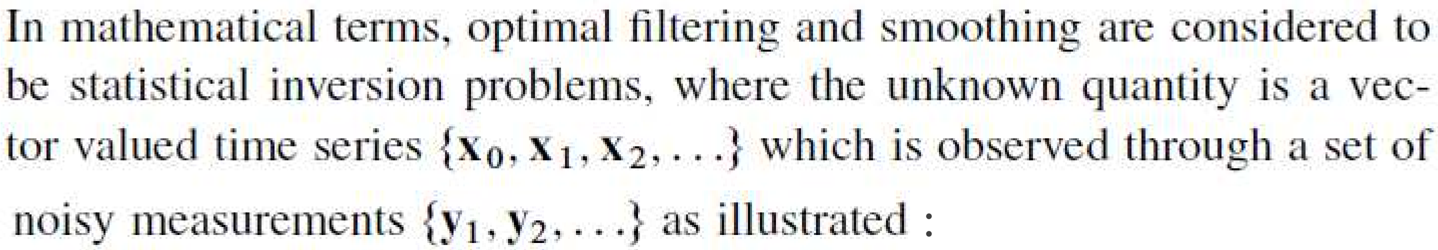
Statistical Filtering and Smoothing - I

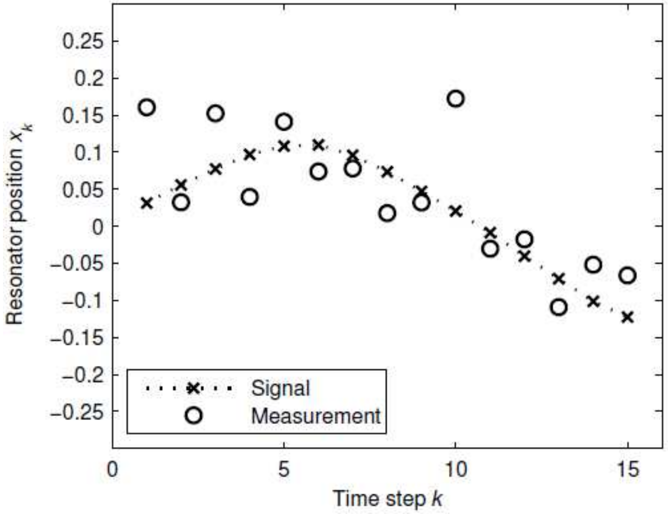
1. PRESENTATION OF OPTIMAL FILTERING AND SMOOTHING



In mathematical terms, optimal filtering and smoothing are considered to be statistical inversion problems, where the unknown quantity is a vector valued time series $\{\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots\}$ which is observed through a set of noisy measurements $\{\mathbf{y}_1, \mathbf{y}_2, \ldots\}$ as illustrated :

$$\begin{tikzcd}
\text{hidden:} \arrow[r, dashed] & \mathbf{x}_{k-1} \arrow[r] \arrow[d] & \mathbf{x}_k \arrow[r] \arrow[d] & \mathbf{x}_{k+1} \arrow[r, dashed] \arrow[d] & {} \\
\text{observed:} & \mathbf{y}_{k-1} & \mathbf{y}_k & \mathbf{y}_{k+1} &
\end{tikzcd}$$

Here is an example of time series which models a discrete-time resonator. The actual signal is hidden and observed through a noisy measurement:



The purpose of the *statistical inversion* at hand is to estimate the hidden states $\mathbf{x}_{0:T} = \{\mathbf{x}_0, \dots, \mathbf{x}_T\}$ from the observed measurements $\mathbf{y}_{1:T} = \{\mathbf{y}_1, \dots, \mathbf{y}_T\}$, which means that in the Bayesian sense we want to compute the joint *posterior distribution* of all the states given all the measurements. In principle, this can be done by a straightforward application of Bayes' rule

$$p(\mathbf{x}_{0:T} \mid \mathbf{y}_{1:T}) = \frac{p(\mathbf{y}_{1:T} \mid \mathbf{x}_{0:T})\, p(\mathbf{x}_{0:T})}{p(\mathbf{y}_{1:T})}, \tag{1.1}$$

where

- $p(\mathbf{x}_{0:T})$, is the *prior distribution* defined by the dynamic model,
- $p(\mathbf{y}_{1:T} \mid \mathbf{x}_{0:T})$ is the likelihood model for the measurements,
- $p(\mathbf{y}_{1:T})$ is the normalization constant defined as

$$p(\mathbf{y}_{1:T}) = \int p(\mathbf{y}_{1:T} \mid \mathbf{x}_{0:T})\, p(\mathbf{x}_{0:T})\, \mathrm{d}\mathbf{x}_{0:T}. \tag{1.2}$$

The model for the states and measurements will be assumed to be of the following type.

- **An initial distribution** specifies the *prior probability distribution* $p(\mathbf{x}_0)$ of the hidden state $\mathbf{x}_0$ at the initial time step $k = 0$.

- **An initial distribution** specifies the *prior probability distribution* $p(\mathbf{x}_0)$ of the hidden state $\mathbf{x}_0$ at the initial time step $k = 0$.
- **A dynamic model** describes the system dynamics and its uncertainties as a *Markov sequence*, defined in terms of the transition probability distribution $p(\mathbf{x}_k \mid \mathbf{x}_{k-1})$.
- **A measurement model** describes how the measurement $\mathbf{y}_k$ depends on the current state $\mathbf{x}_k$. This dependence is modeled by specifying the conditional probability distribution of the measurement given the state, which is denoted as $p(\mathbf{y}_k \mid \mathbf{x}_k)$.

Thus a general probabilistic *state space model* is usually written in the following form:

$$\begin{aligned}
\mathbf{x}_0 &\sim p(\mathbf{x}_0), \\
\mathbf{x}_k &\sim p(\mathbf{x}_k \mid \mathbf{x}_{k-1}), \\
\mathbf{y}_k &\sim p(\mathbf{y}_k \mid \mathbf{x}_k).
\end{aligned} \tag{1.3}$$

Because computing the full joint distribution of the states at all time steps is computationally very inefficient and unnecessary in real-time applications, in *Bayesian filtering and smoothing* the following marginal distributions are considered instead:

- *Filtering distributions* computed by the *Bayesian filter* are the marginal distributions of *the current state* $\mathbf{x}_k$ given *the current and previous measurements* $\mathbf{y}_{1:k} = \{\mathbf{y}_1, \ldots, \mathbf{y}_k\}$:

$$p(\mathbf{x}_k \mid \mathbf{y}_{1:k}), \qquad k = 1, \ldots, T. \tag{1.4}$$

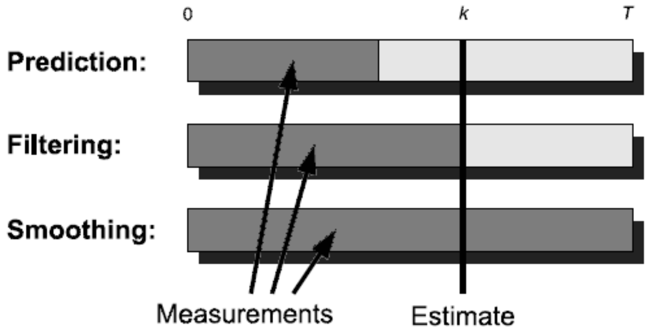
  The result of applying the Bayesian filter to the resonator time series in Figure 1.6 is shown in Figure 1.8.

- *Prediction distributions* which can be computed with the *prediction step of the Bayesian filter* are the marginal distributions of the *future state* $\mathbf{x}_{k+n}$, n steps after the current time step:

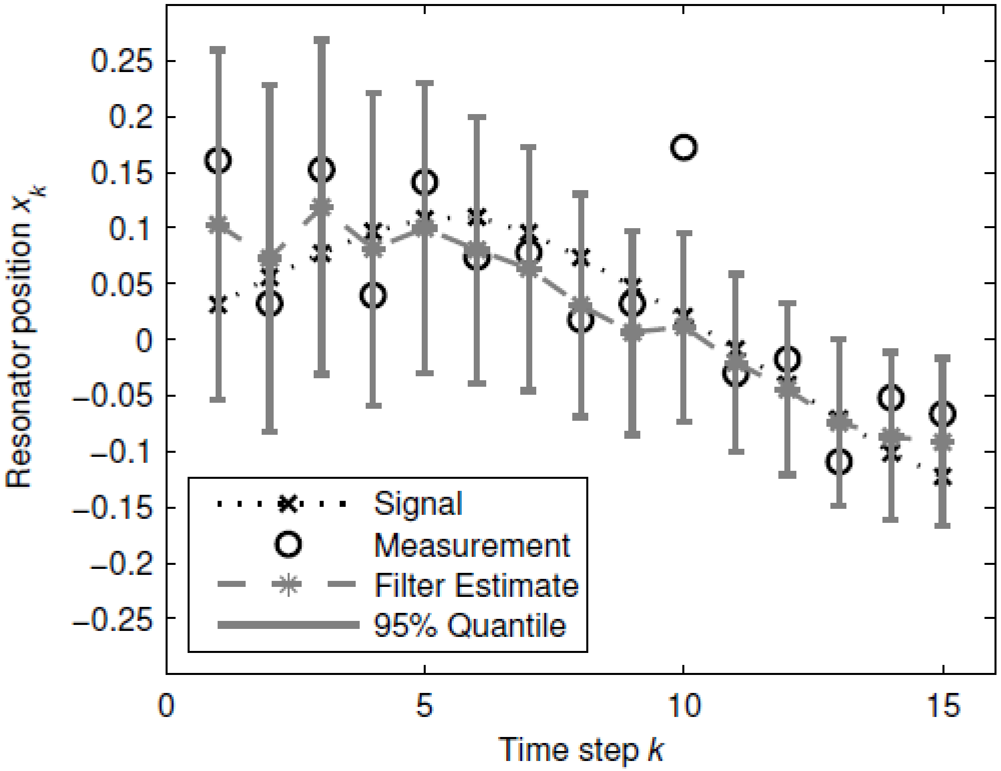
$$p(\mathbf{x}_{k+n} \mid \mathbf{y}_{1:k}), \qquad k = 1, \ldots, T, \quad n = 1, 2, \ldots. \tag{1.5}$$

- *Smoothing distributions* computed by the *Bayesian smoother* are the marginal distributions of the state $\mathbf{x}_k$ given a certain interval $\mathbf{y}_{1:T} = \{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$ of measurements with $T > k$:

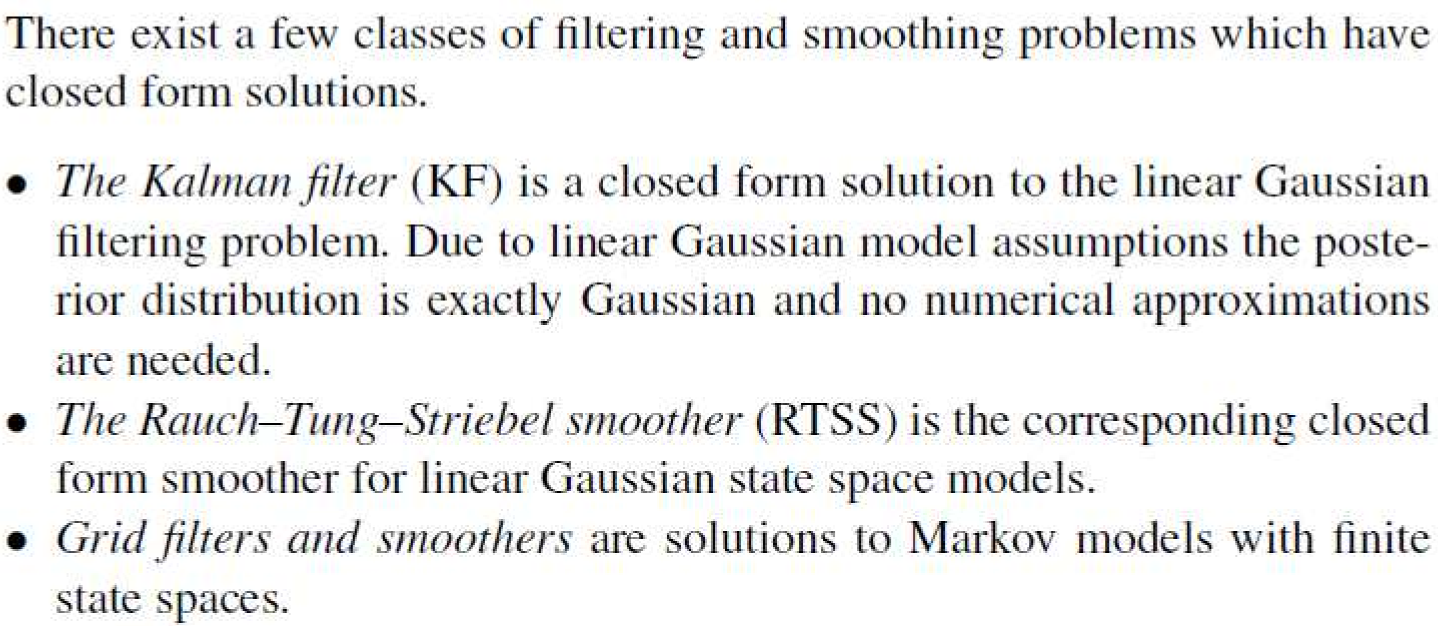
$$p(\mathbf{x}_k \mid \mathbf{y}_{1:T}), \qquad k = 1, \ldots, T. \tag{1.6}$$



Here is shown the result of the filtering distributions for the discrete-time resonator model. The estimates are the means of the filtering distributions and the quantiles are the 95% quantiles of the filtering distributions:



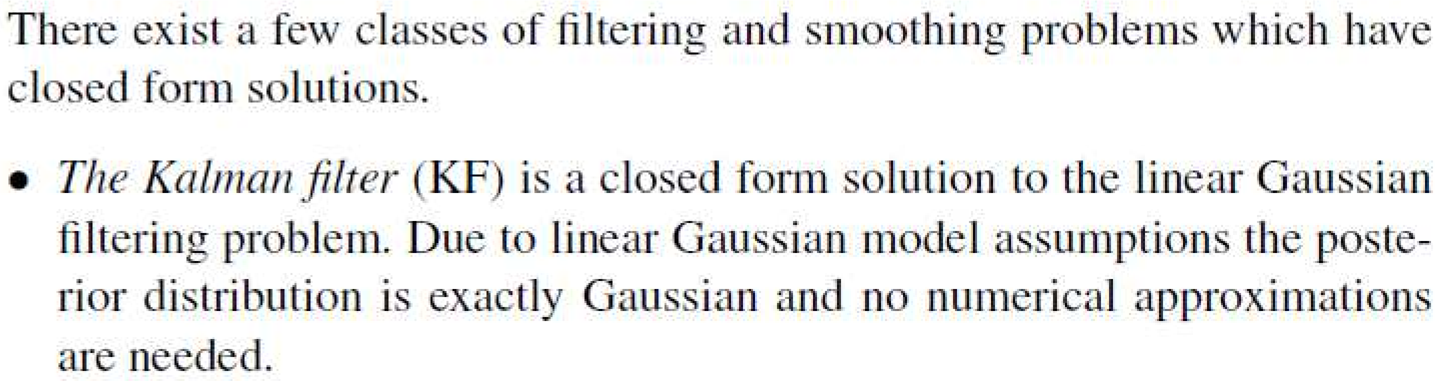
.



There exist a few classes of filtering and smoothing problems which have closed form solutions.

- *The Kalman filter* (KF) is a closed form solution to the linear Gaussian filtering problem. Due to linear Gaussian model assumptions the posterior distribution is exactly Gaussian and no numerical approximations are needed.
- *The Rauch–Tung–Striebel smoother* (RTSS) is the corresponding closed form smoother for linear Gaussian state space models.
- *Grid filters and smoothers* are solutions to Markov models with finite state spaces.

ALGORITHMS FOR FILTERING AND SMOOTHING



There exist a few classes of filtering and smoothing problems which have closed form solutions.

- *The Kalman filter* (KF) is a closed form solution to the linear Gaussian filtering problem. Due to linear Gaussian model assumptions the posterior distribution is exactly Gaussian and no numerical approximations are needed.

- *The Rauch–Tung–Striebel smoother* (RTSS) is the corresponding closed form smoother for linear Gaussian state space models.
- *Grid filters and smoothers* are solutions to Markov models with finite state spaces.

But because the Bayesian optimal filtering and smoothing equations are generally computationally intractable, many kinds of numerical approximation methods have been developed, for example:

- *The extended Kalman filter* (EKF) approximates the non-linear and non-Gaussian measurement and dynamic models by linearization, that is, by forming a Taylor series expansion at the nominal (or maximum a posteriori, MAP) solution. This results in a Gaussian approximation to the filtering distribution.
- *The extended Rauch–Tung–Striebel smoother* (ERTSS) is the approximate non-linear smoothing algorithm corresponding to EKF.
- *The unscented Kalman filter* (UKF) approximates the propagation of densities through the non-linearities of measurement and noise processes using the *unscented transform*. This also results in a Gaussian approximation.
- *The unscented Rauch–Tung–Striebel smoother* (URTSS) is the approximate non-linear smoothing algorithm corresponding to UKF.

- *Sequential Monte Carlo methods* or *particle filters and smoothers* represent the posterior distribution as a weighted set of Monte Carlo samples.
- *The unscented particle filter* (UPF) and *local linearization* based particle filtering methods use UKFs and EKFs, respectively, for approximating the optimal importance distributions in particle filters.
- *Rao–Blackwellized particle filters and smoothers* use closed form integration (e.g., Kalman filters and RTS smoothers) for some of the state variables and Monte Carlo integration for others.
- *Grid based approximation methods* approximate the filtering and smoothing distributions as discrete distributions on a finite grid.
- *Other methods* also exist, for example, based on Gaussian mixtures, series expansions, describing functions, basis function expansions, exponential family of distributions, variational Bayesian methods, and batch Monte Carlo (e.g., Markov chain Monte Carlo, MCMC, methods).

In state space models of dynamic systems, there are often *unknown or uncertain parameters* $\boldsymbol{\theta}$ which should be estimated along with the state itself. For example, in a stochastic resonator model, the frequency of the resonator might be unknown. Also the noise variances might be only known approximately or they can be completely unknown. Although, formally, we can always augment unknown parameters as part of the state, in practice it is often useful to consider parameter estimation separately.

In a Bayesian setting, the proper way to estimate the parameters is by setting a prior distribution on the parameters $p(\boldsymbol{\theta})$ and treating them as additional random variables in the model. When unknown parameters are present, the state space model in Equation (1.3) becomes

$$\begin{aligned}
\boldsymbol{\theta} &\sim p(\boldsymbol{\theta}), \\
\mathbf{x}_0 &\sim p(\mathbf{x}_0 \mid \boldsymbol{\theta}), \\
\mathbf{x}_k &\sim p(\mathbf{x}_k \mid \mathbf{x}_{k-1}, \boldsymbol{\theta}), \\
\mathbf{y}_k &\sim p(\mathbf{y}_k \mid \mathbf{x}_k, \boldsymbol{\theta}).
\end{aligned} \tag{1.7}$$

The full Bayesian solution to this problem would require the computation of the full *joint posterior distribution of states and parameters* $p(\mathbf{x}_{0:T}, \boldsymbol{\theta} \mid \mathbf{y}_{1:T})$. Unfortunately, computing this joint posterior of the states and parameters is even harder than computation of the joint distribution of states alone, and thus this task is intractable.

