Ville Väänänen

Gaussian filtering based maximum likelihood and maximum a posteriori estimation in discrete-time state-space models

School of Electrical Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo July 5, 2012

Thesis supervisor:

Prof. Jouko Lampinen

Thesis instructor:

D.Sc. (Tech.) Simo Särkkä

Aalto University
School of Electrical
Engineering

Contents

1 Introduction

(Murphy 2002)

- SSMs vs Box-Jenkins
- Role of static parameters
- Importance of estimating static parameters
- Overview of different approaches

2 Background

2.1 State space models

State space models (SSMs) provide a unified probabilistic methodology for modeling sequential data (Ljung et al. 1994; Durbin et al. 2012; Cappé et al. 2005; Barber et al. 2011). Sequential data arise in numerous applications, typically in the form of time-series measurements . However it is not necessary for the sequence index to have a temporal meaning. In probabilistic terms a time-series can be described by a *stochastic process* $\mathbf{z} = \{\mathbf{z}_k : k \in K\}$, where $\mathbf{z}_k$ is a random variable and $K \subset \mathbb{R}$ for continuous time or $K \subset \mathbb{N}$ for discrete time sequences. In this thesis we will only be concerned with discerete time processes and the sample space of $\mathbf{z}_k$ will be $\mathbb{R}^d$. The shorthand $\mathbf{z}_{1:k}$ will be used to mean the subset $\{\mathbf{z}_1, \ldots, \mathbf{z}_k\}$. We will also denote by $\mathbf{Z}$ the $d \times T$ matrix, that has all T values of the process $\mathbf{z}$ as columns.

examples

A fundamental question in probabilistic models for sequential data is how to model the dependence between variables. It is infeasible to assume that every random variable in the process depends on all the others. Thus it is common to assume a *Markov chain*, where the distribution of the process at the current timestep depends only on the distribution in the previous timestep. A further assumption in SSMs is that the process of interest, the dynamic process $\mathbf{x}$, is not directly observed but only through another stochastic process, the *measurement process* $\mathbf{y}$. Since $\mathbf{x}$ is not observed, SSMs belong to the class of *latent variable models*. Sometimes, as in Cappé et al. 2005, SSMs are called *hidden Markov models* (HMM) but usually this implies that the sample space of $\mathbf{x}$ is discrete. Another assumption is that the values of the measurement process are conditionally independent given the latent Markov process. An intuitive way to present conditional independence properties between random variables is a *Bayes network* presented by a directed acyclic graph (DAG) (Pearl 1988; Bishop 2006). A Bayes network presentation of a discrete-time SSM is given in figure 1.

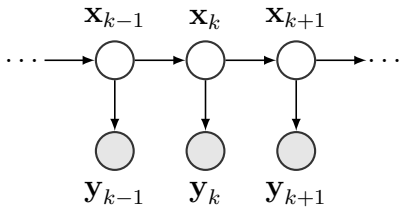

Figure 1: SSM as a graphical model presented with a directed acyclic graph

The value $\mathbf{x}_k$ of the dynamic process at time k is called the *state*at time k. Taking into account the Markov property

Explain state

$$p(\mathbf{x}_k | \mathbf{x}_{1:k-1}) = p(\mathbf{x}_k | \mathbf{x}_{k-1}) \tag{1}$$

of the dynamic process and the conditional independence property

$$p(\mathbf{y}_k | \mathbf{x}_{1:k}, \mathbf{y}_{1:k-1}) = p(\mathbf{y}_k | \mathbf{x}_k) \tag{2}$$

of the measurement process, the joint distribution of states and measurements factorises as

$$p(\mathbf{X}, \mathbf{Y} \,|\, \boldsymbol{\theta}) = p(\mathbf{x}_0 \,|\, \boldsymbol{\theta}) \prod_{k=1}^{T} p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \boldsymbol{\theta})\, p(\mathbf{y}_k \,|\, \mathbf{x}_k, \boldsymbol{\theta})\,. \tag{3}$$

Thus in order to describe a SSM one needs to specify three probability distributions:

Prior distribution $p(\mathbf{x}_0 \,|\, \boldsymbol{\theta})$ is the distribution assumed for the state prior to observing any measurements. The sensitivity of the posterior distributions to the prior depends on the amount of data (the more data the less sensitivity).

Dynamic model $p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \boldsymbol{\theta})$ dictates the time evolution of the states

Measurement model $p(\mathbf{y}_k \,|\, \mathbf{x}_k, \boldsymbol{\theta})$ models how the observations depend on the state and the statistics of the noise

In this thesis it is assumed that the parametric form of these distributions is known for example by physical modeling (Ljung et al. 1994). However the distributions are dependent on the vector parameter $\boldsymbol{\theta}$ whose value is uncertain.

