Gaussian noise, $\mathbf{w}$, with known covariance. From the dynamics, we deduce the discrete form of the matrix,

$$\mathbf{M} = \begin{bmatrix} 1 & \mathrm{d}t \\ 0 & 1 \end{bmatrix},$$

where $\mathrm{d}t$ is the time step increment and corresponds to the instants when measurements are taken. The observation is scalar,

$$y_k = \mathbf{H}\mathbf{x}_k + v_k,$$

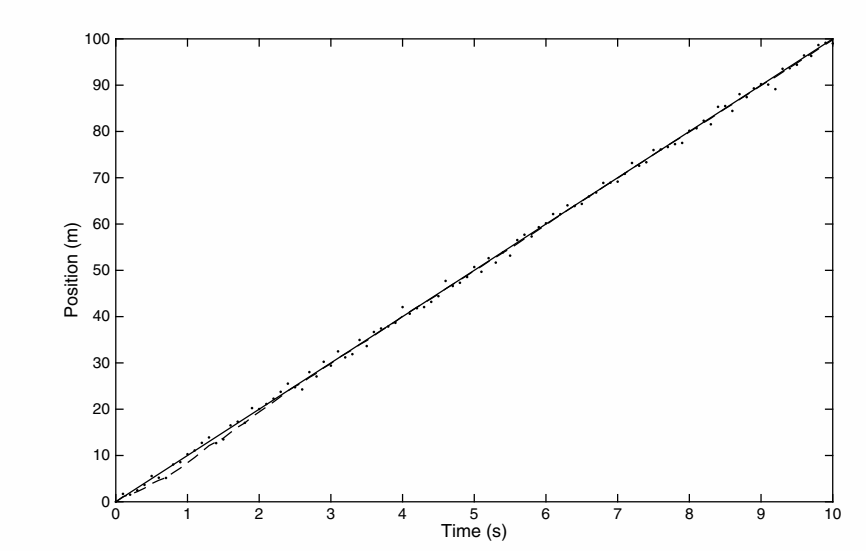

Figure 3.14. *Position estimation for constant-velocity dynamics. True value (solid); measurements (dots); KF estimation (dashed).*

with

$$\mathbf{H} = \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

We initialize our problem with $x_0 = 0$ and $\dot{x}_0 = 0.5\dot{x}^{\mathrm{t}}$, where the (unknown) true velocity $\dot{x}^{\mathrm{t}} = 10\ \mathrm{m\,s^{-1}}$. We assume that measurements are taken every $\mathrm{d}t = 0.1$ s from $t = 0$ until $t = 10$ and that they are subject to a noise with standard deviation of $1\ \mathrm{ms^{-1}}$. We will suppose that the initial error covariance matrix $\mathbf{P}_0 = \mathbf{I}$ and that the process noise is small ($R = 0.0001$). We want to predict the train's position 2 seconds ahead, that is, at $t = 12$ s. We will use the KF, as we need an accurate and smooth estimate for the velocity to predict the train's position in the future.

The numerical computation gives the results and predictions shown in Figures 3.14, 3.15, 3.16, 3.17, and 3.18. The filter does a good job and "steers" a smoother path among the noisy measurements. When compared to an extrapolation, based on a running average for the velocity, the KF clearly outperforms other methods when it comes to the prediction of the position in the future. The convergence of the filter parameters to zero is a sign that the model is adequate and that the data and model are consistent and in good agreement.

The reader is strongly encouraged to modify the code of the previous example to reproduce these results and then to adjust the problem's parameters and observe the effects—as was done in the previous example. ∎

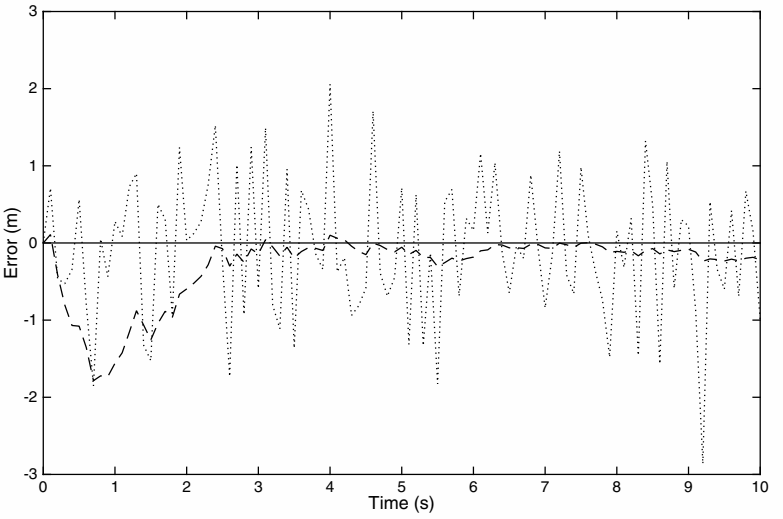

Figure 3.15. *Position estimation errors for constant-velocity dynamics. True value (solid); measurements (dotted); KF estimation (dashed).*