Fortunately, when run with fixed parameters $\boldsymbol{\theta}$, the Bayesian filter algorithm produces the sequence of distributions $p(\mathbf{y}_k \mid \mathbf{y}_{1:k-1}, \boldsymbol{\theta})$ for $k = 1, \ldots, T$ as side products. Once we have these, we can form the *marginal posterior distribution of parameters* as follows:

$$p(\boldsymbol{\theta} \mid \mathbf{y}_{1:T}) \propto p(\boldsymbol{\theta}) \prod_{k=1}^{T} p(\mathbf{y}_k \mid \mathbf{y}_{1:k-1}, \boldsymbol{\theta}), \tag{1.8}$$

where we have denoted $p(\mathbf{y}_1 \mid \mathbf{y}_{1:0}, \boldsymbol{\theta}) \triangleq p(\mathbf{y}_1 \mid \boldsymbol{\theta})$ for notational convenience. When combined with the smoothing distributions, we can form all the marginal joint distributions of states and parameters as follows:

$$p(\mathbf{x}_k, \boldsymbol{\theta} \mid \mathbf{y}_{1:T}) = p(\mathbf{x}_k \mid \mathbf{y}_{1:T}, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathbf{y}_{1:T}), \tag{1.9}$$

for $k = 1, \ldots, T$, where $p(\mathbf{x}_k \mid \mathbf{y}_{1:T}, \boldsymbol{\theta})$ is the smoothing distribution of the states with fixed model parameters $\boldsymbol{\theta}$. However, we cannot compute the full joint posterior distribution of states and parameters, which is the price of only using a constant number of computations per time step.

Remark:

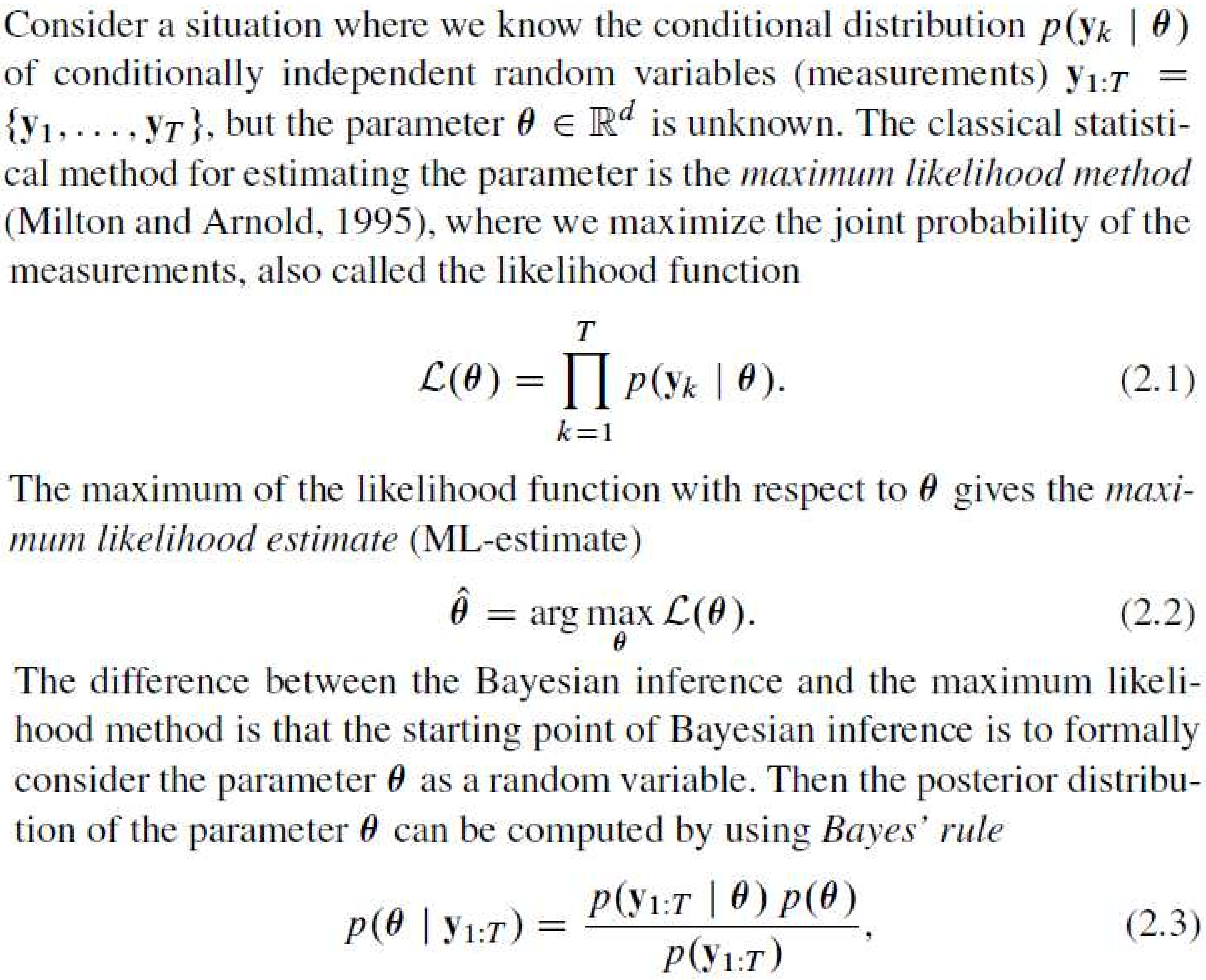
Although here we use the term *parameter estimation*, it might sometimes be the case that we are not actually interested in the values of the parameters as such, but we just do not know the values of them. In that case the proper Bayesian approach is to *integrate out* the parameters. For example, to compute the smoothing distributions in the presence of unknown parameters we can integrate out the parameters from the joint distribution in Equation (1.9):

$$
\begin{aligned}
p(\mathbf{x}_k \mid \mathbf{y}_{1:T}) &= \int p(\mathbf{x}_k, \boldsymbol{\theta} \mid \mathbf{y}_{1:T}) \, \mathrm{d}\boldsymbol{\theta} \\
&= \int p(\mathbf{x}_k \mid \mathbf{y}_{1:T}, \boldsymbol{\theta}) \, p(\boldsymbol{\theta} \mid \mathbf{y}_{1:T}) \, \mathrm{d}\boldsymbol{\theta}. \qquad (1.10)
\end{aligned}
$$

Many of the Bayesian methods for parameter estimation indeed allow this to be done (approximately). For example, by using the parameter samples produced by a Markov chain Monte Carlo (MCMC) method, it is possible to form a Monte Carlo approximation to the above integral.

2. BAYESIAN INFERENCE

2.1 Connection to maximum likelihood estimation



Consider a situation where we know the conditional distribution $p(\mathbf{y}_k \mid \boldsymbol{\theta})$ of conditionally independent random variables (measurements) $\mathbf{y}_{1:T} = \{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$, but the parameter $\boldsymbol{\theta} \in \mathbb{R}^d$ is unknown. The classical statistical method for estimating the parameter is the *maximum likelihood method* (Milton and Arnold, 1995), where we maximize the joint probability of the measurements, also called the likelihood function

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{k=1}^{T} p(\mathbf{y}_k \mid \boldsymbol{\theta}). \tag{2.1}$$

The maximum of the likelihood function with respect to $\boldsymbol{\theta}$ gives the *maximum likelihood estimate* (ML-estimate)

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}). \tag{2.2}$$

The difference between the Bayesian inference and the maximum likelihood method is that the starting point of Bayesian inference is to formally consider the parameter $\boldsymbol{\theta}$ as a random variable. Then the posterior distribution of the parameter $\boldsymbol{\theta}$ can be computed by using *Bayes' rule*

$$p(\boldsymbol{\theta} \mid \mathbf{y}_{1:T}) = \frac{p(\mathbf{y}_{1:T} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathbf{y}_{1:T})}, \tag{2.3}$$

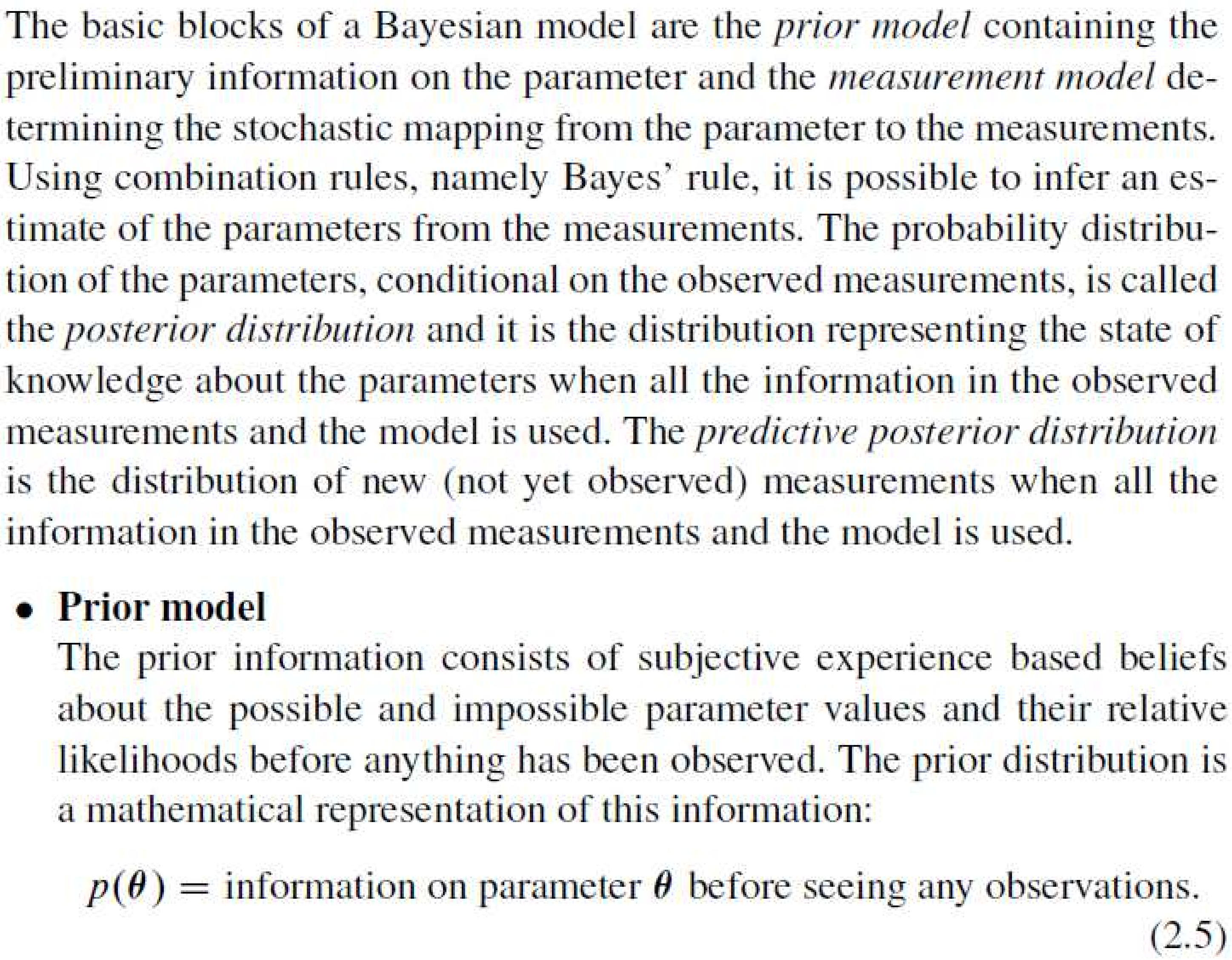
where $p(\boldsymbol{\theta})$ is the prior distribution which models the prior beliefs on the parameter before we have seen any data, and $p(\mathbf{y}_{1:T})$ is a normalization term which is independent of the parameter $\boldsymbol{\theta}$. This normalization constant is often left out and if the measurements $\mathbf{y}_{1:T}$ are conditionally independent given $\boldsymbol{\theta}$, the posterior distribution of the parameter can be written as

$$p(\boldsymbol{\theta} \mid \mathbf{y}_{1:T}) \propto p(\boldsymbol{\theta}) \prod_{k=1}^{T} p(\mathbf{y}_k \mid \boldsymbol{\theta}). \tag{2.4}$$

Because we are dealing with a distribution, we might now choose the most probable value of the random variable, the *maximum a posteriori* (MAP) estimate, which is given by the maximum of the posterior distribution. The optimal estimate in the mean squared sense is the posterior mean of the parameter (MMSE-estimate). There are an infinite number of other ways of choosing the point estimate from the distribution and the best way depends on the assumed loss or cost function (or utility function). The ML-estimate can be seen as a MAP-estimate with uniform prior $p(\boldsymbol{\theta}) \propto 1$ on the parameter $\boldsymbol{\theta}$.

.

2.2 The structure of Bayesian models



The basic blocks of a Bayesian model are the *prior model* containing the preliminary information on the parameter and the *measurement model* determining the stochastic mapping from the parameter to the measurements. Using combination rules, namely Bayes' rule, it is possible to infer an estimate of the parameters from the measurements. The probability distribution of the parameters, conditional on the observed measurements, is called the *posterior distribution* and it is the distribution representing the state of knowledge about the parameters when all the information in the observed measurements and the model is used. The *predictive posterior distribution* is the distribution of new (not yet observed) measurements when all the information in the observed measurements and the model is used.

- **Prior model**

  The prior information consists of subjective experience based beliefs about the possible and impossible parameter values and their relative likelihoods before anything has been observed. The prior distribution is a mathematical representation of this information:

  $$p(\theta) = \text{information on parameter } \boldsymbol{\theta} \text{ before seeing any observations.} \tag{2.5}$$

.

- **Measurement model**
  Between the true parameters and the measurements there is often a causal, but inaccurate or noisy relationship. This relationship is mathematically modeled using the measurement model:

  $$p(\mathbf{y} \mid \boldsymbol{\theta}) = \text{distribution of observation } \mathbf{y} \text{ given the parameters } \boldsymbol{\theta}. \tag{2.6}$$

- **Posterior distribution**
  The posterior distribution is the conditional distribution of the parameters given the observations. It represents the information we have after the measurement $\mathbf{y}$ has been obtained. It can be computed by using Bayes' rule

  $$p(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathbf{y})} \propto p(\mathbf{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}), \tag{2.7}$$

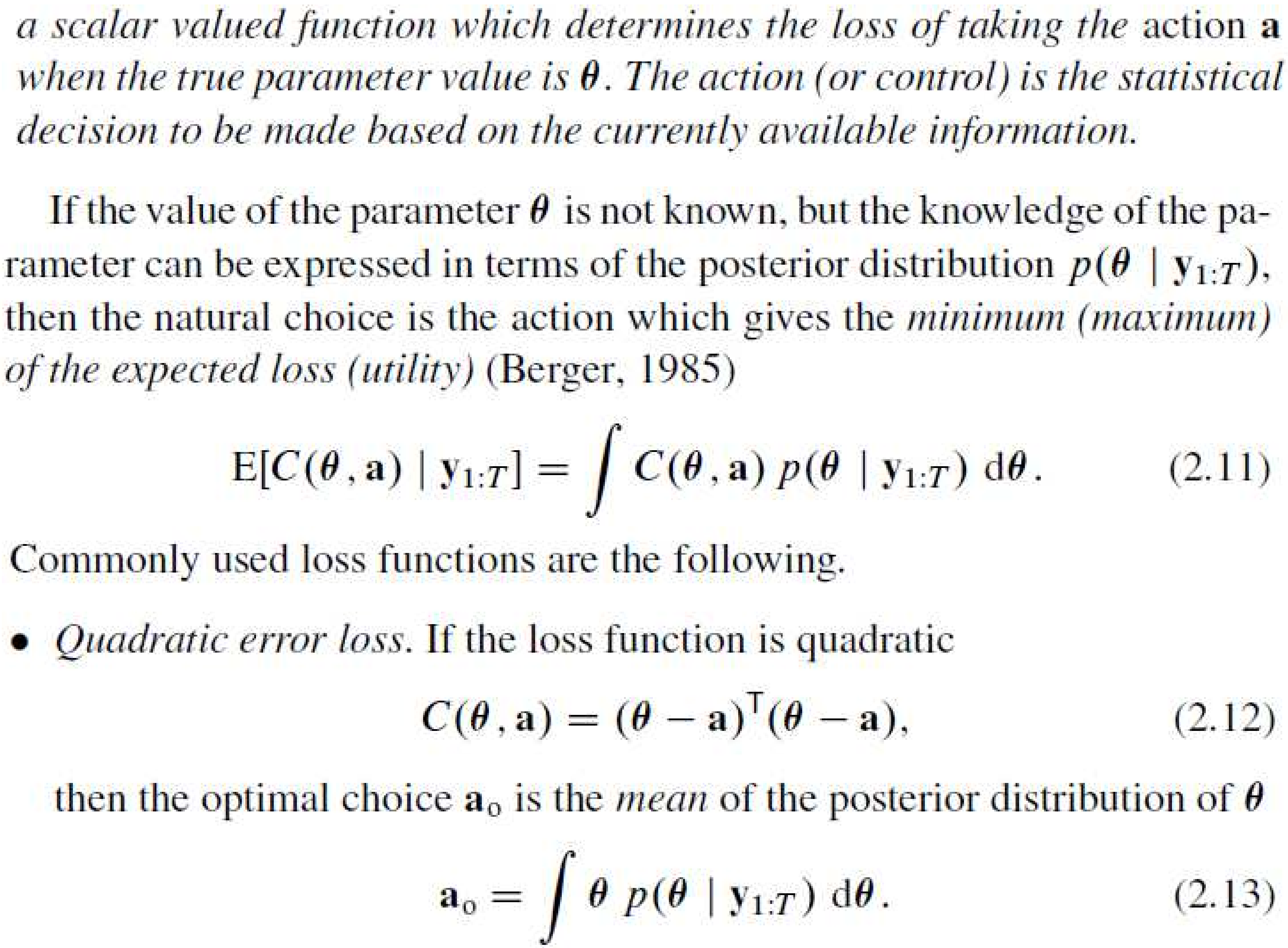
where the normalization constant is given as

$$p(\mathbf{y}) = \int p(\mathbf{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}. \tag{2.8}$$

.

2.3 Point Estimates

Definition (Loss function): *A loss function or cost function $C(\theta, \mathbf{a})$ is*



a scalar valued function which determines the loss of taking the action $\mathbf{a}$ *when the true parameter value is $\boldsymbol{\theta}$. The action (or control) is the statistical decision to be made based on the currently available information.*

If the value of the parameter $\boldsymbol{\theta}$ is not known, but the knowledge of the parameter can be expressed in terms of the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{y}_{1:T})$, then the natural choice is the action which gives the *minimum (maximum) of the expected loss (utility)* (Berger, 1985)

$$\mathrm{E}[C(\boldsymbol{\theta}, \mathbf{a}) \mid \mathbf{y}_{1:T}] = \int C(\boldsymbol{\theta}, \mathbf{a})\, p(\boldsymbol{\theta} \mid \mathbf{y}_{1:T})\, \mathrm{d}\boldsymbol{\theta}. \tag{2.11}$$

Commonly used loss functions are the following.

- *Quadratic error loss*. If the loss function is quadratic

$$C(\boldsymbol{\theta}, \mathbf{a}) = (\boldsymbol{\theta} - \mathbf{a})^\mathsf{T} (\boldsymbol{\theta} - \mathbf{a}), \tag{2.12}$$

then the optimal choice $\mathbf{a}_\mathrm{o}$ is the *mean* of the posterior distribution of $\boldsymbol{\theta}$

$$\mathbf{a}_\mathrm{o} = \int \boldsymbol{\theta}\, p(\boldsymbol{\theta} \mid \mathbf{y}_{1:T})\, \mathrm{d}\boldsymbol{\theta}. \tag{2.13}$$

This posterior mean based estimate is often called the *minimum mean squared error (MMSE)* estimate of the parameter $\boldsymbol{\theta}$. The quadratic loss is the most commonly used loss function, because it is easy to handle mathematically and because in the case of Gaussian posterior distribution the MAP estimate and the median coincide with the posterior mean.

- *Absolute error loss*. The loss function of the form

$$C(\boldsymbol{\theta}, \mathbf{a}) = \sum_i |\theta_i - a_i| \tag{2.14}$$

  is called an absolute error loss and in this case the optimal choice is the *median* of the distribution (the medians of the marginal distributions in the multi-dimensional case).