Example: 1D random walk

The simplest example is a one dimensional random-walk observed in Gaussian noise. We will assume $p(\mathbf{x}_0) = \mathrm{N}(0, P_0)$. In an alternative (but equivalent) notation the dynamics model is now

$$p(x_k \,|\, x_{k-1}) = x_{k-1} + q_{k-1}, \tag{4}$$

where $q_{k-1} \sim \mathrm{N}(0, Q)$ and the measurement model is

$$p(y_k \,|\, x_k) = x_k + r_k, \tag{5}$$

where $r_k \sim \mathrm{N}(0, R)$. A simulation from the model is presented in figure 2.

In a SSM it is assumed that at time k the system is in *state* $\mathbf{x}_k \in \mathbb{R}^{d_x}$.The state vectors are random variables which contain the quantities of interest in the system, such as position and velocity in case of a kinetics model. The state is not observed directly, instead at time k we acquire a *noisy* measurement $\mathbf{y}_k \in \mathbb{R}^{d_y}$. Since the states are not observed, they are called hidden or latent variables.. The system state evolves according to the *dynamics* equation

$$\mathbf{x}_k = \mathbf{f}(\mathbf{x}_{k-1}) + \mathbf{q}_{k-1} \tag{6}$$

$$\mathbf{q}_{k-1} \sim \mathrm{N}(0, \mathbf{Q})\,, \tag{7}$$

where $f : \mathbb{R}^{d_x} \to \mathbb{R}^{d_x}$ and in this thesis we restrict ourselves to additive Gaussian noise. It is assumed that the states form a first order *Markov chain*, so that the

This whole chapter needs to be refactored

stationarity

explain hidden variables

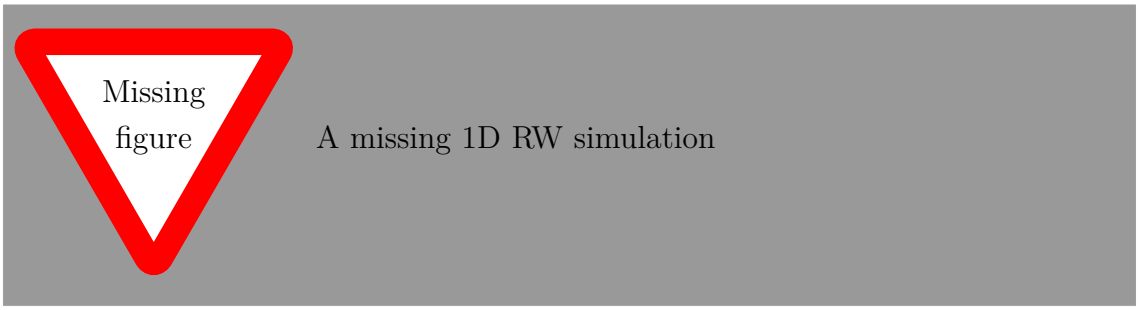

Figure 2: Simulation from the 1D RW model

current state is conditionally independent of the earlier states given the previous state:

$$p(\mathbf{x}_k \,|\, \mathbf{x}_{1:k-1}) = p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}) \tag{8}$$

The observations depend on the state through the measurement equation

$$\mathbf{y}_k = \mathbf{h}(\mathbf{x_k}) + \mathbf{r}_k \tag{9}$$
$$\mathbf{r}_k \sim \mathrm{N}(0, \mathbf{R})\,, \tag{10}$$

where $h : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$. An equivalent way of presenting equations (25a), (25c), (25b) and (25d) is

$$p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}) = \mathrm{N}\big(\mathbf{f}(\mathbf{x}_{k-1}), \mathbf{Q}\big) \tag{11}$$
$$p(\mathbf{y}_k \,|\, \mathbf{x}_k) = \mathrm{N}\big(\mathbf{h}(\mathbf{x}_k), \mathbf{R}\big)\,. \tag{12}$$

2.2 Bayesian optimal filtering and smoothing

Inference can be defined as answering questions of interest with a probability distribution (Barber et al. 2011). In case of SSMs there are many questions of interest, but most commonly one would like to know the *marginal posterior distribution* of the states. State inference can be divided into subcategories based on the temporal relationship between the state and the observations (Särkkä 2006):

Predictive distribution $p\big(\mathbf{x}_k \,\big|\, \mathbf{y}_{1:k-1}\big)$ is the predicted distribution of the state in the next timestep (or more generally at timestep $k + h$, where $h > 0$) given the previous measurements

Filtering distribution $p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k})$ is the marginal posterior distribution of any state $\mathbf{x}_k$ given the measurements up to and including $\mathbf{y}_k$

Smoothing distribution $p(\mathbf{x}_k \,|\, \mathbf{y}_{1:T})$ is the marginal posterior distribution of any state $\mathbf{x}_k$ given the measurements up to and including $\mathbf{y}_T$ where $k < T$

Predictive distribution

Let us now derive a recursive formulation for computing the filtering distribution at time k. Let $p\big(\mathbf{x}_{k-1} \,\big|\, \mathbf{y}_{1:k-1}\big)$ be the filtering distribution of the previous step. Then

$$\begin{aligned} p\big(\mathbf{x}_k \,\big|\, \mathbf{y}_{1:k-1}\big) &= \int p\big(\mathbf{x}_k, \mathbf{x}_{k-1} \,\big|\, \mathbf{y}_{1:k-1}\big)\ \mathrm{d}\mathbf{x}_{k-1} \\ &= \int p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1})\, p\big(\mathbf{x}_{k-1} \,\big|\, \mathbf{y}_{1:k-1}\big)\ \mathrm{d}\mathbf{x}_{k-1}, \end{aligned} \tag{13}$$

which is known as the *Chapman-Kolmogorov equation* (Särkkä 2006).