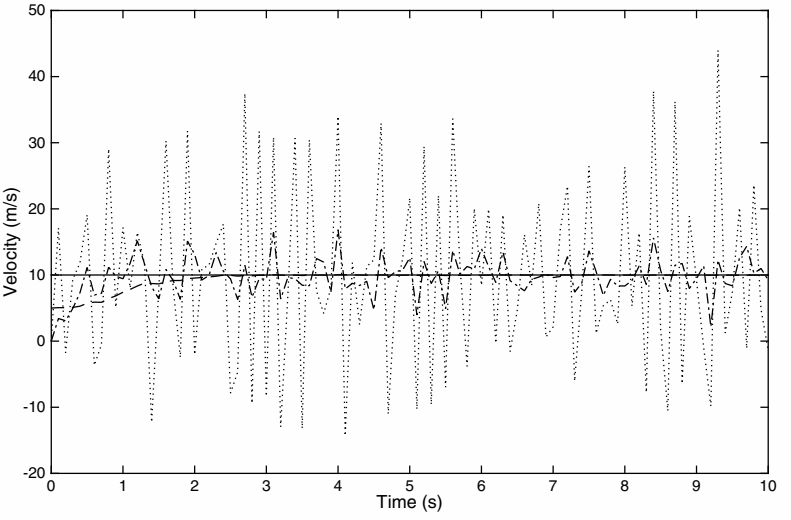

Figure 3.16. *Velocity estimation results for constant-velocity dynamics. True value (solid); from raw measurements (dotted); from running average of raw measurements (dotted-dashed); KF estimation (dash).*

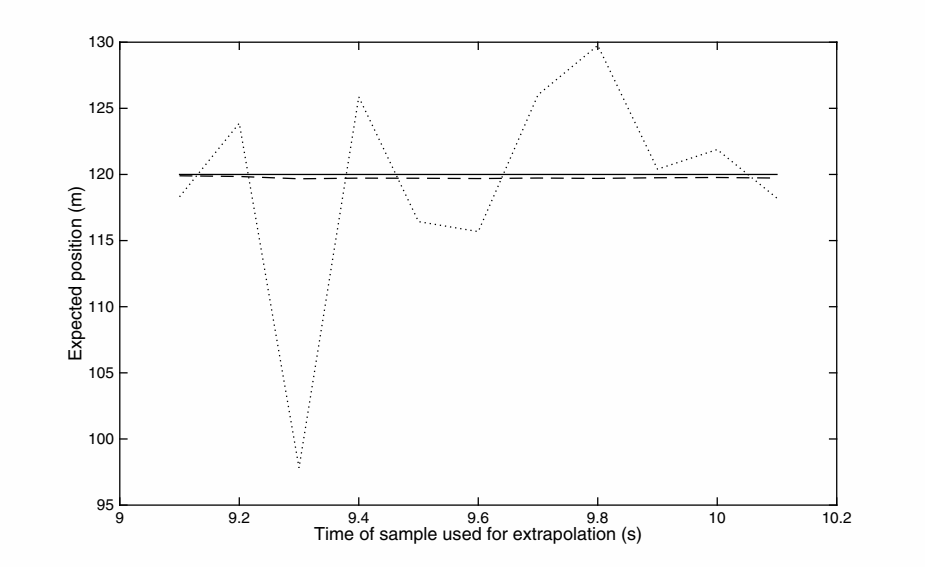

Figure 3.17. *Extrapolation of position 2 seconds ahead for constant-velocity dynamics. True value (solid); from running average of raw measurements (dotted); KF estimation (dashed).*

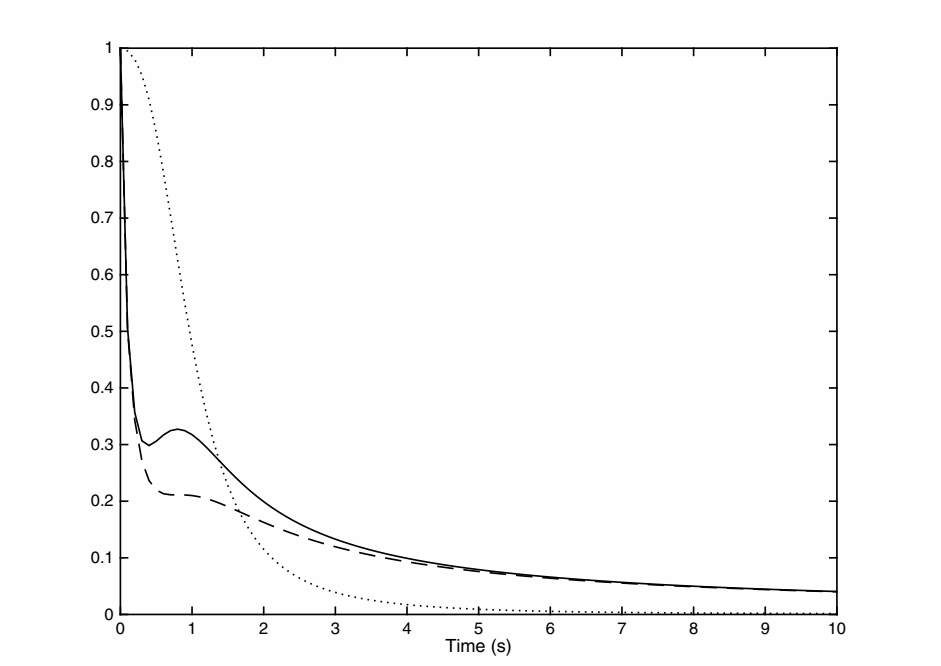

Figure 3.18. *Convergence of the KF parameters for constant-velocity dynamics. K (solid); P_{11} (dashed); P_{22} (dotted).*