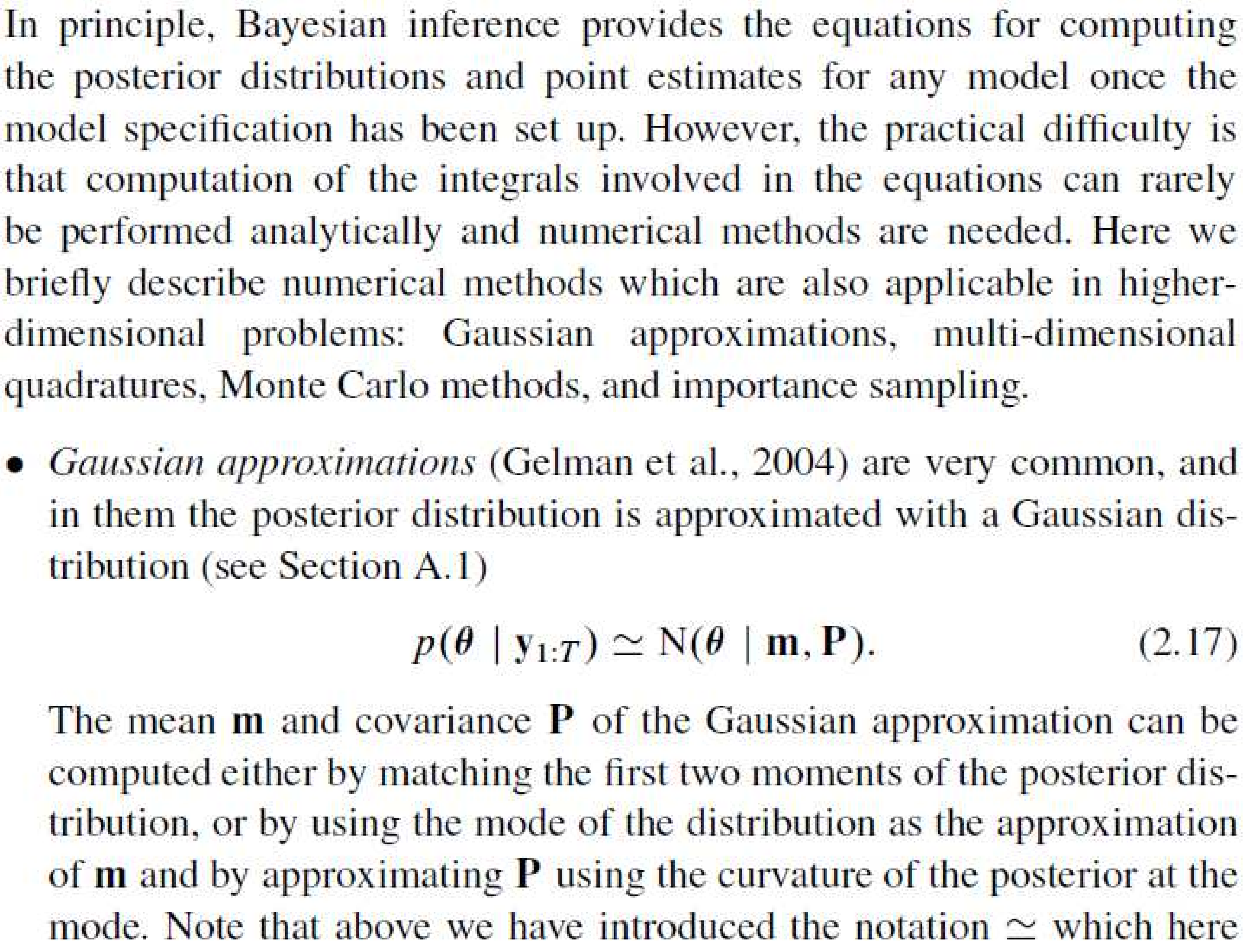
- *0–1 loss*. If the loss function is of the form

$$C(\boldsymbol{\theta}, \mathbf{a}) = -\delta(\mathbf{a} - \boldsymbol{\theta}), \tag{2.15}$$

  where $\delta(\cdot)$ is the Dirac's delta function, then the optimal choice is the maximum (mode) of the posterior distribution, that is, the *maximum a posterior (MAP)* estimate of the parameter. If the random variable $\boldsymbol{\theta}$ is discrete the corresponding loss function can be defined as

$$C(\boldsymbol{\theta}, \mathbf{a}) = \begin{cases} 0, & \text{if } \boldsymbol{\theta} = \mathbf{a}, \\ 1, & \text{if } \boldsymbol{\theta} \neq \mathbf{a}. \end{cases} \tag{2.16}$$

2.4 Numerical Approximation Methods



In principle, Bayesian inference provides the equations for computing the posterior distributions and point estimates for any model once the model specification has been set up. However, the practical difficulty is that computation of the integrals involved in the equations can rarely be performed analytically and numerical methods are needed. Here we briefly describe numerical methods which are also applicable in higher-dimensional problems: Gaussian approximations, multi-dimensional quadratures, Monte Carlo methods, and importance sampling.

- *Gaussian approximations* (Gelman et al., 2004) are very common, and in them the posterior distribution is approximated with a Gaussian distribution (see Section A.1)

  $$p(\boldsymbol{\theta} \mid \mathbf{y}_{1:T}) \simeq \mathrm{N}(\boldsymbol{\theta} \mid \mathbf{m}, \mathbf{P}). \tag{2.17}$$

  The mean $\mathbf{m}$ and covariance $\mathbf{P}$ of the Gaussian approximation can be computed either by matching the first two moments of the posterior distribution, or by using the mode of the distribution as the approximation of $\mathbf{m}$ and by approximating $\mathbf{P}$ using the curvature of the posterior at the mode. Note that above we have introduced the notation $\simeq$ which here means that the left-hand side is *assumed* to be approximately equal to the right-hand side, even though we know that it will not be true in most practical situations nor can we control the approximation error in any practical way.

- *Multi-dimensional quadrature or cubature integration methods* such as Gauss–Hermite quadrature can also be used if the dimensionality of the integral is moderate. The idea is to deterministically form a representative set of sample points $\{\boldsymbol{\theta}^{(i)} : i = 1, \ldots, N\}$ (sometimes called *sigma points*) and form the approximation of the integral as the weighted average

$$\mathrm{E}[\mathbf{g}(\boldsymbol{\theta}) \mid \mathbf{y}_{1:T}] \approx \sum_{i=1}^{N} W_i \, \mathbf{g}(\boldsymbol{\theta}^{(i)}), \tag{2.18}$$

where the numerical values of the weights W_i are determined by the algorithm. The sample points and weights can be selected, for example, to give exact answers for polynomials up to certain degree or to account for the moments up to certain degree. Above we have used the notation $\approx$ to mean that the expressions are approximately equal in some suitable limit (here $N \to \infty$) or in some verifiable conditions.

- In direct *Monte Carlo methods* a set of N samples from the posterior distribution is randomly drawn

$$\boldsymbol{\theta}^{(i)} \sim p(\boldsymbol{\theta} \mid \mathbf{y}_{1:T}), \qquad i = 1, \ldots, N, \tag{2.19}$$

and expectation of any function $\mathbf{g}(\cdot)$ can be then approximated as the sample average

$$\mathrm{E}[\mathbf{g}(\boldsymbol{\theta}) \mid \mathbf{y}_{1:T}] \approx \frac{1}{N} \sum_{i} \mathbf{g}(\boldsymbol{\theta}^{(i)}). \tag{2.20}$$

Another interpretation of this is that Monte Carlo methods form an approximation of the posterior density of the form

$$p(\boldsymbol{\theta} \mid \mathbf{y}_{1:T}) \approx \frac{1}{N} \sum_{i=1}^{N} \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i)}), \tag{2.21}$$

where $\delta(\cdot)$ is the Dirac delta function. The convergence of Monte Carlo approximation is guaranteed by the *central limit theorem (CLT)* (see, e.g., Liu, 2001) and the error term is, at least in theory, under certain ideal conditions, independent of the dimensionality of $\boldsymbol{\theta}$. The rule of thumb is that the error should decrease like the square root of the number of samples, regardless of the dimensions.

-

- Efficient methods for generating Monte Carlo samples are the *Markov chain Monte Carlo* (MCMC) methods (see, e.g., Gilks et al., 1996; Liu, 2001; Brooks et al., 2011). In MCMC methods, a Markov chain is constructed such that it has the target distribution as its stationary distribution. By simulating the Markov chain, samples from the target distribution can be generated.
- *Importance sampling* (see, e.g., Liu, 2001) is a simple algorithm for generating *weighted* samples from the target distribution. The difference between this and direct Monte Carlo sampling and MCMC is that each of the particles has an associated weight, which corrects for the difference between the actual target distribution and the approximate importance distribution $\pi(\cdot)$ from which the sample was drawn.

An importance sampling estimate can be formed by drawing N samples from the *importance distribution*

$$\boldsymbol{\theta}^{(i)} \sim \pi(\boldsymbol{\theta} \mid \mathbf{y}_{1:T}), \qquad i = 1, \ldots, N. \tag{2.22}$$

The *importance weights* are then computed as

$$\tilde{w}^{(i)} = \frac{1}{N} \frac{p(\boldsymbol{\theta}^{(i)} \mid \mathbf{y}_{1:T})}{\pi(\boldsymbol{\theta}^{(i)} \mid \mathbf{y}_{1:T})}, \tag{2.23}$$

and the expectation of any function $\mathbf{g}(\cdot)$ can be then approximated as

$$\mathrm{E}[\mathbf{g}(\boldsymbol{\theta}) \mid \mathbf{y}_{1:T}] \approx \sum_{i=1}^{N} \tilde{w}^{(i)} \, \mathbf{g}(\boldsymbol{\theta}^{(i)}), \tag{2.24}$$

or alternatively as

$$\mathrm{E}[\mathbf{g}(\boldsymbol{\theta}) \mid \mathbf{y}_{1:T}] \approx \frac{\sum_{i=1}^{N} \tilde{w}^{(i)} \, \mathbf{g}(\boldsymbol{\theta}^{(i)})}{\sum_{i=1}^{N} \tilde{w}^{(i)}}. \tag{2.25}$$

2.5 EXERCISES

2.1 Prove that median of distribution $p(\theta)$ minimizes the expected value of the absolute error loss function

$$\mathrm{E}[|\theta - a|] = \int |\theta - a| \, p(\theta) \, \mathrm{d}\theta. \tag{2.26}$$

2.2 Find the optimal point estimate $\mathbf{a}$ which minimizes the expected value of the loss function

$$C(\boldsymbol{\theta}, \mathbf{a}) = (\boldsymbol{\theta} - \mathbf{a})^{\mathsf{T}} \, \mathbf{R} \, (\boldsymbol{\theta} - \mathbf{a}), \tag{2.27}$$

where $\mathbf{R}$ is a positive definite matrix, and the distribution of the parameter is $\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \mathbf{y}_{1:T})$.

2.3 Assume that we have obtained T measurement pairs (x_k, y_k) from the linear regression model

$$y_k = \theta_1 \, x_k + \theta_2, \qquad k = 1, 2, \ldots, T. \tag{2.28}$$

The purpose is now to derive estimates of the parameters θ_1 and θ_2 such that the following error is minimized (least squares estimate):

$$E(\theta_1, \theta_2) = \sum_{k=1}^{T} (y_k - \theta_1 \, x_k - \theta_2)^2. \tag{2.29}$$

(a) Define $\mathbf{y} = (y_1 \ \ldots \ y_T)^\mathsf{T}$ and $\boldsymbol{\theta} = (\theta_1 \ \theta_2)^\mathsf{T}$. Show that the set of Equations (2.28) can be written in matrix form as

$$\mathbf{y} = \mathbf{X} \, \boldsymbol{\theta},$$

with a suitably defined matrix $\mathbf{X}$.

(b) Write the error function in Equation (2.29) in matrix form in terms of $\mathbf{y}$, $\mathbf{X}$, and $\boldsymbol{\theta}$.

(c) Compute the gradient of the matrix form error function and solve the least squares estimate of the parameter $\boldsymbol{\theta}$ by finding the point where the gradient is zero.

2.4 Assume that in the linear regression model above (Equation (2.28)) we set independent Gaussian priors for the parameters θ_1 and θ_2 as follows:

$$\theta_1 \sim \mathrm{N}(0, \sigma^2),$$
$$\theta_2 \sim \mathrm{N}(0, \sigma^2),$$

where the variance σ^2 is known. The measurements y_k are modeled as

$$y_k = \theta_1\, x_k + \theta_2 + \varepsilon_k, \qquad k = 1, 2, \ldots, T,$$

where the terms ε_k are independent Gaussian errors with mean 0 and variance 1, that is, $\varepsilon_k \sim \mathrm{N}(0, 1)$. The values x_k are fixed and known. The posterior distribution can be now written as

$$\begin{aligned} &p(\boldsymbol{\theta} \mid y_1, \ldots, y_T) \\ &\propto \exp\left(-\frac{1}{2}\sum_{k=1}^{T}(y_k - \theta_1\, x_k - \theta_2)^2\right) \exp\left(-\frac{1}{2\sigma^2}\theta_1^2\right) \exp\left(-\frac{1}{2\sigma^2}\theta_2^2\right). \end{aligned}$$

The posterior distribution can be seen to be Gaussian and your task is to derive its mean and covariance.

The posterior distribution can be seen to be Gaussian and your task is to derive its mean and covariance.

(a) Write the exponent of the posterior distribution in matrix form as in Exercise 2.3 (in terms of $\mathbf{y}$, $\mathbf{X}$, $\boldsymbol{\theta}$, and σ^2).

(b) Because a Gaussian distribution is always symmetric, its mean $\mathbf{m}$ is at the maximum of the distribution. Find the posterior mean by computing the gradient of the exponent and finding where it vanishes.

(c) Find the covariance of the distribution by computing the second derivative matrix (Hessian matrix) $\mathbf{H}$ of the exponent. The posterior covariance is then $\mathbf{P} = -\mathbf{H}^{-1}$ (why?).

(d) What is the resulting posterior distribution? What is the relationship with the least squares estimate in Exercise 2.3?

#

3. BLOCK (or BATCH) AND RECURSIVE BAYESIAN ESTIMATION

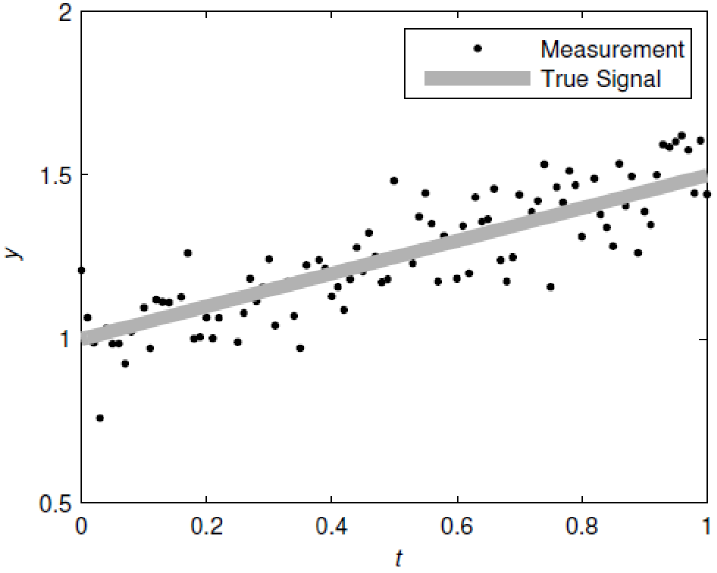
3.1 Block Linear Regression

Consider the *linear regression model*

$$y_k = \theta_1 + \theta_2 \, t_k + \varepsilon_k, \tag{3.1}$$

where we assume that the measurement noise is zero mean Gaussian with a given variance $\varepsilon_k \sim \mathrm{N}(0, \sigma^2)$ and the prior distribution of the parameters $\boldsymbol{\theta} = (\theta_1 \;\; \theta_2)^\mathsf{T}$ is Gaussian with known mean and covariance $\boldsymbol{\theta} \sim \mathrm{N}(\mathbf{m}_0, \mathbf{P}_0)$. In the classical linear regression problem we want to estimate the parameters $\boldsymbol{\theta}$ from a set of measurement data $\mathcal{D} = \{(t_1, y_1), \ldots, (t_T, y_T)\}$. The measurement data and the true linear function used in simulation are illustrated in Figure 3.1.

#



In compact *probabilistic notation* the linear regression model can be written as

$$\begin{aligned} p(y_k \mid \boldsymbol{\theta}) &= \mathrm{N}(y_k \mid \mathbf{H}_k \, \boldsymbol{\theta}, \sigma^2) \\ p(\boldsymbol{\theta}) &= \mathrm{N}(\boldsymbol{\theta} \mid \mathbf{m}_0, \mathbf{P}_0), \end{aligned} \tag{3.2}$$

where we have introduced the row vector $\mathbf{H}_k = (1 \; t_k)$ and $\mathrm{N}(\cdot)$ denotes the Gaussian probability density function (see Section A.1). Note that we denote the row vector $\mathbf{H}_k$ in matrix notation, because it generally is a matrix (when the measurements are vector valued) and we want to avoid using different notations for scalar and vector measurements. The likelihood of y_k is conditional on the regressors t_k also (or equivalently $\mathbf{H}_k$), but because the regressors are assumed to be known, to simplify the notation we will not denote this dependence explicitly and from now on this dependence is assumed to be understood from the context.

The *batch solution* to the linear regression problem in Equation (3.2) can be obtained by a straightforward application of Bayes' rule

$$\begin{aligned} p(\boldsymbol{\theta} \mid y_{1:T}) &\propto p(\boldsymbol{\theta}) \prod_{k=1}^{T} p(y_k \mid \boldsymbol{\theta}) \\ &= \mathrm{N}(\boldsymbol{\theta} \mid \mathbf{m}_0, \mathbf{P}_0) \prod_{k=1}^{T} \mathrm{N}(y_k \mid \mathbf{H}_k \, \boldsymbol{\theta}, \sigma^2). \end{aligned}$$

In the *posterior distribution* above, we assume the conditioning on t_k and $\mathbf{H}_k$, but will not denote it explicitly. Thus the posterior distribution is denoted to be conditional on $y_{1:T}$, and not on the data set $\mathcal{D}$ also containing the regressor values t_k. The reason for this simplification is that the simplified notation will also work in more general filtering problems, where there is no natural way of defining the associated regressor variables.