Filtering distribution

Incorporating the newest measurement can be achieved with the Bayes' rule (see for example Gelman et al. 2004)

check marginal likelihood wording

$$\underbrace{p(\mathbf{x}_k|\mathbf{y}_{1:k})}_{\text{posterior}} = \frac{\overbrace{p(\mathbf{y}_k|\mathbf{x}_k)}^{\text{likelihood}}\overbrace{p\big(\mathbf{x}_k\big|\mathbf{y}_{1:k-1}\big)}^{\text{prior}}}{\underbrace{p\big(\mathbf{y}_k\big|\mathbf{y}_{1:k-1}\big)}_{\text{marginal likelihood}}}$$
$$= \frac{p(\mathbf{y}_k|\mathbf{x}_k)\,p\big(\mathbf{x}_k\big|\mathbf{y}_{1:k-1}\big)}{\int p(\mathbf{y}_k|\mathbf{x}_k)\,p\big(\mathbf{x}_k\big|\mathbf{y}_{1:k-1}\big)\;\mathrm{d}\mathbf{x}_k} \tag{14}$$

which is called the measurement update equation.

Smoothing distribution

The smoothing distributions can also be computed recursively by assuming that the filtering distributions and the smoothing distribution $p(\mathbf{x}_{k+1}|\mathbf{y}_{1:T})$ of the "previous" step are available. Since

$$\begin{aligned} p(\mathbf{x}_k|\mathbf{x}_{k+1},\mathbf{y}_{1:T}) &= p(\mathbf{x}_k|\mathbf{x}_{k+1},\mathbf{y}_{1:k}) \\ &= \frac{p(\mathbf{x}_k,\mathbf{x}_{k+1}|\mathbf{y}_{1:k})}{p(\mathbf{x}_{k+1}|\mathbf{y}_{1:k})} \\ &= \frac{p(\mathbf{x}_{k+1}|\mathbf{x}_k)\,p(\mathbf{x}_k|\mathbf{y}_{1:k})}{p(\mathbf{x}_{k+1}|\mathbf{y}_{1:k})} \end{aligned}$$

we get

$$p(\mathbf{x}_k|\mathbf{y}_{1:T}) = p(\mathbf{x}_k|\mathbf{y}_{1:k}) \int \left[\frac{p(\mathbf{x}_{k+1}|\mathbf{x}_k)\,p(\mathbf{x}_{k+1}|\mathbf{y}_{1:T})}{p(\mathbf{x}_{k+1}|\mathbf{y}_{1:k})}\right] \mathrm{d}\mathbf{x}_{k+1}, \tag{15}$$

where $p(\mathbf{x}_{k+1}|\mathbf{y}_{1:k})$ can be computed by equation (13).

Marginal likelihood

An important quantity concerning parameter estimation is the marginal likelihood $p(\mathbf{y}_{1:T})$. Since

$$p\big(\mathbf{y}_k\big|\mathbf{y}_{1:k-1}\big) = \int p(\mathbf{y}_k|\mathbf{x}_k)\,p\big(\mathbf{x}_k\big|\mathbf{y}_{1:k-1}\big)\;\mathrm{d}\mathbf{x}_k \tag{16}$$

the marginal likelihood can be computed from

$$p(\mathbf{y}_{1:T}) = p(\mathbf{y}_1)\prod_{k=2}^{T} p\big(\mathbf{y}_k\big|\mathbf{y}_{1:k-1}\big) \tag{17}$$

2.3 Parameters in SSM

We will assume that the prior distribution, the dynamics model and the measurement model are known except for a set of parameters $\boldsymbol{\theta}$. Now the joint distribution of all the variables in the SSM can be written as

$$p(\mathbf{x}_{0:T}, \mathbf{y}_{1:T}, \boldsymbol{\theta}) = p(\boldsymbol{\theta})\, p(\mathbf{x}_0|\boldsymbol{\theta}) \prod_{k=1}^{T} p(\mathbf{x}_k|\mathbf{x}_{k-1}, \boldsymbol{\theta})\, p(\mathbf{y}_k|\mathbf{x}_k, \boldsymbol{\theta})\,. \tag{18}$$

3 State estimation

3.1 Linear-Gaussian State Space Models

Linear-Gaussian SSMs can be defined with the following equations

$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} + \mathbf{q}_{k-1} \tag{19a}$$
$$\mathbf{y}_k = \mathbf{H}\mathbf{x}_\mathbf{k} + \mathbf{r}_k \tag{19b}$$
$$\mathbf{q}_{k-1} \sim \mathrm{N}(0, \mathbf{Q}) \tag{19c}$$
$$\mathbf{r}_k \sim \mathrm{N}(0, \mathbf{R}) \tag{19d}$$
$$\mathbf{x}_0 \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{19e}$$

Equations (19a), (19b), (19c) and (19d) together specify the following conditional distributions

$$\mathbf{x}_k|\mathbf{x}_{k-1} \sim \mathrm{N}(\mathbf{A}\mathbf{x}_{k-1}, \mathbf{Q}) \tag{20}$$
$$\mathbf{y}_k|\mathbf{x}_k \sim \mathrm{N}(\mathbf{H}\mathbf{x}_\mathbf{k}, \mathbf{R}) \tag{21}$$

Linearity in this case means that $\mathbf{x}_k$ is a linear combinationof the elements of $\mathbf{x}_{k-1}$ and $\mathbf{y}_k$ is a linear combination of the elements of $\mathbf{x}_k$ (with additive noise in both cases). Since the noise terms $\mathbf{q}_{k-1}$ and $\mathbf{r}_k$ are assumed to be white and Gaussian, these models are called linear-Gaussian.

Better wording

3.1.1 Kalman filter

To derive the expression for the log-likelihood function in our case, let us first see what the Kalman filter calculates. Firstly, the recursions are as follows (Kalman et al. 1960; Jazwinski 2007):

prediction:

$$\mathbf{m}_{k|k-1} = \mathbf{A}\mathbf{m}_{k-1|k-1} \tag{22a}$$
$$\mathbf{P}_{k|k-1} = \mathbf{A}\mathbf{P}_{k-1|k-1}\mathbf{A}^T + \mathbf{Q} \tag{22b}$$

update:

$$\mathbf{v}_k = \mathbf{y}_k - \mathbf{H}\mathbf{m}_{k|k-1} \tag{22c}$$
$$\mathbf{S}_k = \mathbf{H}\mathbf{P}_{k|k-1}\mathbf{H}^T + \mathbf{R} \tag{22d}$$
$$\mathbf{K}_k = \mathbf{P}_{k|k-1}\mathbf{H}^T\mathbf{S}_k^{-1} \tag{22e}$$
$$\mathbf{m}_{k|k} = \mathbf{m}_{k|k-1} + \mathbf{K}_k\mathbf{v}_k \tag{22f}$$
$$\mathbf{P}_{k|k} = \mathbf{P}_{k|k-1} - \mathbf{K}_k\mathbf{S}_k\mathbf{K}_k^T \tag{22g}$$

This includes the sufficient statistics for the T joint distributions

$$p\big(\mathbf{x}_k, \mathbf{y}_k \big| \mathbf{y}_1, \ldots, \mathbf{y}_{k-1}, \boldsymbol{\theta}\big) = N\left(\begin{bmatrix}\mathbf{x}_k \\ \mathbf{y}_k\end{bmatrix} \middle| \begin{bmatrix}\mathbf{m}_{k|k-1} \\ \mathbf{H}\mathbf{m}_{k|k-1}\end{bmatrix}, \begin{bmatrix}\mathbf{P}_{k|k-1} & \mathbf{P}_{k|k-1}\mathbf{H}^T \\ \mathbf{H}\mathbf{P}_{k|k-1}^T & \mathbf{S}_k\end{bmatrix}\right) \tag{23}$$
$$\Rightarrow p\big(\mathbf{y}_k \big| \mathbf{y}_1, \ldots, \mathbf{y}_{k-1}, \boldsymbol{\theta}\big) = N\left(\mathbf{y}_k \,\middle|\, \mathbf{H}\mathbf{m}_{k|k-1}, \mathbf{S}_k\right)$$

3.1.2 RTS Smoother

The standard RTS smoother gives the statistics $\mathbf{m}_{k|N}$ and $\mathbf{P}_{k|N}$ (Jazwinski 2007; Rauch et al. 1965). The cross-timestep variance $\mathbf{C}_{k|N}$ can be computed with an additional recursive formula alongside the usual RTS smoother recursions

$$\mathbf{J}_k = \mathbf{P}_{k|k}\mathbf{A}^T\mathbf{P}_{k|k+1}^{-1} \tag{24a}$$
$$\mathbf{m}_{k|N} = \mathbf{m}_{k|k} + \mathbf{J}_k\left(\mathbf{m}_{k+1|N} - \mathbf{m}_{k+1|k}\right) \tag{24b}$$
$$\mathbf{P}_{k|N} = \mathbf{P}_{k|k} + \mathbf{J}_k\left(\mathbf{P}_{k+1|N} - \mathbf{P}_{k+1|k}\right)\mathbf{J}_k^T \tag{24c}$$
$$\mathbf{C}_{k|N} = \mathbf{P}_{k|k}\mathbf{J}_{k-1}^T + \mathbf{J}_k\left(\mathbf{C}_{k+1|N} - \mathbf{A}\mathbf{P}_{k|k}\right)\mathbf{J}_{k-1}^T \tag{24d}$$
$$\tag{24e}$$

For more specific details see (Gibson et al. 2005). All in all, the E-step of the EM algorithm in linear-Gaussian SSM:s corresponds to computing the matrices in (**??**) with the help of the Kalman filter and the RTS smoother. In (Elliott et al. 1999) a new kind of filter is presented that can compute (**??**) with only forward recursions.