Because the prior and likelihood are Gaussian, the *posterior distribution* will also be Gaussian:

The mean and covariance can be obtained by completing the quadratic form in the exponent, which gives:

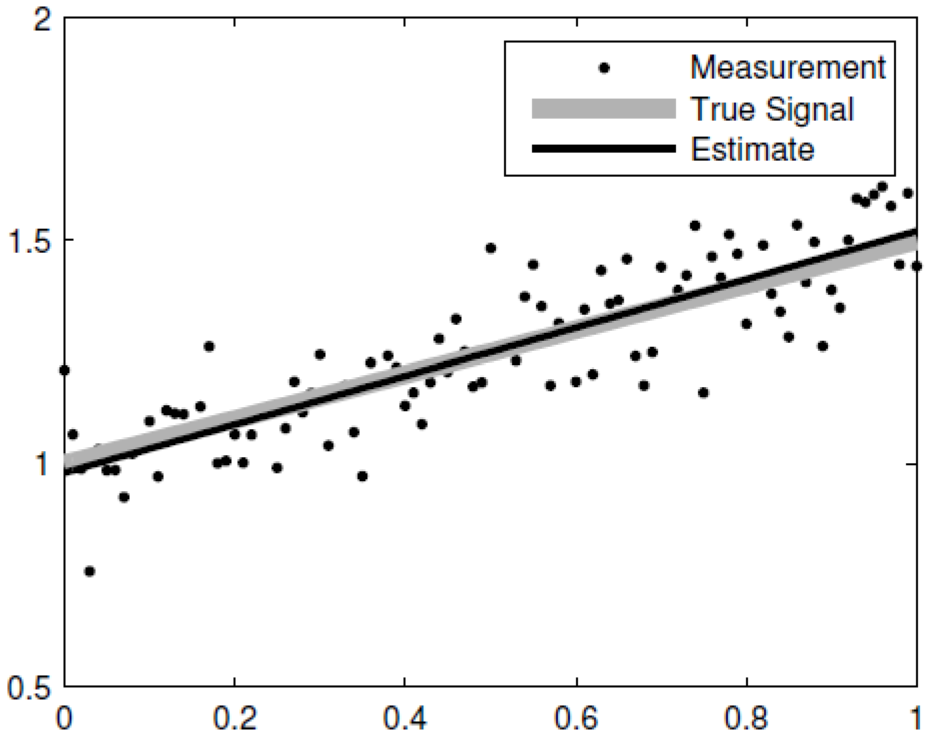
$$\begin{aligned}\mathbf{m}_T &= \left[\mathbf{P}_0^{-1} + \frac{1}{\sigma^2}\mathbf{H}^\mathsf{T}\mathbf{H}\right]^{-1}\left[\frac{1}{\sigma^2}\mathbf{H}^\mathsf{T}\mathbf{y} + \mathbf{P}_0^{-1}\mathbf{m}_0\right], \\ \mathbf{P}_T &= \left[\mathbf{P}_0^{-1} + \frac{1}{\sigma^2}\mathbf{H}^\mathsf{T}\mathbf{H}\right]^{-1}, \end{aligned} \tag{3.4}$$

where $\mathbf{H}_k = (1 \; t_k)$ and

$$\mathbf{H} = \begin{pmatrix}\mathbf{H}_1 \\ \vdots \\ \mathbf{H}_T\end{pmatrix} = \begin{pmatrix}1 & t_1 \\ \vdots & \vdots \\ 1 & t_T\end{pmatrix}, \qquad \mathbf{y} = \begin{pmatrix}y_1 \\ \vdots \\ y_T\end{pmatrix}. \tag{3.5}$$

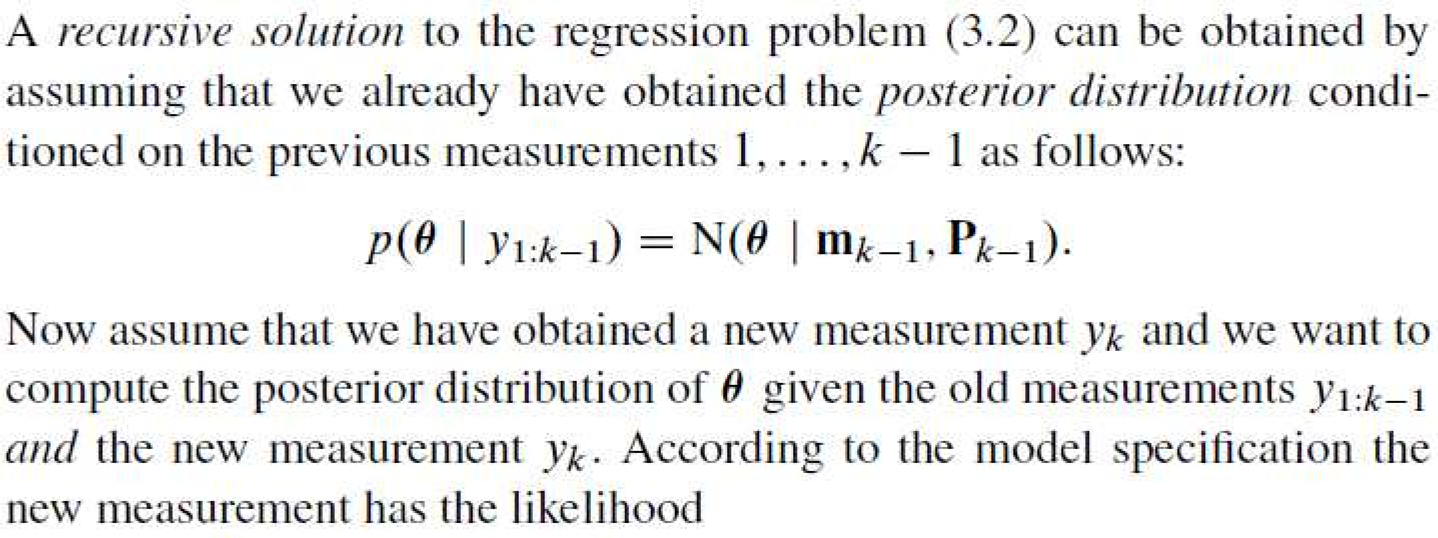
#

Figure 3.2 shows the result of batch linear regression, where the posterior mean parameter values are used as the linear regression parameters.



#

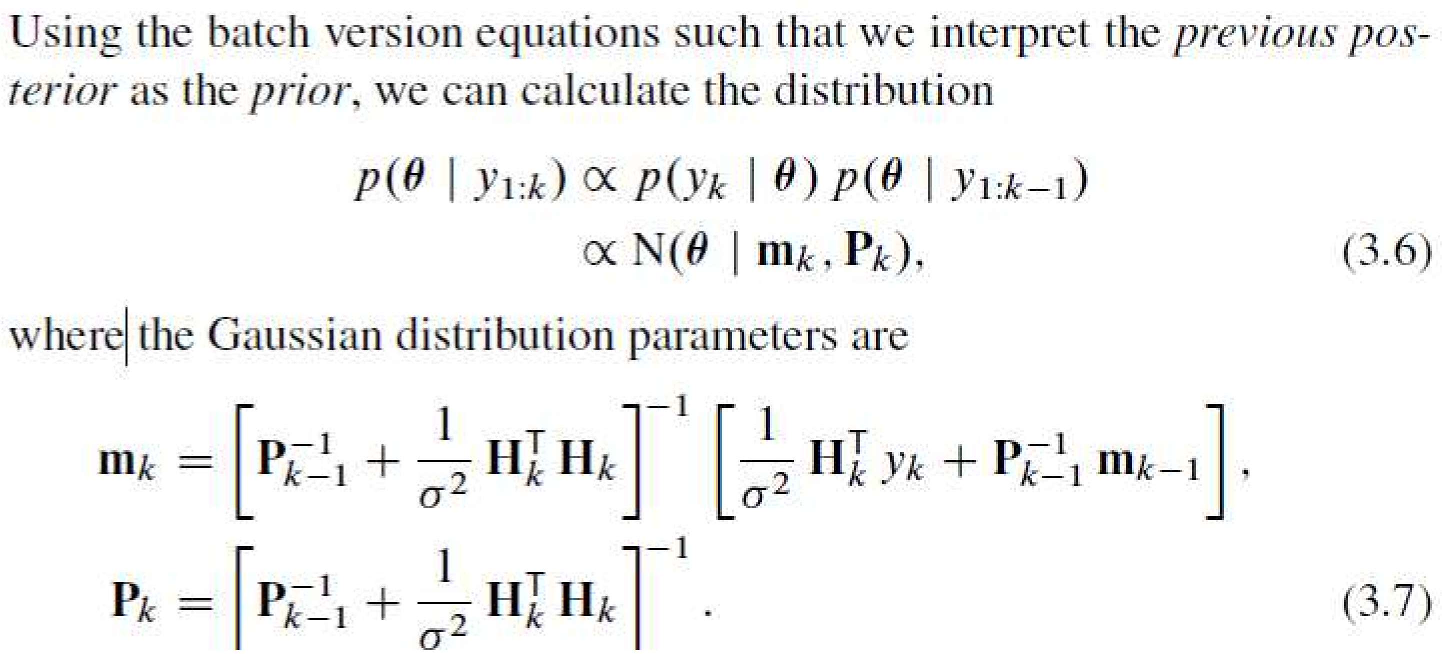
3.2 Recursive Line ar Regression



A *recursive solution* to the regression problem (3.2) can be obtained by assuming that we already have obtained the *posterior distribution* conditioned on the previous measurements $1, \ldots, k-1$ as follows:

$$p(\boldsymbol{\theta} \mid y_{1:k-1}) = \mathrm{N}(\boldsymbol{\theta} \mid \mathbf{m}_{k-1}, \mathbf{P}_{k-1}).$$

Now assume that we have obtained a new measurement y_k and we want to compute the posterior distribution of $\boldsymbol{\theta}$ given the old measurements $y_{1:k-1}$ *and* the new measurement y_k. According to the model specification the new measurement has the likelihood



Using the batch version equations such that we interpret the *previous posterior* as the *prior*, we can calculate the distribution

$$\begin{aligned} p(\boldsymbol{\theta} \mid y_{1:k}) &\propto p(y_k \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid y_{1:k-1}) \\ &\propto \mathrm{N}(\boldsymbol{\theta} \mid \mathbf{m}_k, \mathbf{P}_k), \end{aligned} \tag{3.6}$$

where the Gaussian distribution parameters are

$$\begin{aligned} \mathbf{m}_k &= \left[\mathbf{P}_{k-1}^{-1} + \frac{1}{\sigma^2}\mathbf{H}_k^\mathsf{T}\mathbf{H}_k\right]^{-1}\left[\frac{1}{\sigma^2}\mathbf{H}_k^\mathsf{T} y_k + \mathbf{P}_{k-1}^{-1}\mathbf{m}_{k-1}\right], \\ \mathbf{P}_k &= \left[\mathbf{P}_{k-1}^{-1} + \frac{1}{\sigma^2}\mathbf{H}_k^\mathsf{T}\mathbf{H}_k\right]^{-1}. \end{aligned} \tag{3.7}$$

By using the *matrix inversion lemma*, the covariance calculation can be written as

$$\mathbf{P}_k = \mathbf{P}_{k-1} - \mathbf{P}_{k-1}\,\mathbf{H}_k^\mathsf{T}\left[\mathbf{H}_k\,\mathbf{P}_{k-1}\,\mathbf{H}_k^\mathsf{T} + \sigma^2\right]^{-1}\mathbf{H}_k\,\mathbf{P}_{k-1}.$$

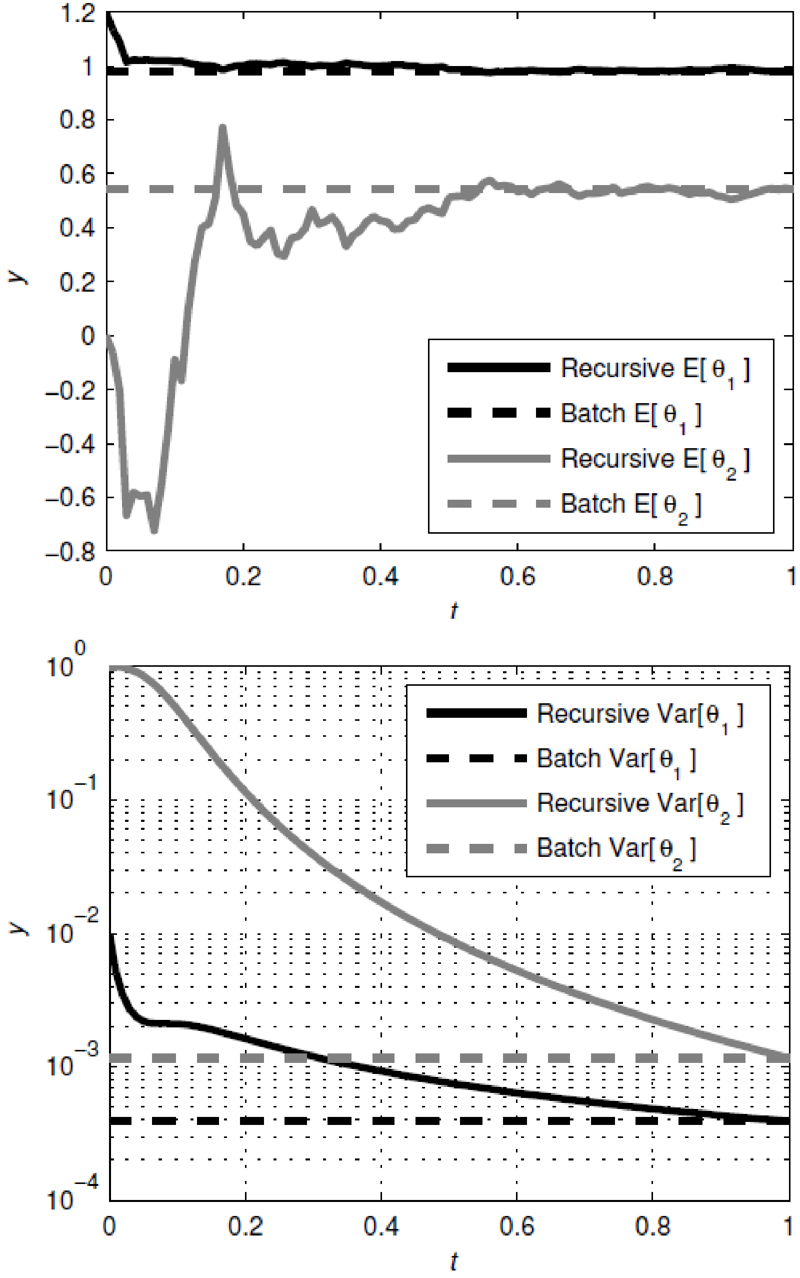
By introducing temporary variables S_k and $\mathbf{K}_k$ the calculation of the mean and covariance can be written in the form

$$\begin{aligned} S_k &= \mathbf{H}_k\,\mathbf{P}_{k-1}\,\mathbf{H}_k^\mathsf{T} + \sigma^2, \\ \mathbf{K}_k &= \mathbf{P}_{k-1}\,\mathbf{H}_k^\mathsf{T}\,S_k^{-1}, \\ \mathbf{m}_k &= \mathbf{m}_{k-1} + \mathbf{K}_k\,[y_k - \mathbf{H}_k\,\mathbf{m}_{k-1}], \\ \mathbf{P}_k &= \mathbf{P}_{k-1} - \mathbf{K}_k\,S_k\,\mathbf{K}_k^\mathsf{T}. \end{aligned} \tag{3.8}$$

Note that $S_k = \mathbf{H}_k\,\mathbf{P}_{k-1}\,\mathbf{H}_k^\mathsf{T} + \sigma^2$ is scalar because the measurements are scalar and thus no matrix inversion is required.

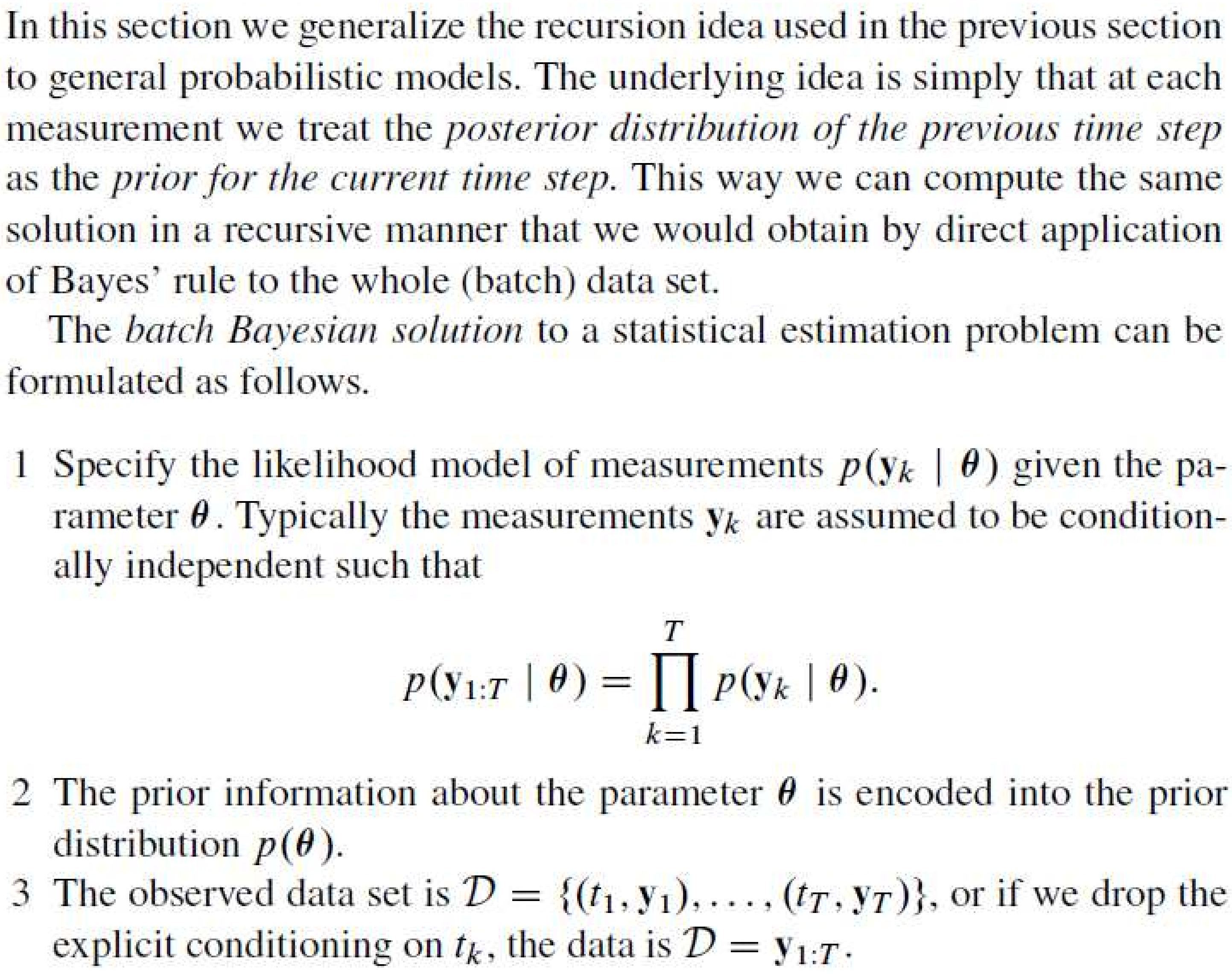
The equations above actually are special cases of the Kalman filter update equations. Only the update part of the equations (as opposed to the prediction and update) is required, because the estimated parameters are assumed to be constant, that is, there is no stochastic dynamics model for the parameters $\boldsymbol{\theta}$. Figures 3.3 and 3.4 illustrate the convergence of the means and variances of the parameters during the recursive estimation.

#


1.2
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
y
0
0.2
0.4
0.6
0.8
1
t
Recursive E[θ_1]
Batch E[θ_1]
Recursive E[θ_2]
Batch E[θ_2]
10^0
10^−1
10^−2
10^−3
10^−4
y
0
0.2
0.4
0.6
0.8
1
t
Recursive Var[θ_1]
Batch Var[θ_1]
Recursive Var[θ_2]
Batch Var[θ_2]

#

3.3 Block Versus Recursive Estimation



In this section we generalize the recursion idea used in the previous section to general probabilistic models. The underlying idea is simply that at each measurement we treat the *posterior distribution of the previous time step* as the *prior for the current time step*. This way we can compute the same solution in a recursive manner that we would obtain by direct application of Bayes' rule to the whole (batch) data set.

The *batch Bayesian solution* to a statistical estimation problem can be formulated as follows.

1 Specify the likelihood model of measurements $p(\mathbf{y}_k \mid \boldsymbol{\theta})$ given the parameter $\boldsymbol{\theta}$. Typically the measurements $\mathbf{y}_k$ are assumed to be conditionally independent such that

$$p(\mathbf{y}_{1:T} \mid \boldsymbol{\theta}) = \prod_{k=1}^{T} p(\mathbf{y}_k \mid \boldsymbol{\theta}).$$

2 The prior information about the parameter $\boldsymbol{\theta}$ is encoded into the prior distribution $p(\boldsymbol{\theta})$.

3 The observed data set is $\mathcal{D} = \{(t_1, \mathbf{y}_1), \ldots, (t_T, \mathbf{y}_T)\}$, or if we drop the explicit conditioning on t_k, the data is $\mathcal{D} = \mathbf{y}_{1:T}$.