3.2 Nonlinear-Gaussian SSMs

The SSM model is now

$$\mathbf{x}_k = \mathbf{f}(\mathbf{x}_{k-1}) + \mathbf{q}_{k-1} \tag{25a}$$
$$\mathbf{y}_k = \mathbf{h}(\mathbf{x_k}) + \mathbf{r}_k \tag{25b}$$
$$\mathbf{q}_{k-1} \sim \mathrm{N}(0, \mathbf{Q}) \tag{25c}$$
$$\mathbf{r}_k \sim \mathrm{N}(0, \mathbf{R}) \tag{25d}$$
$$\mathbf{x}_0 \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{25e}$$

We assume an implicit dependence of f and h on the parameter $\boldsymbol{\theta}$.

3.2.1 Gaussian filtering and smoothing

One approach to forming Gaussian approximations is to assume a Gaussian probability density function with mean and variance that match the actual ones (Ito 2000; Särkkä 2006). Let

$$\mathbf{a} \sim \mathrm{N}(\mathbf{m}, \boldsymbol{\Sigma}_a) \tag{26}$$

$$\mathbf{b}|\mathbf{a} \sim \mathrm{N}\big(\mathbf{f}(\mathbf{a}), \boldsymbol{\Sigma}_{b|a}\big) \tag{27}$$

then

$$p(\mathbf{a}, \mathbf{b}) = \mathrm{N}(\mathbf{m}, \boldsymbol{\Sigma}_a)\, \mathrm{N}\big(\mathbf{f}(\mathbf{a}), \boldsymbol{\Sigma}_{b|a}\big) \tag{28}$$

is only Gaussian if $\mathbf{f}(\mathbf{a})$ is linear. Let the Gaussian approximation to (28) be

$$p\left(\begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix}\right) \approx \mathrm{N}\left(\begin{bmatrix}\boldsymbol{\mu}_a\\ \boldsymbol{\mu}_b\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab}\\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb}\end{bmatrix}\right) \tag{29}$$

Then since the marginal distributions of a Gaussian distribution are also Gaussian, we have to have

$$\boldsymbol{\mu}_a = \mathbf{m} \tag{30}$$

$$\boldsymbol{\Sigma}_{aa} = \boldsymbol{\Sigma}_a \tag{31}$$

$$\boldsymbol{\mu}_b = \int \mathbf{b} p(\mathbf{b})\ \mathrm{d}\mathbf{b} \tag{32}$$

$$\boldsymbol{\Sigma}_{bb} = \int (\mathbf{b} - \boldsymbol{\mu}_b)(\mathbf{b} - \boldsymbol{\mu}_b)^T p(\mathbf{b})\ \mathrm{d}\mathbf{b} \tag{33}$$

Both (32) and (33) can be written in terms of (26) and (27). To see this, let us rewrite (32) as

$$\begin{aligned}
\boldsymbol{\mu}_b &= \int \mathbf{b} p(\mathbf{b})\ \mathrm{d}\mathbf{b} \\
&= \int \mathbf{b} \int p(\mathbf{b}|\mathbf{a})\, p(\mathbf{a})\ \mathrm{d}\mathbf{a}\ \mathrm{d}\mathbf{b} \\
&= \int \int \mathbf{b} p(\mathbf{b}|\mathbf{a})\ \mathrm{d}\mathbf{b} p(\mathbf{a})\ \mathrm{d}\mathbf{a} \\
&= \int \mathbf{f}(\mathbf{a}) \mathrm{N}(\mathbf{m}, \boldsymbol{\Sigma}_a)\ \mathrm{d}\mathbf{a}
\end{aligned} \tag{34}$$

and (33) as

$$\begin{aligned}\boldsymbol{\Sigma}_{bb} &= \int \mathbf{b}\mathbf{b}^T p(\mathbf{b})\ \mathrm{d}\mathbf{b} - \boldsymbol{\mu}_b\boldsymbol{\mu}_b^T \\ &= \int \mathbf{f}(\mathbf{a})\mathbf{f}(\mathbf{a})^T p(\mathbf{a})\ \mathrm{d}\mathbf{a} - \boldsymbol{\mu}_b\boldsymbol{\mu}_b^T \\ &\quad + \int\int [(\mathbf{b}-\mathbf{f}(\mathbf{a}))(\mathbf{b}-\mathbf{f}(\mathbf{a}))^T] p(\mathbf{b}\,|\,\mathbf{a})\ \mathrm{d}\mathbf{b} p(\mathbf{a})\ \mathrm{d}\mathbf{a} \\ &= \int (\mathbf{f}(\mathbf{a})-\boldsymbol{\mu}_b)(\mathbf{f}(\mathbf{a})-\boldsymbol{\mu}_b)^T \mathrm{N}(\mathbf{m},\boldsymbol{\Sigma}_a)\ \mathrm{d}\mathbf{a} + \boldsymbol{\Sigma}_{b|a}. \end{aligned} \tag{35}$$