#

4. The batch Bayesian solution to the statistical estimation problem can be computed by applying Bayes' rule:

$$p(\boldsymbol{\theta} \mid \mathbf{y}_{1:T}) = \frac{1}{Z} p(\boldsymbol{\theta}) \prod_{k=1}^{T} p(\mathbf{y}_k \mid \boldsymbol{\theta}),$$

   where Z is the *normalization constant*

$$Z = \int p(\boldsymbol{\theta}) \prod_{k=1}^{T} p(\mathbf{y}_k \mid \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}.$$

For example, the batch solution of the above kind to the linear regression problem (3.2) was given by Equations (3.3) and (3.4).

The *recursive Bayesian solution* to the above statistical estimation problem can be formulated as follows.

1. The distribution of measurements is again modeled by the likelihood function $p(\mathbf{y}_k \mid \boldsymbol{\theta})$ and the measurements are assumed to be conditionally independent.
2. In the beginning of estimation (i.e., at step 0), all the information about the parameter $\boldsymbol{\theta}$ we have is contained in the prior distribution $p(\boldsymbol{\theta})$.

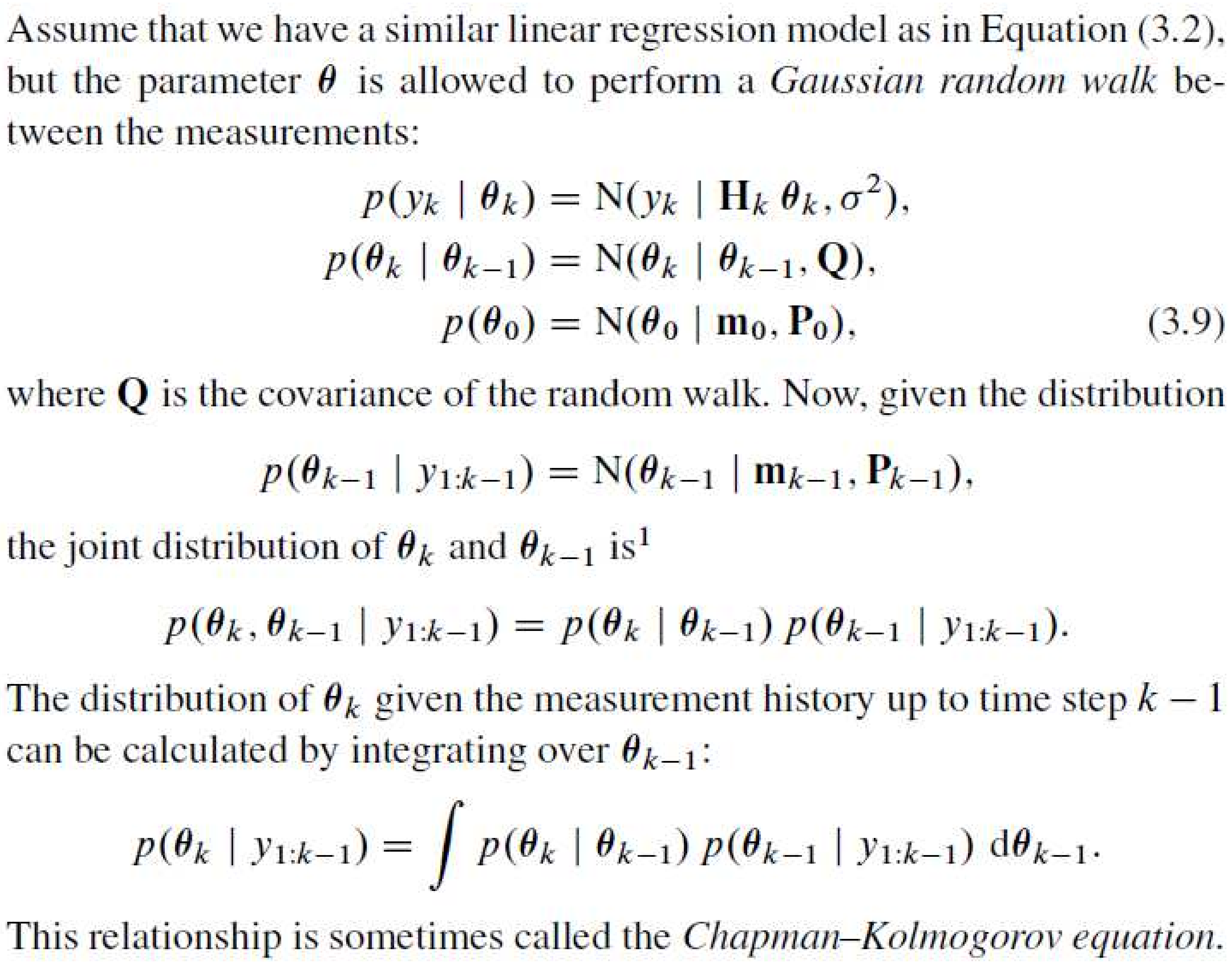
#

3 The measurements are assumed to be obtained one at a time, first $\mathbf{y}_1$, then $\mathbf{y}_2$ and so on. At each step we use the posterior distribution from the previous time step as the current prior distribution:

$$
\begin{aligned}
p(\theta \mid \mathbf{y}_1) &= \frac{1}{Z_1} p(\mathbf{y}_1 \mid \theta) p(\theta), \\
p(\theta \mid \mathbf{y}_{1:2}) &= \frac{1}{Z_2} p(\mathbf{y}_2 \mid \theta) p(\theta \mid \mathbf{y}_1), \\
p(\theta \mid \mathbf{y}_{1:3}) &= \frac{1}{Z_3} p(\mathbf{y}_3 \mid \theta) p(\theta \mid \mathbf{y}_{1:2}), \\
&\vdots \\
p(\theta \mid \mathbf{y}_{1:T}) &= \frac{1}{Z_T} p(\mathbf{y}_T \mid \theta) p(\theta \mid \mathbf{y}_{1:T-1}).
\end{aligned}
$$

#

3.4 Drift Model for Linear Regression



Assume that we have a similar linear regression model as in Equation (3.2), but the parameter $\boldsymbol{\theta}$ is allowed to perform a *Gaussian random walk* between the measurements:

$$\begin{aligned} p(y_k \mid \boldsymbol{\theta}_k) &= \mathrm{N}(y_k \mid \mathbf{H}_k\, \boldsymbol{\theta}_k, \sigma^2), \\ p(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_{k-1}) &= \mathrm{N}(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_{k-1}, \mathbf{Q}), \\ p(\boldsymbol{\theta}_0) &= \mathrm{N}(\boldsymbol{\theta}_0 \mid \mathbf{m}_0, \mathbf{P}_0), \end{aligned} \tag{3.9}$$

where $\mathbf{Q}$ is the covariance of the random walk. Now, given the distribution

$$p(\boldsymbol{\theta}_{k-1} \mid y_{1:k-1}) = \mathrm{N}(\boldsymbol{\theta}_{k-1} \mid \mathbf{m}_{k-1}, \mathbf{P}_{k-1}),$$

the joint distribution of $\boldsymbol{\theta}_k$ and $\boldsymbol{\theta}_{k-1}$ is[1]

$$p(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{k-1} \mid y_{1:k-1}) = p(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_{k-1})\, p(\boldsymbol{\theta}_{k-1} \mid y_{1:k-1}).$$

The distribution of $\boldsymbol{\theta}_k$ given the measurement history up to time step $k-1$ can be calculated by integrating over $\boldsymbol{\theta}_{k-1}$:

$$p(\boldsymbol{\theta}_k \mid y_{1:k-1}) = \int p(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_{k-1})\, p(\boldsymbol{\theta}_{k-1} \mid y_{1:k-1})\, \mathrm{d}\boldsymbol{\theta}_{k-1}.$$

This relationship is sometimes called the *Chapman–Kolmogorov equation*.

#

Because $p(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_{k-1})$ and $p(\boldsymbol{\theta}_{k-1} \mid y_{1:k-1})$ are Gaussian, the result of the marginalization is Gaussian,

$$p(\boldsymbol{\theta}_k \mid y_{1:k-1}) = \mathrm{N}(\boldsymbol{\theta}_k \mid \mathbf{m}_k^-, \mathbf{P}_k^-),$$

where

$$\begin{aligned}
\mathbf{m}_k^- &= \mathbf{m}_{k-1}, \\
\mathbf{P}_k^- &= \mathbf{P}_{k-1} + \mathbf{Q}.
\end{aligned}$$

By using this as the prior distribution for the measurement likelihood $p(y_k \mid \boldsymbol{\theta}_k)$ we get the parameters of the posterior distribution

$$p(\boldsymbol{\theta}_k \mid y_{1:k}) = \mathrm{N}(\boldsymbol{\theta}_k \mid \mathbf{m}_k, \mathbf{P}_k),$$

which are given by Equations (3.8), when $\mathbf{m}_{k-1}$ and $\mathbf{P}_{k-1}$ are replaced by $\mathbf{m}_k^-$ and $\mathbf{P}_k^-$:

$$\begin{aligned}
S_k &= \mathbf{H}_k \, \mathbf{P}_k^- \, \mathbf{H}_k^\mathsf{T} + \sigma^2, \\
\mathbf{K}_k &= \mathbf{P}_k^- \, \mathbf{H}_k^\mathsf{T} \, S_k^{-1}, \\
\mathbf{m}_k &= \mathbf{m}_k^- + \mathbf{K}_k \, [y_k - \mathbf{H}_k \, \mathbf{m}_k^-], \\
\mathbf{P}_k &= \mathbf{P}_k^- - \mathbf{K}_k \, S_k \, \mathbf{K}_k^\mathsf{T}.
\end{aligned} \tag{3.10}$$

#

This recursive computational algorithm for the time-varying linear regression weights is again a special case of the Kalman filter algorithm. Figure 3.5 shows the result of recursive estimation of a sine signal assuming a small diagonal Gaussian drift model for the parameters.

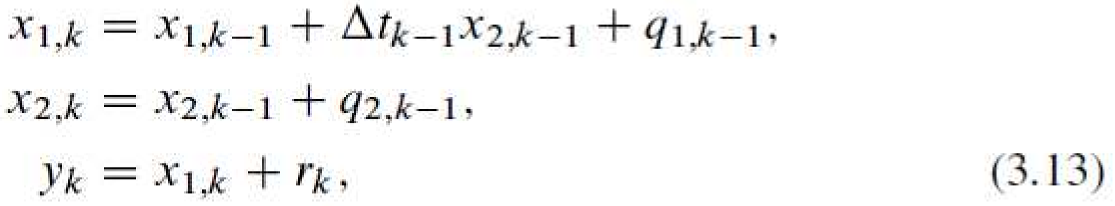
At this point we change from the *regression notation* used so far into *state space model notation*, which is commonly used in Kalman filtering and related dynamic estimation literature. Because this notation easily causes confusion to people who have got used to regression notation, this point is emphasized.

- In *state space notation* $\mathbf{x}$ means the unknown state of the system, that is, the vector of *unknown parameters in the system*. It is *not* the regressor, covariate or input variable of the system.
- For example, the time-varying linear regression model with drift presented in this section can be transformed into the more standard *state space model notation* by replacing the variable $\boldsymbol{\theta}_k = (\theta_{1,k} \; \theta_{2,k})^\mathsf{T}$ with the variable $\mathbf{x}_k = (x_{1,k} \; x_{2,k})^\mathsf{T}$:

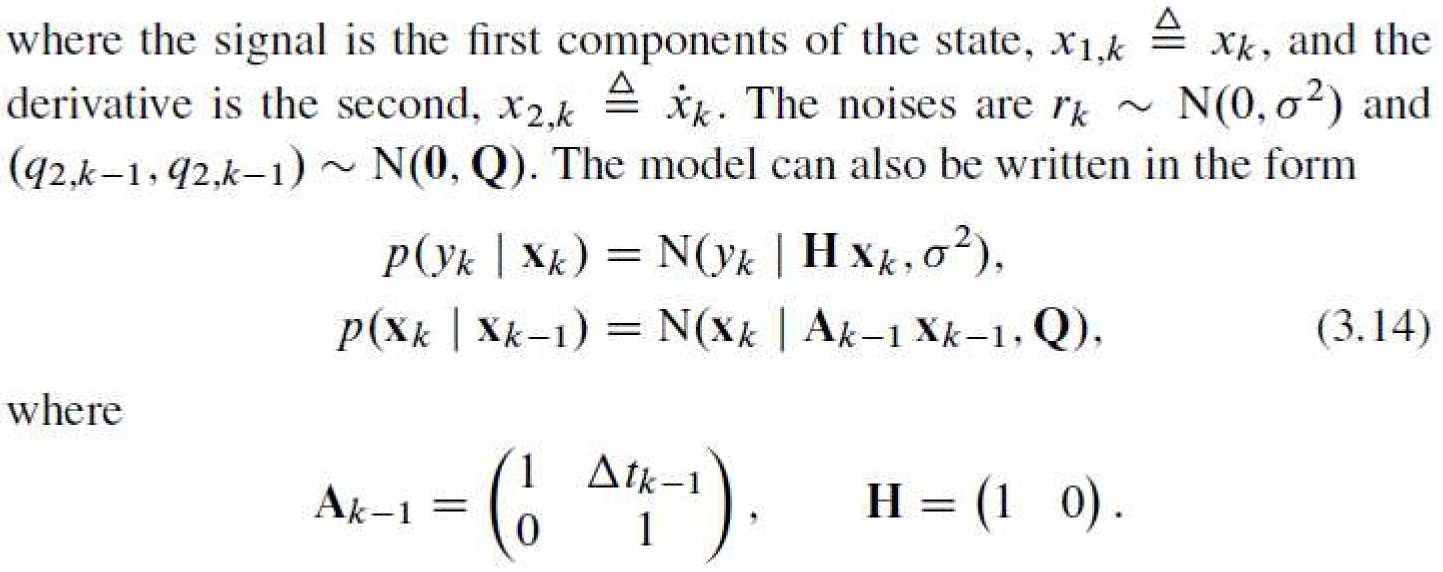
$$\begin{aligned}
p(y_k \mid \mathbf{x}_k) &= \mathrm{N}(y_k \mid \mathbf{H}_k \, \mathbf{x}_k, \sigma^2), \\
p(\mathbf{x}_k \mid \mathbf{x}_{k-1}) &= \mathrm{N}(\mathbf{x}_k \mid \mathbf{x}_{k-1}, \mathbf{Q}), \\
p(\mathbf{x}_0) &= \mathrm{N}(\mathbf{x}_0 \mid \mathbf{m}_0, \mathbf{P}_0).
\end{aligned} \tag{3.11}$$

3.5 State Space Model for Linear Regression with Drift

This model can be written as follows:



$$
\begin{aligned}
x_{1,k} &= x_{1,k-1} + \Delta t_{k-1} x_{2,k-1} + q_{1,k-1}, \\
x_{2,k} &= x_{2,k-1} + q_{2,k-1}, \\
y_k &= x_{1,k} + r_k,
\end{aligned}
\tag{3.13}
$$



where the signal is the first components of the state, $x_{1,k} \triangleq x_k$, and the derivative is the second, $x_{2,k} \triangleq \dot{x}_k$. The noises are $r_k \sim \mathrm{N}(0, \sigma^2)$ and $(q_{2,k-1}, q_{2,k-1}) \sim \mathrm{N}(\mathbf{0}, \mathbf{Q})$. The model can also be written in the form

$$
\begin{aligned}
p(y_k \mid \mathbf{x}_k) &= \mathrm{N}(y_k \mid \mathbf{H}\,\mathbf{x}_k, \sigma^2), \\
p(\mathbf{x}_k \mid \mathbf{x}_{k-1}) &= \mathrm{N}(\mathbf{x}_k \mid \mathbf{A}_{k-1}\,\mathbf{x}_{k-1}, \mathbf{Q}),
\end{aligned}
\tag{3.14}
$$

where

$$
\mathbf{A}_{k-1} = \begin{pmatrix} 1 & \Delta t_{k-1} \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{H} = \begin{pmatrix} 1 & 0 \end{pmatrix}.
$$

#

We could now explicitly derive the recursion equations in the same manner as we did in the previous sections. However, we can also use the *Kalman filter*, which is a readily derived recursive solution to generic linear Gaussian models of the form

$$\begin{aligned} p(\mathbf{y}_k \mid \mathbf{x}_k) &= \mathrm{N}(\mathbf{y}_k \mid \mathbf{H}_k \, \mathbf{x}_k, \mathbf{R}_k), \\ p(\mathbf{x}_k \mid \mathbf{x}_{k-1}) &= \mathrm{N}(\mathbf{x}_k \mid \mathbf{A}_{k-1} \, \mathbf{x}_{k-1}, \mathbf{Q}_{k-1}). \end{aligned}$$

Our alternative linear regression model in Equation (3.13) can be seen to be a special case of these models. The *Kalman filter equations* are often expressed as *prediction and update steps* as follows.

1 *Prediction step:*

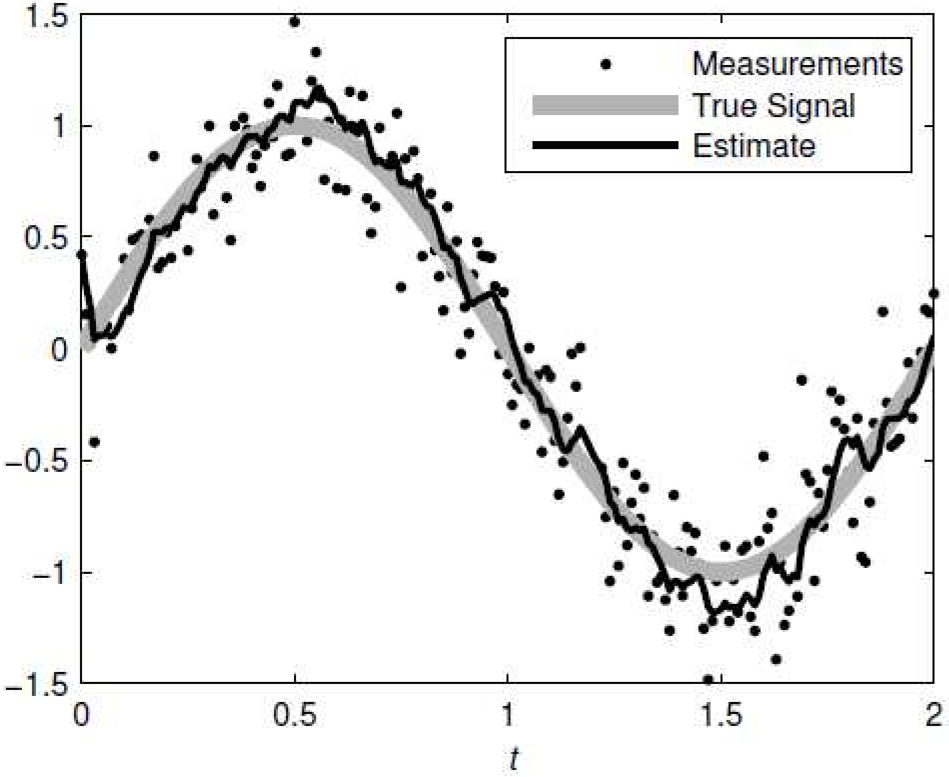
$$\begin{aligned} \mathbf{m}_k^- &= \mathbf{A}_{k-1} \, \mathbf{m}_{k-1}, \\ \mathbf{P}_k^- &= \mathbf{A}_{k-1} \, \mathbf{P}_{k-1} \, \mathbf{A}_{k-1}^\mathsf{T} + \mathbf{Q}_{k-1}. \end{aligned}$$

2 *Update step:*

$$\begin{aligned} \mathbf{S}_k &= \mathbf{H}_k \, \mathbf{P}_k^- \, \mathbf{H}_k^\mathsf{T} + \mathbf{R}_k, \\ \mathbf{K}_k &= \mathbf{P}_k^- \, \mathbf{H}_k^\mathsf{T} \, \mathbf{S}_k^{-1}, \\ \mathbf{m}_k &= \mathbf{m}_k^- + \mathbf{K}_k \, [\mathbf{y}_k - \mathbf{H}_k \, \mathbf{m}_k^-], \\ \mathbf{P}_k &= \mathbf{P}_k^- - \mathbf{K}_k \, \mathbf{S}_k \, \mathbf{K}_k^\mathsf{T}. \end{aligned}$$

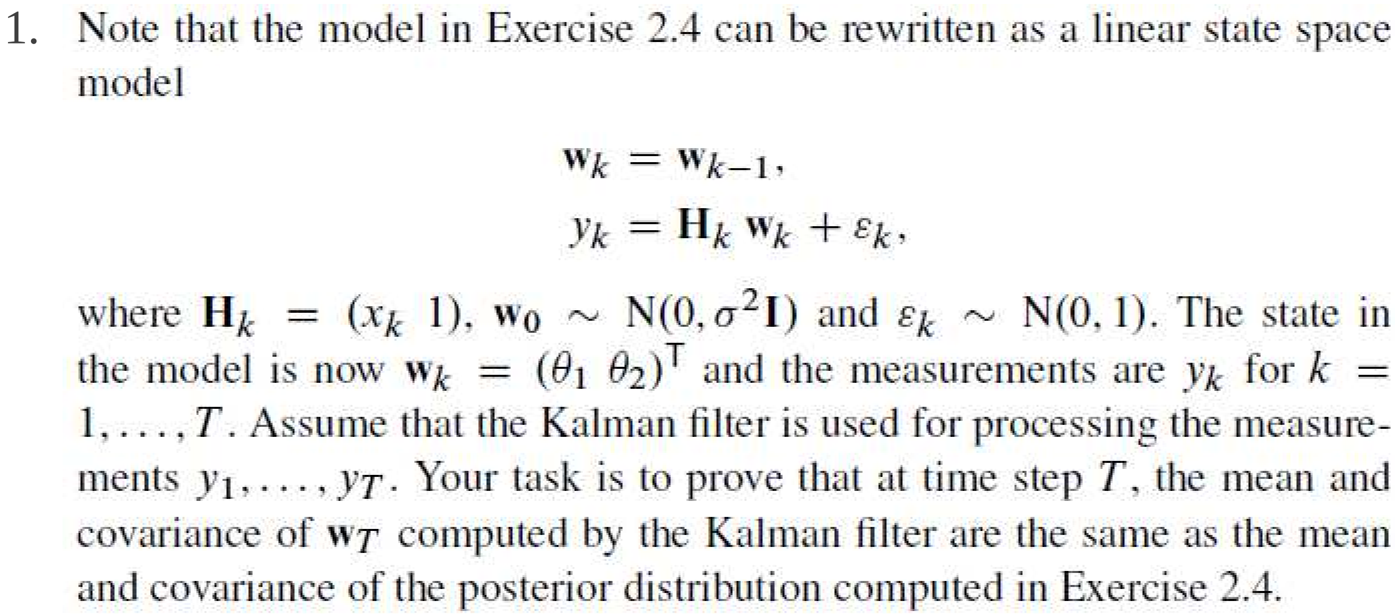
The result of tracking the sine signal with Kalman filter is shown in Figure 3.6. All the mean and covariance calculation equations given in this book so far have been special cases of the above equations, including the batch solution to the scalar measurement case (which is a one-step solution). The Kalman filter recursively computes the mean and covariance of the posterior distributions of the form

$$p(\mathbf{x}_k \mid \mathbf{y}_{1:k}) = \mathrm{N}(\mathbf{x}_k \mid \mathbf{m}_k, \mathbf{P}_k).$$



#

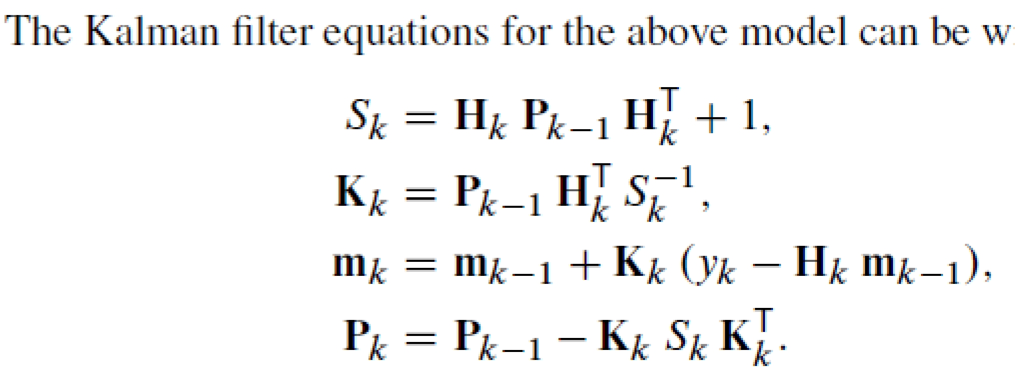
EXERCISES



1. Note that the model in Exercise 2.4 can be rewritten as a linear state space model

$$\begin{aligned} \mathbf{w}_k &= \mathbf{w}_{k-1}, \\ y_k &= \mathbf{H}_k\, \mathbf{w}_k + \varepsilon_k, \end{aligned}$$

where $\mathbf{H}_k = (x_k \;\; 1)$, $\mathbf{w}_0 \sim \mathrm{N}(0, \sigma^2 \mathbf{I})$ and $\varepsilon_k \sim \mathrm{N}(0, 1)$. The state in the model is now $\mathbf{w}_k = (\theta_1 \;\; \theta_2)^\mathsf{T}$ and the measurements are y_k for $k = 1, \ldots, T$. Assume that the Kalman filter is used for processing the measurements $y_1, \ldots, y_T$. Your task is to prove that at time step T, the mean and covariance of $\mathbf{w}_T$ computed by the Kalman filter are the same as the mean and covariance of the posterior distribution computed in Exercise 2.4.



The Kalman filter equations for the above model can be written as:

$$\begin{aligned} S_k &= \mathbf{H}_k\, \mathbf{P}_{k-1}\, \mathbf{H}_k^\mathsf{T} + 1, \\ \mathbf{K}_k &= \mathbf{P}_{k-1}\, \mathbf{H}_k^\mathsf{T}\, S_k^{-1}, \\ \mathbf{m}_k &= \mathbf{m}_{k-1} + \mathbf{K}_k\, (y_k - \mathbf{H}_k\, \mathbf{m}_{k-1}), \\ \mathbf{P}_k &= \mathbf{P}_{k-1} - \mathbf{K}_k\, S_k\, \mathbf{K}_k^\mathsf{T}. \end{aligned}$$

#

(a) Write formulas for the posterior mean $\mathbf{m}_{k-1}$ and covariance $\mathbf{P}_{k-1}$ assuming that they are the same as those which would be obtained if the pairs $\{(x_i, y_i) \ : \ i = 1, \ldots, k-1\}$ were (batch) processed as in Exercise 2.4. Write similar equations for the mean $\mathbf{m}_k$ and covariance $\mathbf{P}_k$. Show that the posterior means can be expressed in the form

$$\begin{aligned} \mathbf{m}_{k-1} &= \mathbf{P}_{k-1}\,\mathbf{X}_{k-1}^{\mathsf{T}}\,\mathbf{y}_{k-1}, \\ \mathbf{m}_k &= \mathbf{P}_k\,\mathbf{X}_k^{\mathsf{T}}\,\mathbf{y}_k, \end{aligned}$$

where $\mathbf{X}_{k-1}$ and $\mathbf{y}_{k-1}$ have been constructed as $\mathbf{X}$ and $\mathbf{y}$ in Exercise 2.4, except that only the pairs $\{(x_i, y_i) \ : \ i = 1, \ldots, k-1\}$ have been used, and $\mathbf{X}_k$ and $\mathbf{y}_k$ have been constructed similarly from pairs up to the step k.

(b) Rewrite the expressions $\mathbf{X}_k^{\mathsf{T}}\,\mathbf{X}_k$ and $\mathbf{X}_k^{\mathsf{T}}\,\mathbf{y}_k$ in terms of $\mathbf{X}_{k-1}$, $\mathbf{y}_{k-1}$, $\mathbf{H}_k$ and y_k. Substitute these into the expressions of $\mathbf{m}_k$ and $\mathbf{P}_k$ obtained in (a).

(c) Expand the expression of the covariance $\mathbf{P}_k = \mathbf{P}_{k-1} - \mathbf{K}_k\,S_k\,\mathbf{K}_k^{\mathsf{T}}$ by substituting the expressions for $\mathbf{K}_k$ and S_k. Convert it to a simpler form by applying the matrix inversion lemma

$$\mathbf{P}_{k-1} - \mathbf{P}_{k-1}\,\mathbf{H}_k^{\mathsf{T}}\,(\mathbf{H}_k\,\mathbf{P}_{k-1}\,\mathbf{H}_k^{\mathsf{T}} + 1)^{-1}\,\mathbf{H}_k\,\mathbf{P}_{k-1} = (\mathbf{P}_{k-1}^{-1} + \mathbf{H}_k^{\mathsf{T}}\,\mathbf{H}_k)^{-1}.$$

Show that this expression for $\mathbf{P}_k$ is equivalent to the expression in (a).

(d) Expand the expression of the mean $\mathbf{m}_k = \mathbf{m}_{k-1} + \mathbf{K}_k\,(y_k - \mathbf{H}_k\,\mathbf{m}_{k-1})$ and show that the result is equivalent to the expression obtained in (a).
‡
Hint: the Kalman gain can also be written as $\mathbf{K}_k = \mathbf{P}_k\,\mathbf{H}_k^{\mathsf{T}}$.

2. Recall that the Gaussian probability density is defined as

$$\mathrm{N}(\mathbf{x} \mid \mathbf{m}, \mathbf{P}) = \frac{1}{(2\,\pi)^{n/2}\,|\mathbf{P}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathsf{T}}\,\mathbf{P}^{-1}\,(\mathbf{x}-\mathbf{m})\right).$$

Derive the following Gaussian identities.

(a) Let **x** and **y** have the Gaussian densities

$$p(\mathbf{x}) = \mathrm{N}(\mathbf{x} \mid \mathbf{m}, \mathbf{P}), \qquad p(\mathbf{y} \mid \mathbf{x}) = \mathrm{N}(\mathbf{y} \mid \mathbf{H}\,\mathbf{x}, \mathbf{R}),$$

then the joint distribution of **x** and **y** is

$$\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mathbf{m} \\ \mathbf{H}\,\mathbf{m} \end{pmatrix}, \begin{pmatrix} \mathbf{P} & \mathbf{P}\,\mathbf{H}^{\mathsf{T}} \\ \mathbf{H}\,\mathbf{P} & \mathbf{H}\,\mathbf{P}\,\mathbf{H}^{\mathsf{T}} + \mathbf{R} \end{pmatrix}\right)$$

and the marginal distribution of **y** is

$$\mathbf{y} \sim \mathrm{N}(\mathbf{H}\,\mathbf{m}, \mathbf{H}\,\mathbf{P}\,\mathbf{H}^{\mathsf{T}} + \mathbf{R}).$$

Hint: use the properties of expectation $\mathrm{E}[\mathbf{H}\,\mathbf{x} + \mathbf{r}] = \mathbf{H}\ \mathrm{E}[\mathbf{x}] + \mathrm{E}[\mathbf{r}]$ and $\mathrm{Cov}[\mathbf{H}\,\mathbf{x} + \mathbf{r}] = \mathbf{H}\ \mathrm{Cov}[\mathbf{x}]\,\mathbf{H}^{\mathsf{T}} + \mathrm{Cov}[\mathbf{r}]$ (if **x** and **r** are independent).

(b) Write down the explicit expression for the joint and marginal probability densities above:

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y} \mid \mathbf{x})\; p(\mathbf{x}) = ?$$

$$p(\mathbf{y}) = \int p(\mathbf{y} \mid \mathbf{x})\; p(\mathbf{x})\; d\mathbf{x} = ?$$

#

(c) If the random variables $\mathbf{x}$ and $\mathbf{y}$ have the joint Gaussian probability density

$$\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}, \begin{pmatrix}\mathbf{A} & \mathbf{C}\\ \mathbf{C}^{\mathsf{T}} & \mathbf{B}\end{pmatrix}\right),$$

then the conditional density of $\mathbf{x}$ given $\mathbf{y}$ is

$$\mathbf{x} \mid \mathbf{y} \sim \mathrm{N}(\mathbf{a} + \mathbf{C}\,\mathbf{B}^{-1}\,(\mathbf{y} - \mathbf{b}), \mathbf{A} - \mathbf{C}\,\mathbf{B}^{-1}\mathbf{C}^{\mathsf{T}}).$$

Hints:

- Denote inverse covariance as $\mathbf{D} = \begin{pmatrix}\mathbf{D}_{11} & \mathbf{D}_{12}\\ \mathbf{D}_{12}^{\mathsf{T}} & \mathbf{D}_{22}\end{pmatrix}$ and expand the quadratic form in the Gaussian exponent.
- Compute the derivative with respect to $\mathbf{x}$ and set it to zero. Conclude that due to symmetry the point where the derivative vanishes is the mean.
- Check from a linear algebra book that the inverse of $\mathbf{D}_{11}$ is given by the Schur complement:

$$\mathbf{D}_{11}^{-1} = \mathbf{A} - \mathbf{C}\,\mathbf{B}^{-1}\,\mathbf{C}^{\mathsf{T}}$$

  and that $\mathbf{D}_{12}$ can be then written as

$$\mathbf{D}_{12} = -\mathbf{D}_{11}\,\mathbf{C}\,\mathbf{B}^{-1}.$$

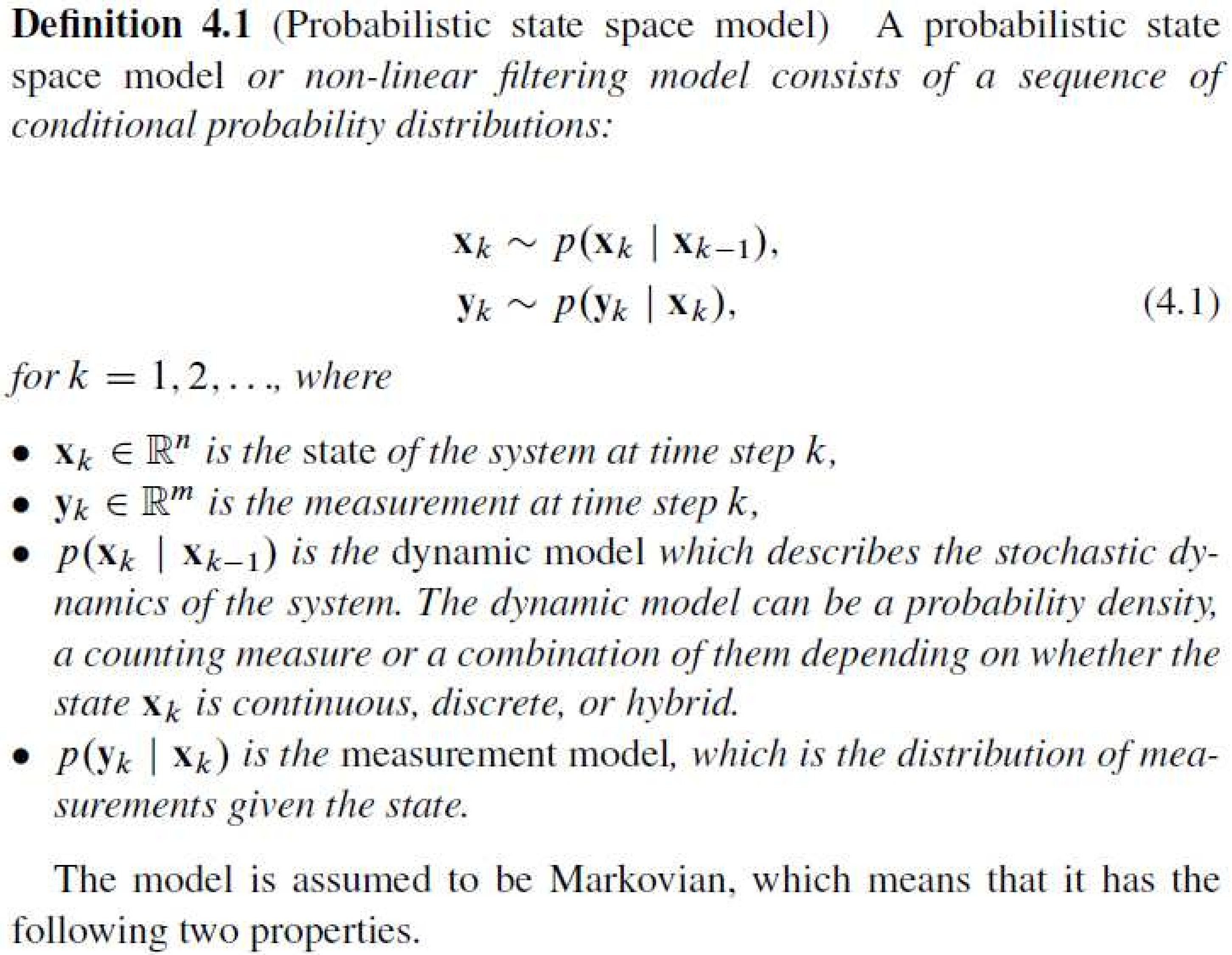
- Find the simplified expression for the mean by applying the identities above.

#

- Find the second derivative of the negative Gaussian exponent with respect to $\mathbf{x}$. Conclude that it must be the inverse conditional covariance of $\mathbf{x}$.
- Use the Schur complement expression above for computing the conditional covariance.

4. EXACT SOLUTIONS TO THE BAYESIAN FILTERING PROBLEM

4.1 State Space Models



Definition 4.1 (Probabilistic state space model) A probabilistic state space model *or non-linear filtering model consists of a sequence of conditional probability distributions:*

$$\begin{aligned} \mathbf{x}_k &\sim p(\mathbf{x}_k \mid \mathbf{x}_{k-1}), \\ \mathbf{y}_k &\sim p(\mathbf{y}_k \mid \mathbf{x}_k), \end{aligned} \tag{4.1}$$

for $k = 1, 2, \ldots,$ *where*

- $\mathbf{x}_k \in \mathbb{R}^n$ *is the* state *of the system at time step* k,
- $\mathbf{y}_k \in \mathbb{R}^m$ *is the measurement at time step* k,
- $p(\mathbf{x}_k \mid \mathbf{x}_{k-1})$ *is the* dynamic model *which describes the stochastic dynamics of the system. The dynamic model can be a probability density, a counting measure or a combination of them depending on whether the state* $\mathbf{x}_k$ *is continuous, discrete, or hybrid.*
- $p(\mathbf{y}_k \mid \mathbf{x}_k)$ *is the* measurement model, *which is the distribution of measurements given the state.*

The model is assumed to be Markovian, which means that it has the following two properties.

Property 4.1 (Markov property of states)
The states $\{\mathbf{x}_k : k = 0, 1, 2, \ldots\}$ *form a Markov sequence (or Markov chain if the state is discrete). This* Markov property *means that* $\mathbf{x}_k$ *(and actually the whole future* $\mathbf{x}_{k+1}, \mathbf{x}_{k+2}, \ldots$*) given* $\mathbf{x}_{k-1}$ *is independent of anything that has happened before the time step* $k - 1$*:*

$$p(\mathbf{x}_k \mid \mathbf{x}_{1:k-1}, \mathbf{y}_{1:k-1}) = p(\mathbf{x}_k \mid \mathbf{x}_{k-1}). \tag{4.2}$$

Also the past is independent of the future given the present:

$$p(\mathbf{x}_{k-1} \mid \mathbf{x}_{k:T}, \mathbf{y}_{k:T}) = p(\mathbf{x}_{k-1} \mid \mathbf{x}_k). \tag{4.3}$$

Property 4.2 (Conditional independence of measurements)
The current measurement $\mathbf{y}_k$ *given the current state* $\mathbf{x}_k$ *is conditionally independent of the measurement and state histories:*

$$p(\mathbf{y}_k \mid \mathbf{x}_{1:k}, \mathbf{y}_{1:k-1}) = p(\mathbf{y}_k \mid \mathbf{x}_k). \tag{4.4}$$

A simple example of a Markovian sequence is the Gaussian random walk. When this is combined with noisy measurements, we obtain the following example of a probabilistic state space model.