Finally, the cross-covariance $\boldsymbol{\Sigma}_{ab} = \boldsymbol{\Sigma}_{ba}^T$ similarly reads

$$\begin{aligned}\boldsymbol{\Sigma}_{ab} &= \int\int (\mathbf{a}-\boldsymbol{\mu}_a)(\mathbf{b}-\boldsymbol{\mu}_b)^T p(\mathbf{a},\mathbf{b})\ \mathrm{d}\mathbf{a}\ \mathrm{d}\mathbf{b} \\ &= \int\int (\mathbf{a}-\boldsymbol{\mu}_a)(\mathbf{b}-\boldsymbol{\mu}_b)^T p(\mathbf{a})\, p(\mathbf{b}\,|\,\mathbf{a})\ \mathrm{d}\mathbf{a}\ \mathrm{d}\mathbf{b} \\ &= \int (\mathbf{a}-\boldsymbol{\mu}_a)\Big(\int \mathbf{b} p(\mathbf{b}\,|\,\mathbf{a})\ \mathrm{d}\mathbf{b} - \boldsymbol{\mu}_b\Big)^T p(\mathbf{a})\ \mathrm{d}\mathbf{a} \\ &= \int (\mathbf{a}-\mathbf{m})(\mathbf{f}(\mathbf{a})-\boldsymbol{\mu}_b)^T \mathrm{N}(\mathbf{m},\boldsymbol{\Sigma}_a)\ \mathrm{d}\mathbf{a} \end{aligned} \tag{36}$$

To see how this idea can be used to form a Gaussian approximation to (**??**), let us rewrite (**??**) as

$$\begin{aligned} p(\mathbf{x}_{k-1},\mathbf{x}_k\,|\,\mathbf{Y}) &= p(\mathbf{x}_{k-1}\,|\,\mathbf{x}_k,\mathbf{Y}_{1:k-1})\, p(\mathbf{x}_k\,|\,\mathbf{Y}) \\ &= \frac{p(\mathbf{x}_{k-1},\mathbf{x}_k\,|\,\mathbf{Y}_{1:k-1})\, p(\mathbf{x}_k\,|\,\mathbf{Y})}{p(\mathbf{x}_k\,|\,\mathbf{Y}_{1:k-1})}, \end{aligned} \tag{37}$$

where the dependance on the current estimate of the parameter $\hat{\boldsymbol{\theta}}_j$ is suppressed for clarity. Since the Gaussian approximation to (**??**) will be calculated by forward (filtering) and backward (smoothing) recursions, let us assume that we already have available the Gaussian approximation

$$p\Big(\mathbf{x}_{k-1}\Big|\,\mathbf{Y}_{1:k-1},\hat{\boldsymbol{\theta}}_j\Big) \approx \mathrm{N}\big(\mathbf{m}_{k-1|k-1},\mathbf{P}_{k-1|k-1}\big)\,. \tag{38}$$

The Gaussian approximation to

$$p(\mathbf{x}_{k-1},\mathbf{x}_k\,|\,\mathbf{Y}_{1:k-1}) = \mathrm{N}\big(\mathbf{x}_k\,\big|\,\mathbf{f}(\mathbf{x}_{k-1}),\mathbf{Q}\big)\, p(\mathbf{x}_{k-1}\,|\,\mathbf{Y}_{1:k-1}) \tag{39}$$

is then given by application of equations (34), (35) and (36)

$$\mathbf{m}_{k|k-1} = \int \mathbf{f}(\mathbf{x}_{k-1})\mathrm{N}\big(\mathbf{x}_{k-1}\,\big|\,\mathbf{m}_{k-1|k-1}, \mathbf{P}_{k-1|k-1}\big)\ \mathrm{d}\mathbf{x}_{k-1} \tag{40}$$

$$\begin{aligned}\mathbf{P}_{k|k-1} = \int &(\mathbf{f}(\mathbf{x}_{k-1}) - \mathbf{m}_{k|k-1})(\mathbf{f}(\mathbf{x}_{k-1}) - \mathbf{m}_{k|k-1})^T \\ &\mathrm{N}\big(\mathbf{x}_{k-1}\,\big|\,\mathbf{m}_{k-1|k-1}, \mathbf{P}_{k-1|k-1}\big)\ \mathrm{d}\mathbf{x}_{k-1} + \mathbf{Q}\end{aligned} \tag{41}$$

$$\begin{aligned}\mathbf{C}_{k^-} = \int &(\mathbf{x}_{k-1} - \mathbf{m}_{k-1|k-1})(\mathbf{f}(\mathbf{x}_{k-1}) - \mathbf{m}_{k|k-1})^T \\ &\mathrm{N}\big(\mathbf{x}_{k-1}\,\big|\,\mathbf{m}_{k-1|N}, \mathbf{P}_{k-1|N}\big)\ \mathrm{d}\mathbf{x}_{k-1}\end{aligned} \tag{42}$$

so that the approximation is

$$p\Big(\mathbf{x}_{k-1}, \mathbf{x}_k\,\Big|\,\mathbf{Y}_{1:k-1}, \hat{\boldsymbol{\theta}}_j\Big) \approx \mathrm{N}\left(\begin{bmatrix}\mathbf{m}_{k-1|k-1}\\ \mathbf{m}_{k|k-1}\end{bmatrix}, \begin{bmatrix}\mathbf{P}_{k-1|k-1} & \mathbf{C}_{k^-}\\ \mathbf{C}_{k^-}^T & \mathbf{P}_{k|k-1}\end{bmatrix}\right) \tag{43}$$

In order to calculate (40) and (41) we also need a Gaussian approximation for the joint distribution of the current state and measurement given the previous measurements

$$\begin{aligned}p(\mathbf{x}_k, \mathbf{y}_k\,|\,\mathbf{Y}_{1:k-1}) &= \mathrm{N}\big(\mathbf{y}_k\,\big|\,\mathbf{h}(\mathbf{x}_k), \mathbf{R}\big)\, p(\mathbf{x}_k\,|\,\mathbf{Y}_{1:k-1}) \\ &\approx \mathrm{N}\left(\begin{bmatrix}\mathbf{m}_{k|k-1}\\ \boldsymbol{\mu}_k\end{bmatrix}, \begin{bmatrix}\mathbf{P}_{k|k-1} & \mathbf{C}_k\\ \mathbf{C}_k^T & \mathbf{S}_k\end{bmatrix}\right).\end{aligned} \tag{44}$$

Applying equations (34), (35) and (36) again, we get

$$\boldsymbol{\mu}_k = \int \mathbf{h}(\mathbf{x}_k)\mathrm{N}\big(\mathbf{x}_k\,\big|\,\mathbf{m}_{k|k-1}, \mathbf{P}_{k|k-1}\big)\ \mathrm{d}\mathbf{x}_k \tag{45}$$

$$\mathbf{S}_k = \int (\mathbf{h}(\mathbf{x}_k) - \boldsymbol{\mu}_k)(\mathbf{h}(\mathbf{x}_k) - \boldsymbol{\mu}_k)^T\mathrm{N}\big(\mathbf{x}_k\,\big|\,\mathbf{m}_{k|k-1}, \mathbf{P}_{k|k-1}\big)\ \mathrm{d}\mathbf{x}_k + \mathbf{R} \tag{46}$$

$$\mathbf{C}_k = \int (\mathbf{x}_k - \mathbf{m}_{k|k-1})(\mathbf{h}(\mathbf{x}_k) - \boldsymbol{\mu}_k)^T\mathrm{N}\big(\mathbf{x}_k\,\big|\,\mathbf{m}_{k|k-1}, \mathbf{P}_{k|k-1}\big)\ \mathrm{d}\mathbf{x}_k \tag{47}$$

and by using the well known formula for calculating the conditional distribution of jointly Gaussian variables we have

$$\mathbf{m}_{k|k} = \mathbf{m}_{k|k-1} + \mathbf{C}_k\mathbf{S}_k^{-1}\left(\mathbf{y}_k - \boldsymbol{\mu}_k\right) \tag{48}$$

$$\mathbf{P}_{k|k} = \mathbf{P}_{k|k-1} - \mathbf{C}_k\mathbf{S}_k^{-1}\mathbf{C}_k^T. \tag{49}$$

Again using the formula for the conditional of jointly Gaussian variables we get from (43)

$$p(\mathbf{x}_{k-1}\,|\,\mathbf{x}_k, \mathbf{Y}_{1:k-1}) \approx \mathrm{N}(\mathbf{m}_2, \mathbf{P}_2) \tag{50}$$

$$\mathbf{G}_{k-1} = \mathbf{C}_{k^-}\mathbf{P}_{k|k-1}^{-1} \tag{51}$$

$$\mathbf{m}_2 = \mathbf{m}_{k-1|k-1} + \mathbf{G}_{k-1}(\mathbf{x}_k - \mathbf{m}_{k|k-1}) \tag{52}$$

$$\mathbf{P}_2 = \mathbf{P}_{k-1|k-1} - \mathbf{G}_{k-1}\mathbf{P}_{k|k-1}\mathbf{G}_{k-1}^T \tag{53}$$

and then finally we can write the Gaussian approximation to the joint distribution of consecutive states given all the measurements as

$$\begin{aligned} p(\mathbf{x}_{k-1}, \mathbf{x}_k \,|\, \mathbf{Y}) &= p(\mathbf{x}_{k-1} \,|\, \mathbf{x}_k, \mathbf{Y}_{1:k-1})\, p(\mathbf{x}_k \,|\, \mathbf{Y}) \\ &\approx \mathrm{N}\left(\begin{bmatrix} \mathbf{m}_{k-1|N} \\ \mathbf{m}_{k|N} \end{bmatrix}, \begin{bmatrix} \mathbf{P}_{k-1|N} & \mathbf{D}_k \\ \mathbf{D}_k^T & \mathbf{P}_{k|N} \end{bmatrix} \right) \end{aligned} \tag{54}$$