Example 4.1 (Gaussian random walk) *A Gaussian random walk model can be written as*

#

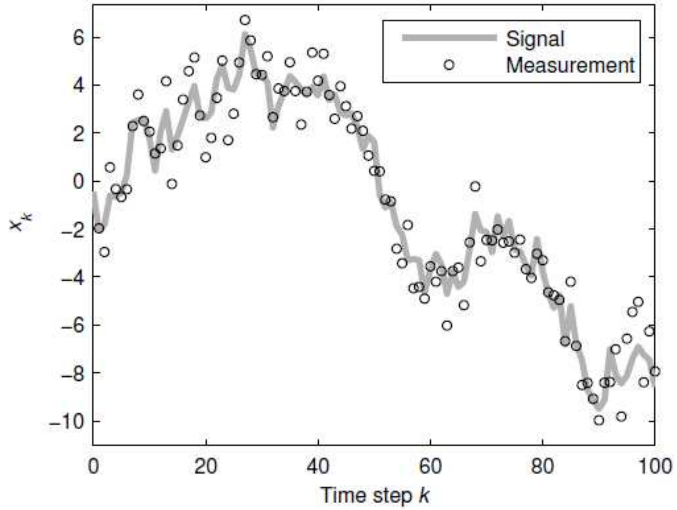
$$\begin{aligned} x_k &= x_{k-1} + q_{k-1}, \quad & q_{k-1} &\sim \mathrm{N}(0, Q), \\ y_k &= x_k + r_k, & r_k &\sim \mathrm{N}(0, R), \end{aligned} \tag{4.5}$$

where x_k *is the hidden state (or signal) and* y_k *is the measurement. In terms of probability densities the model can be written as*

$$\begin{aligned}
p(x_k \mid x_{k-1}) &= \mathrm{N}(x_k \mid x_{k-1}, Q) \\
&= \frac{1}{\sqrt{2\pi Q}} \exp\left(-\frac{1}{2Q}(x_k - x_{k-1})^2\right), \\
p(y_k \mid x_k) &= \mathrm{N}(y_k \mid x_k, R) \\
&= \frac{1}{\sqrt{2\pi R}} \exp\left(-\frac{1}{2R}(y_k - x_k)^2\right),
\end{aligned} \tag{4.6}$$

which is a probabilistic state space model. Example realizations of the signal x_k *and measurements* y_k *are shown in Figure 4.1. The parameter values in the simulation were* $Q = R = 1$.

#



With the Markovian assumption and the filtering model (4.1), the *joint prior distribution of the states* $\mathbf{x}_{0:T} = \{\mathbf{x}_0, \ldots, \mathbf{x}_T\}$, and the *joint likelihood of the measurements* $\mathbf{y}_{1:T} = \{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$ are, respectively,

$$p(\mathbf{x}_{0:T}) = p(\mathbf{x}_0) \prod_{k=1}^{T} p(\mathbf{x}_k \mid \mathbf{x}_{k-1}), \tag{4.7}$$

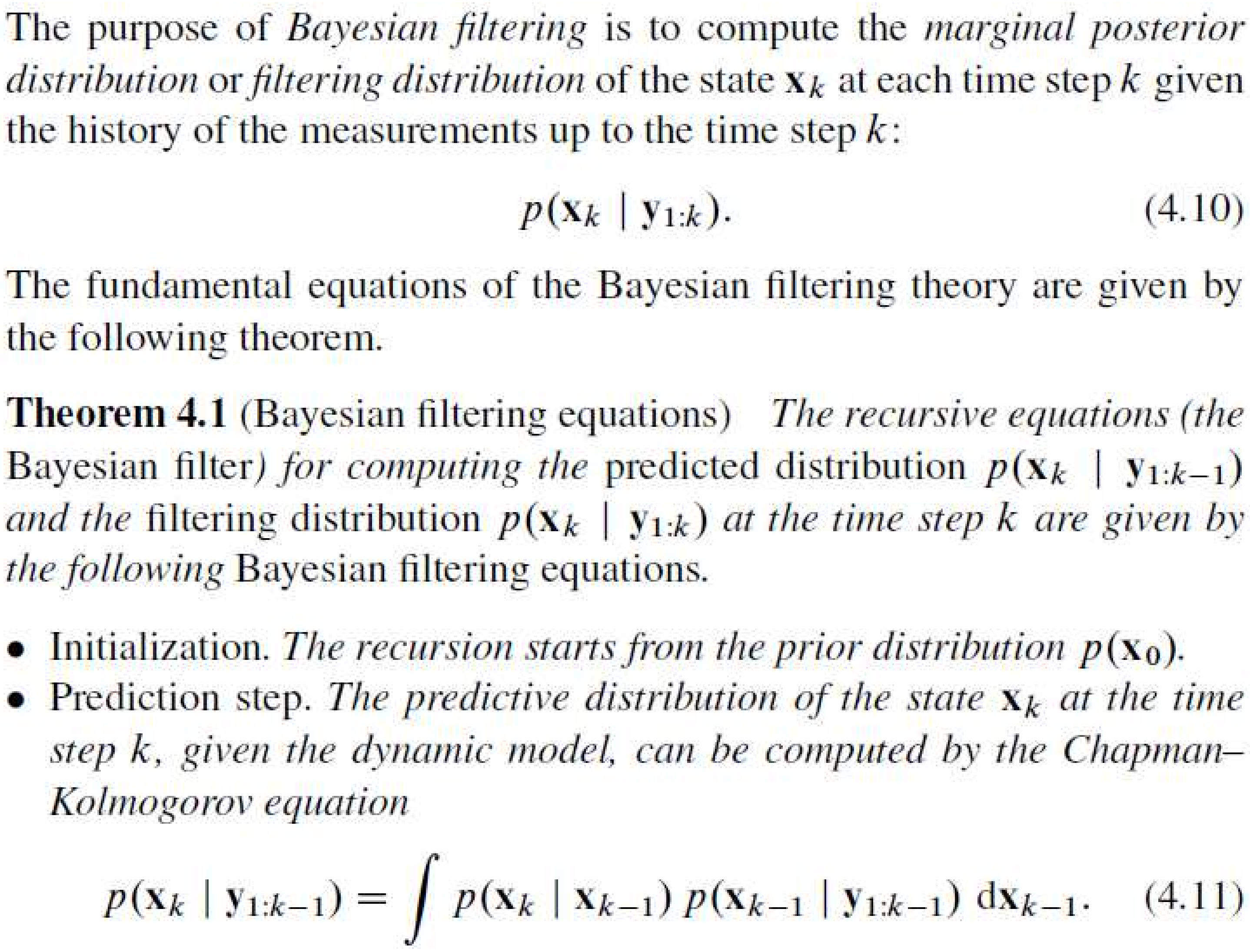
$$p(\mathbf{y}_{1:T} \mid \mathbf{x}_{0:T}) = \prod_{k=1}^{T} p(\mathbf{y}_k \mid \mathbf{x}_k). \tag{4.8}$$

In principle, for a given T we could simply compute the posterior distribution of the states by Bayes' rule:

$$\begin{aligned} p(\mathbf{x}_{0:T} \mid \mathbf{y}_{1:T}) &= \frac{p(\mathbf{y}_{1:T} \mid \mathbf{x}_{0:T})\, p(\mathbf{x}_{0:T})}{p(\mathbf{y}_{1:T})} \\ &\propto p(\mathbf{y}_{1:T} \mid \mathbf{x}_{0:T})\, p(\mathbf{x}_{0:T}). \end{aligned} \tag{4.9}$$

However, this kind of explicit usage of the full Bayes' rule is not feasible in real-time applications, because the number of computations per time step increases as new observations arrive. Thus, this way we could only work with small data sets, because if the amount of data is unbounded (as in real-time sensing applications), then at some point of time the computations will become intractable. To cope with real-time data we need to have an # algorithm which does a constant number of computations per time step.

4.2 Filtering Equations



The purpose of *Bayesian filtering* is to compute the *marginal posterior distribution* or *filtering distribution* of the state $\mathbf{x}_k$ at each time step k given the history of the measurements up to the time step k:

$$p(\mathbf{x}_k \mid \mathbf{y}_{1:k}). \tag{4.10}$$

The fundamental equations of the Bayesian filtering theory are given by the following theorem.

Theorem 4.1 (Bayesian filtering equations) *The recursive equations (the* Bayesian filter*) for computing the* predicted distribution $p(\mathbf{x}_k \mid \mathbf{y}_{1:k-1})$ *and the* filtering distribution $p(\mathbf{x}_k \mid \mathbf{y}_{1:k})$ *at the time step k are given by the following* Bayesian filtering equations.

- Initialization. *The recursion starts from the prior distribution* $p(\mathbf{x}_0)$.
- Prediction step. *The predictive distribution of the state* $\mathbf{x}_k$ *at the time step k, given the dynamic model, can be computed by the Chapman–Kolmogorov equation*

$$p(\mathbf{x}_k \mid \mathbf{y}_{1:k-1}) = \int p(\mathbf{x}_k \mid \mathbf{x}_{k-1})\, p(\mathbf{x}_{k-1} \mid \mathbf{y}_{1:k-1})\, \mathrm{d}\mathbf{x}_{k-1}. \tag{4.11}$$

#

- Update step. *Given the measurement* $\mathbf{y}_k$ *at time step* k *the posterior distribution of the state* $\mathbf{x}_k$ *can be computed by Bayes' rule*

$$p(\mathbf{x}_k \mid \mathbf{y}_{1:k}) = \frac{1}{Z_k} p(\mathbf{y}_k \mid \mathbf{x}_k)\, p(\mathbf{x}_k \mid \mathbf{y}_{1:k-1}), \tag{4.12}$$

where the normalization constant Z_k *is given as*

$$Z_k = \int p(\mathbf{y}_k \mid \mathbf{x}_k)\, p(\mathbf{x}_k \mid \mathbf{y}_{1:k-1})\, \mathrm{d}\mathbf{x}_k. \tag{4.13}$$

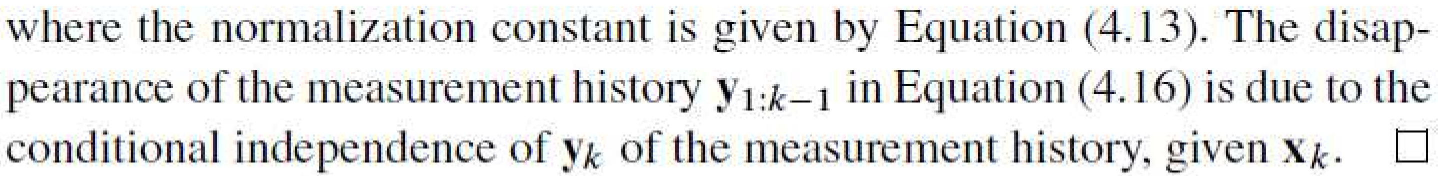
Proof The joint distribution of $\mathbf{x}_k$ and $\mathbf{x}_{k-1}$ given $\mathbf{y}_{1:k-1}$ can be computed as

$$\begin{aligned} p(\mathbf{x}_k, \mathbf{x}_{k-1} \mid \mathbf{y}_{1:k-1}) &= p(\mathbf{x}_k \mid \mathbf{x}_{k-1}, \mathbf{y}_{1:k-1})\, p(\mathbf{x}_{k-1} \mid \mathbf{y}_{1:k-1}) \\ &= p(\mathbf{x}_k \mid \mathbf{x}_{k-1})\, p(\mathbf{x}_{k-1} \mid \mathbf{y}_{1:k-1}), \end{aligned} \tag{4.14}$$

where the disappearance of the measurement history $\mathbf{y}_{1:k-1}$ is due to the Markov property of the sequence $\{\mathbf{x}_k : k = 1, 2, \ldots\}$. The marginal distribution of $\mathbf{x}_k$ given $\mathbf{y}_{1:k-1}$ can be obtained by integrating the distribution (4.14) over $\mathbf{x}_{k-1}$, which gives the *Chapman–Kolmogorov equation*

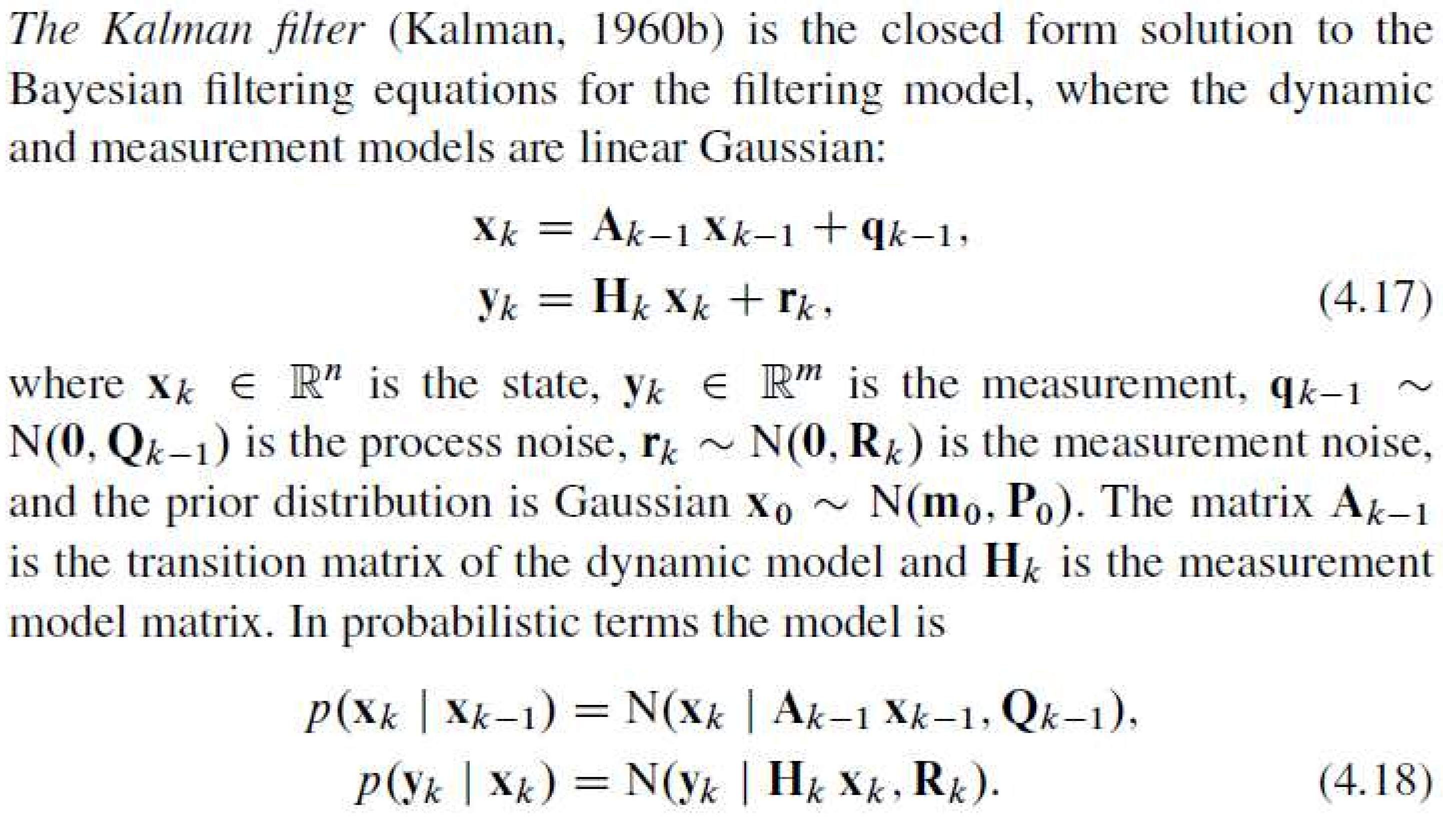
$$p(\mathbf{x}_k \mid \mathbf{y}_{1:k-1}) = \int p(\mathbf{x}_k \mid \mathbf{x}_{k-1})\, p(\mathbf{x}_{k-1} \mid \mathbf{y}_{1:k-1})\, \mathrm{d}\mathbf{x}_{k-1}. \tag{4.15}$$

If $\mathbf{x}_{k-1}$ is discrete, then the above integral is replaced with summation over $\mathbf{x}_{k-1}$. The distribution of $\mathbf{x}_k$ given $\mathbf{y}_k$ and $\mathbf{y}_{1:k-1}$, that is, given $\mathbf{y}_{1:k}$, can be # computed by *Bayes' rule*



where the normalization constant is given by Equation (4.13). The disappearance of the measurement history $\mathbf{y}_{1:k-1}$ in Equation (4.16) is due to the conditional independence of $\mathbf{y}_k$ of the measurement history, given $\mathbf{x}_k$. □

4.3 Kalman Filter



The Kalman filter (Kalman, 1960b) is the closed form solution to the Bayesian filtering equations for the filtering model, where the dynamic and measurement models are linear Gaussian:

$$\begin{aligned} \mathbf{x}_k &= \mathbf{A}_{k-1}\,\mathbf{x}_{k-1} + \mathbf{q}_{k-1}, \\ \mathbf{y}_k &= \mathbf{H}_k\,\mathbf{x}_k + \mathbf{r}_k, \end{aligned} \tag{4.17}$$

where $\mathbf{x}_k \in \mathbb{R}^n$ is the state, $\mathbf{y}_k \in \mathbb{R}^m$ is the measurement, $\mathbf{q}_{k-1} \sim \mathrm{N}(\mathbf{0}, \mathbf{Q}_{k-1})$ is the process noise, $\mathbf{r}_k \sim \mathrm{N}(\mathbf{0}, \mathbf{R}_k)$ is the measurement noise, and the prior distribution is Gaussian $\mathbf{x}_0 \sim \mathrm{N}(\mathbf{m}_0, \mathbf{P}_0)$. The matrix $\mathbf{A}_{k-1}$ is the transition matrix of the dynamic model and $\mathbf{H}_k$ is the measurement model matrix. In probabilistic terms the model is

$$\begin{aligned} p(\mathbf{x}_k \mid \mathbf{x}_{k-1}) &= \mathrm{N}(\mathbf{x}_k \mid \mathbf{A}_{k-1}\,\mathbf{x}_{k-1}, \mathbf{Q}_{k-1}), \\ p(\mathbf{y}_k \mid \mathbf{x}_k) &= \mathrm{N}(\mathbf{y}_k \mid \mathbf{H}_k\,\mathbf{x}_k, \mathbf{R}_k). \end{aligned} \tag{4.18}$$

#

Theorem 4.2 (Kalman filter) *The Bayesian filtering equations for the linear filtering model* (4.17) *can be evaluated in closed form and the resulting distributions are Gaussian:*

$$
\begin{aligned}
p(\mathbf{x}_k \mid \mathbf{y}_{1:k-1}) &= \mathrm{N}(\mathbf{x}_k \mid \mathbf{m}_k^-, \mathbf{P}_k^-), \\
p(\mathbf{x}_k \mid \mathbf{y}_{1:k}) &= \mathrm{N}(\mathbf{x}_k \mid \mathbf{m}_k, \mathbf{P}_k), \\
p(\mathbf{y}_k \mid \mathbf{y}_{1:k-1}) &= \mathrm{N}(\mathbf{y}_k \mid \mathbf{H}_k\, \mathbf{m}_k^-, \mathbf{S}_k).
\end{aligned}
\tag{4.19}
$$

The parameters of the distributions above can be computed with the following Kalman filter prediction *and* update steps.

- The prediction step *is*

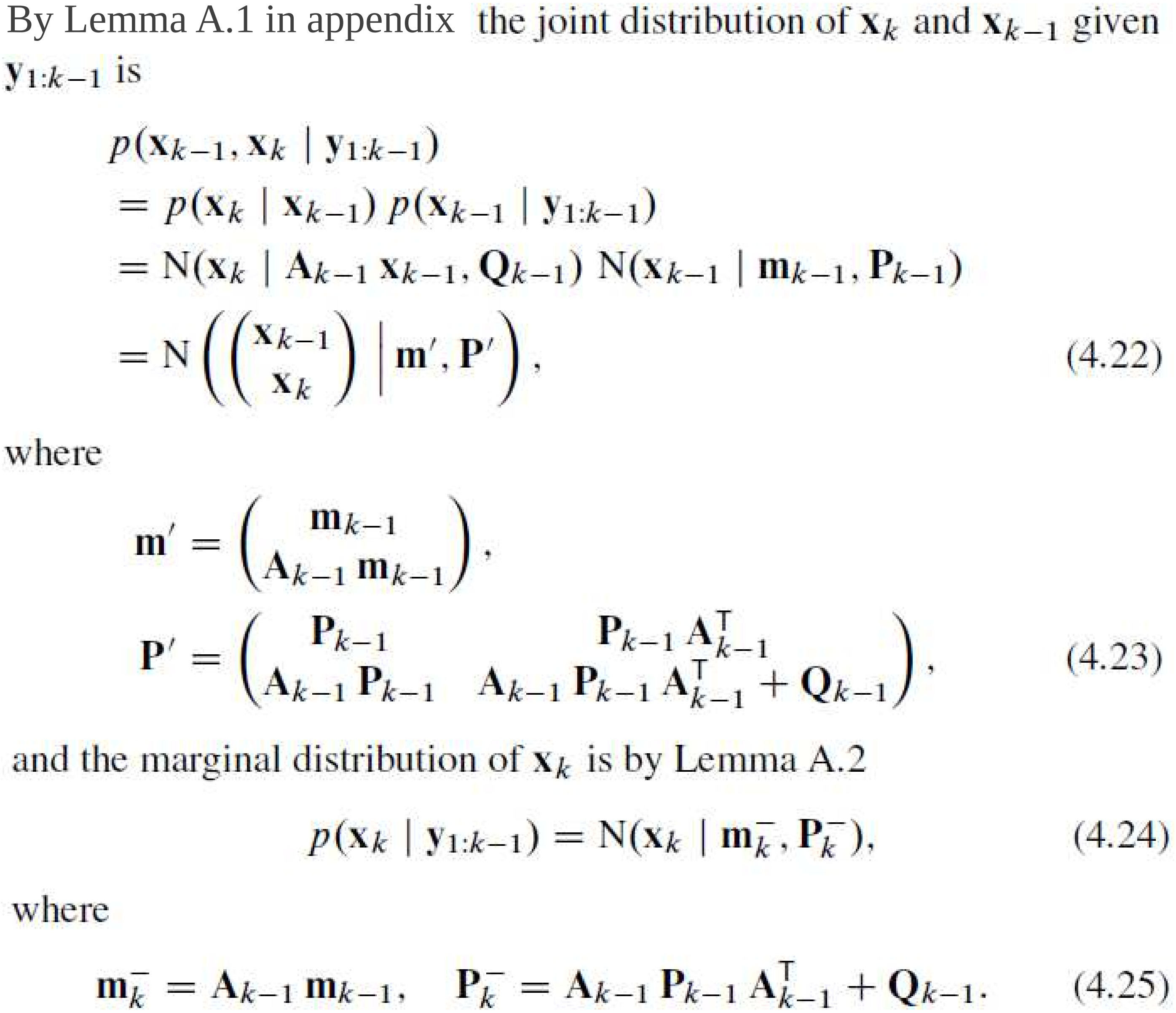
$$
\begin{aligned}
\mathbf{m}_k^- &= \mathbf{A}_{k-1}\, \mathbf{m}_{k-1}, \\
\mathbf{P}_k^- &= \mathbf{A}_{k-1}\, \mathbf{P}_{k-1}\, \mathbf{A}_{k-1}^\top + \mathbf{Q}_{k-1}.
\end{aligned}
\tag{4.20}
$$

- The update step *is*

$$
\begin{aligned}
\mathbf{v}_k &= \mathbf{y}_k - \mathbf{H}_k\, \mathbf{m}_k^-, \\
\mathbf{S}_k &= \mathbf{H}_k\, \mathbf{P}_k^-\, \mathbf{H}_k^\top + \mathbf{R}_k, \\
\mathbf{K}_k &= \mathbf{P}_k^-\, \mathbf{H}_k^\top\, \mathbf{S}_k^{-1}, \\
\mathbf{m}_k &= \mathbf{m}_k^- + \mathbf{K}_k\, \mathbf{v}_k, \\
\mathbf{P}_k &= \mathbf{P}_k^- - \mathbf{K}_k\, \mathbf{S}_k\, \mathbf{K}_k^\top.
\end{aligned}
\tag{4.21}
$$

#

Proof



1 By Lemma A.1 in appendix the joint distribution of $\mathbf{x}_k$ and $\mathbf{x}_{k-1}$ given $\mathbf{y}_{1:k-1}$ is

$$\begin{aligned} & p(\mathbf{x}_{k-1}, \mathbf{x}_k \mid \mathbf{y}_{1:k-1}) \\ & = p(\mathbf{x}_k \mid \mathbf{x}_{k-1})\, p(\mathbf{x}_{k-1} \mid \mathbf{y}_{1:k-1}) \\ & = \mathrm{N}(\mathbf{x}_k \mid \mathbf{A}_{k-1}\, \mathbf{x}_{k-1}, \mathbf{Q}_{k-1})\, \mathrm{N}(\mathbf{x}_{k-1} \mid \mathbf{m}_{k-1}, \mathbf{P}_{k-1}) \\ & = \mathrm{N}\left(\begin{pmatrix} \mathbf{x}_{k-1} \\ \mathbf{x}_k \end{pmatrix} \,\middle|\, \mathbf{m}', \mathbf{P}' \right), \end{aligned} \qquad (4.22)$$

where

$$\begin{aligned} \mathbf{m}' &= \begin{pmatrix} \mathbf{m}_{k-1} \\ \mathbf{A}_{k-1}\, \mathbf{m}_{k-1} \end{pmatrix}, \\ \mathbf{P}' &= \begin{pmatrix} \mathbf{P}_{k-1} & \mathbf{P}_{k-1}\, \mathbf{A}_{k-1}^{\mathsf{T}} \\ \mathbf{A}_{k-1}\, \mathbf{P}_{k-1} & \mathbf{A}_{k-1}\, \mathbf{P}_{k-1}\, \mathbf{A}_{k-1}^{\mathsf{T}} + \mathbf{Q}_{k-1} \end{pmatrix}, \end{aligned} \qquad (4.23)$$

and the marginal distribution of $\mathbf{x}_k$ is by Lemma A.2

$$p(\mathbf{x}_k \mid \mathbf{y}_{1:k-1}) = \mathrm{N}(\mathbf{x}_k \mid \mathbf{m}_k^-, \mathbf{P}_k^-), \qquad (4.24)$$

where

$$\mathbf{m}_k^- = \mathbf{A}_{k-1}\, \mathbf{m}_{k-1}, \quad \mathbf{P}_k^- = \mathbf{A}_{k-1}\, \mathbf{P}_{k-1}\, \mathbf{A}_{k-1}^{\mathsf{T}} + \mathbf{Q}_{k-1}. \qquad (4.25)$$

#

2 By Lemma A.1, the joint distribution of $\mathbf{y}_k$ and $\mathbf{x}_k$ is

$$\begin{aligned} p(\mathbf{x}_k, \mathbf{y}_k \mid \mathbf{y}_{1:k-1}) &= p(\mathbf{y}_k \mid \mathbf{x}_k)\, p(\mathbf{x}_k \mid \mathbf{y}_{1:k-1}) \\ &= \mathrm{N}(\mathbf{y}_k \mid \mathbf{H}_k\, \mathbf{x}_k, \mathbf{R}_k)\, \mathrm{N}(\mathbf{x}_k \mid \mathbf{m}_k^-, \mathbf{P}_k^-) \\ &= \mathrm{N}\left(\begin{pmatrix} \mathbf{x}_k \\ \mathbf{y}_k \end{pmatrix} \,\middle|\, \mathbf{m}'', \mathbf{P}'' \right), \end{aligned} \tag{4.26}$$

where

$$\mathbf{m}'' = \begin{pmatrix} \mathbf{m}_k^- \\ \mathbf{H}_k\, \mathbf{m}_k^- \end{pmatrix}, \qquad \mathbf{P}'' = \begin{pmatrix} \mathbf{P}_k^- & \mathbf{P}_k^-\, \mathbf{H}_k^\top \\ \mathbf{H}_k\, \mathbf{P}_k^- & \mathbf{H}_k\, \mathbf{P}_k^-\, \mathbf{H}_k^\top + \mathbf{R}_k \end{pmatrix}. \tag{4.27}$$

3 By Lemma A.2 the conditional distribution of $\mathbf{x}_k$ is

$$\begin{aligned} p(\mathbf{x}_k \mid \mathbf{y}_k, \mathbf{y}_{1:k-1}) &= p(\mathbf{x}_k \mid \mathbf{y}_{1:k}) \\ &= \mathrm{N}(\mathbf{x}_k \mid \mathbf{m}_k, \mathbf{P}_k), \end{aligned} \tag{4.28}$$

where

$$\begin{aligned} \mathbf{m}_k &= \mathbf{m}_k^- + \mathbf{P}_k^-\, \mathbf{H}_k^\top (\mathbf{H}_k\, \mathbf{P}_k^-\, \mathbf{H}_k^\top + \mathbf{R}_k)^{-1} [\mathbf{y}_k - \mathbf{H}_k\, \mathbf{m}_k^-], \\ \mathbf{P}_k &= \mathbf{P}_k^- - \mathbf{P}_k^-\, \mathbf{H}_k^\top\, (\mathbf{H}_k\, \mathbf{P}_k^-\, \mathbf{H}_k^\top + \mathbf{R}_k)^{-1}\, \mathbf{H}_k\, \mathbf{P}_k^-, \end{aligned} \tag{4.29}$$

which can be also written in the form (4.21). □

#

Example 4.2 (Kalman filter for a Gaussian random walk) *Assume that we are observing measurements y_k of the Gaussian random walk model given in Example 4.1 and we want to estimate the state x_k at each time step. The information obtained up to time step k is summarized by the Gaussian filtering density*

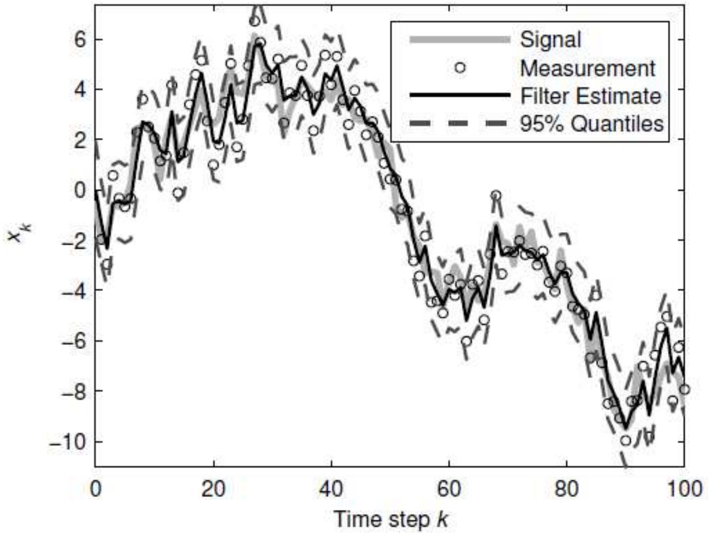
$$p(x_k \mid y_{1:k}) = \mathrm{N}(x_k \mid m_k, P_k). \tag{4.30}$$

The Kalman filter prediction and update equations are now given as

$$\begin{aligned}
m_k^- &= m_{k-1}, \\
P_k^- &= P_{k-1} + Q, \\
m_k &= m_k^- + \frac{P_k^-}{P_k^- + R}(y_k - m_k^-), \\
P_k &= P_k^- - \frac{(P_k^-)^2}{P_k^- + R}.
\end{aligned} \tag{4.31}$$

The result of applying this Kalman filter to the data in Figure 4.1 is shown in Figure 4.4.

#



EXERCISES

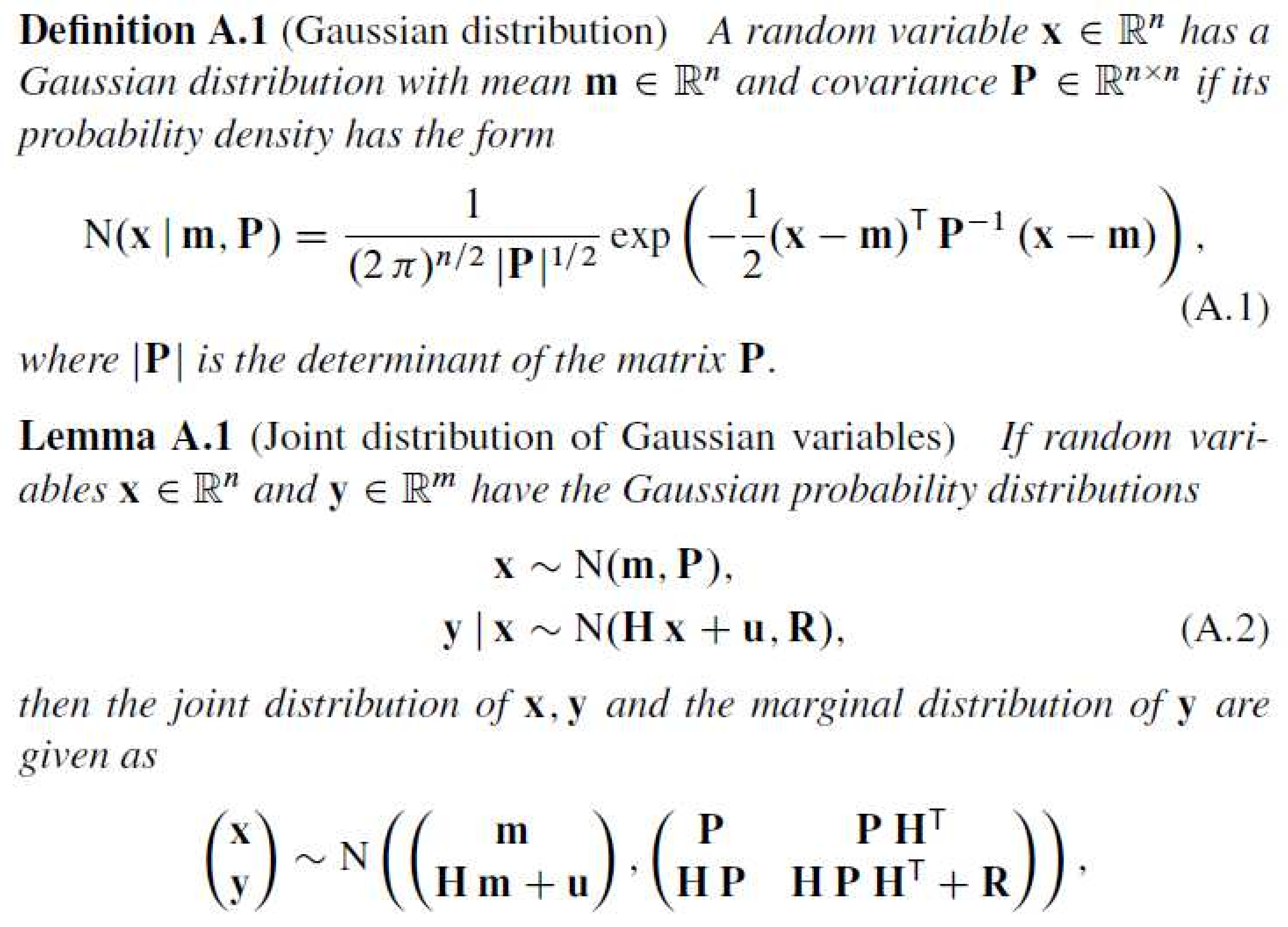
4.1 Derive the Kalman filter equations for the following linear-Gaussian filtering model with non-zero-mean noises:

$$\begin{aligned} \mathbf{x}_k &= \mathbf{A}\,\mathbf{x}_{k-1} + \mathbf{q}_{k-1}, \\ \mathbf{y}_k &= \mathbf{H}\,\mathbf{x}_k + \mathbf{r}_k, \end{aligned} \tag{4.34}$$

where $\mathbf{q}_{k-1} \sim \mathrm{N}(\mathbf{m}_q, \mathbf{Q})$ and $\mathbf{r}_k \sim \mathrm{N}(\mathbf{m}_r, \mathbf{R})$.

4.2 Write down the Bayesian filtering equations for finite-state hidden Markov models (HMM), that is, for models where the state only takes values from a finite set $x_k \in \{1, \ldots, N_x\}$.

APPENDIX



Definition A.1 (Gaussian distribution) *A random variable* $\mathbf{x} \in \mathbb{R}^n$ *has a Gaussian distribution with mean* $\mathbf{m} \in \mathbb{R}^n$ *and covariance* $\mathbf{P} \in \mathbb{R}^{n \times n}$ *if its probability density has the form*

$$\mathrm{N}(\mathbf{x} \mid \mathbf{m}, \mathbf{P}) = \frac{1}{(2\,\pi)^{n/2}\,|\mathbf{P}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathsf{T}}\,\mathbf{P}^{-1}\,(\mathbf{x}-\mathbf{m})\right), \tag{A.1}$$

where $|\mathbf{P}|$ *is the determinant of the matrix* $\mathbf{P}$.

Lemma A.1 (Joint distribution of Gaussian variables) *If random variables* $\mathbf{x} \in \mathbb{R}^n$ *and* $\mathbf{y} \in \mathbb{R}^m$ *have the Gaussian probability distributions*

$$\begin{aligned} \mathbf{x} &\sim \mathrm{N}(\mathbf{m}, \mathbf{P}), \\ \mathbf{y} \mid \mathbf{x} &\sim \mathrm{N}(\mathbf{H}\,\mathbf{x} + \mathbf{u}, \mathbf{R}), \end{aligned} \tag{A.2}$$

then the joint distribution of $\mathbf{x}, \mathbf{y}$ *and the marginal distribution of* $\mathbf{y}$ *are given as*

$$\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mathbf{m} \\ \mathbf{H}\,\mathbf{m} + \mathbf{u} \end{pmatrix}, \begin{pmatrix} \mathbf{P} & \mathbf{P}\,\mathbf{H}^{\mathsf{T}} \\ \mathbf{H}\,\mathbf{P} & \mathbf{H}\,\mathbf{P}\,\mathbf{H}^{\mathsf{T}} + \mathbf{R} \end{pmatrix}\right),$$

#

Lemma A.2 (Conditional distribution of Gaussian variables) *If the random variables* $\mathbf{x}$ *and* $\mathbf{y}$ *have the joint Gaussian probability distribution*

$$\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix}, \begin{pmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^{\mathsf{T}} & \mathbf{B} \end{pmatrix} \right), \tag{A.4}$$

then the marginal and conditional distributions of $\mathbf{x}$ *and* $\mathbf{y}$ *are given as follows:*

$$\begin{aligned} \mathbf{x} &\sim \mathrm{N}(\mathbf{a}, \mathbf{A}), \\ \mathbf{y} &\sim \mathrm{N}(\mathbf{b}, \mathbf{B}), \\ \mathbf{x} \mid \mathbf{y} &\sim \mathrm{N}(\mathbf{a} + \mathbf{C}\,\mathbf{B}^{-1}\,(\mathbf{y} - \mathbf{b}), \mathbf{A} - \mathbf{C}\,\mathbf{B}^{-1}\mathbf{C}^{\mathsf{T}}), \\ \mathbf{y} \mid \mathbf{x} &\sim \mathrm{N}(\mathbf{b} + \mathbf{C}^{\mathsf{T}}\,\mathbf{A}^{-1}\,(\mathbf{x} - \mathbf{a}), \mathbf{B} - \mathbf{C}^{\mathsf{T}}\,\mathbf{A}^{-1}\,\mathbf{C}). \end{aligned} \tag{A.5}$$

#