where

$$\mathbf{D}_k = \mathbf{G}_{k-1} \mathbf{P}_{k|N} \tag{55}$$

$$\mathbf{m}_{k-1|N} = \mathbf{m}_{k-1|k-1} + \mathbf{G}_{k-1} \left(\mathbf{m}_{k|N} - \mathbf{m}_{k|k-1} \right) \tag{56}$$

$$\mathbf{P}_{k-1|N} = \mathbf{P}_{k-1|k-1} + \mathbf{G}_{k-1} \left(\mathbf{P}_{k|N} - \mathbf{P}_{k|k-1} \right) \mathbf{G}_{k-1}^T \tag{57}$$

3.2.2 Numerical integration

(Arasaratnam et al. 2009)

3.2.3 Gauss-Hermite Kalman Filter and Smoother

(Ito 2000)

3.2.4 Unscented Kalman Filter and Smoother

(Julier et al. 1997; Merwe 2004)

3.2.5 Cubature Kalman Filter and Smoother

(Arasaratnam et al. 2009; Arasaratnam et al. 2011; Jia et al. 2012)

4 Parameter estimation

4.1 Maximum likelihood and maximum a posteriori estimation

In the Bayesian sense the complete answer to the parameter estimation problem is the marginal posterior probability of the parameters given the measurements

$$\begin{aligned} p(\boldsymbol{\theta} \,|\, \mathbf{Y}) &= \frac{p(\mathbf{Y} \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathbf{Y})} \\ \Rightarrow \log p(\boldsymbol{\theta} \,|\, \mathbf{Y}) &\propto \log p(\mathbf{Y} \,|\, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \end{aligned} \tag{58}$$

Here the matrices of all the states and all the observations are denoted with

notation changes

$$\mathbf{X} = \mathbf{X}_{1:N} = \begin{bmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_N \end{bmatrix} \tag{59}$$

$$\mathbf{Y} = \mathbf{Y}_{1:N} = \begin{bmatrix} \mathbf{y}_1 & \dots & \mathbf{y}_N \end{bmatrix} \tag{60}$$

respectively. Computing the posterior distribution of the parameters is usually intractable . A much easier problem is finding a suitable *point estimate* $\hat{\boldsymbol{\theta}}$. This effectively means that we don't need to worry about the normalizing term $p(\mathbf{Y})$. A point estimate that maximizes the posterior distribution is called a *maximum a posteriori* (MAP) estimate. In the case of a flat prior distribution $p(\boldsymbol{\theta})$ the MAP estimate converges to the *maximum likelihood* (ML) estimate, that maximizes the likelihood $p(\mathbf{Y}|\boldsymbol{\theta})$ (or equivalently its logarithm) . Going further we we will only be concerned with finding the ML estimate, but it should be remembered that both of the methods we consider can be extended to the estimation of the MAP estimate in a straightforward fashion.

why? examples?

elaborate

Since our model contains the latent (unobserved) states, evaluating the likelihood $p(\mathbf{Y}|\boldsymbol{\theta})$ is a problem in itself. Mathematically speaking, we need to integrate out the states (marginalization) from the complete-data likelihood. Because of the Markov conditional independence properties that are implicit in the model, the complete-data likelihood factorizes as

$$p(\mathbf{X}, \mathbf{Y}|\boldsymbol{\theta}) = p(\mathbf{x}_0) \prod_{k=1}^{N} p(\mathbf{y}_k|\mathbf{x}_k)\, p(\mathbf{x}_k|\mathbf{x}_{k-1}) \tag{61}$$

so that the marginal likelihood is obtained by integration:

$$p(\mathbf{Y}|\boldsymbol{\theta}) = \int_{\mathbb{R}^{d_x}} p(\mathbf{X}, \mathbf{Y}|\boldsymbol{\theta})\ \mathrm{d}\mathbf{X} \tag{62}$$

Since $\mathbf{Y}$ is observed, equation (62) is a function of the parameters only. In this linear-Gaussian case, the Kalman filter forward recursions give us the means to perform the integration over the states analytically, so that (62) can be evaluated for any given $\boldsymbol{\theta}$.

4.2 Gradient based numerical optimization

This is the classical way of solving the parameter estimation problem. It consists of computing the gradient of the log-likelihood function and then using some non-linear optimization method to find a *local* maximum to it.(Mbalawata et al. 2011). An efficient non-linear optimization algorithm is the scaled conjugate gradient method (Mbalawata et al. 2011).

To derive the expression for the log-likelihood function in our case, let us first see what the Kalman filter calculates. Firstly, the recursions are as follows (Mbalawata et al. 2011):

prediction:

$$\mathbf{m}_{k|k-1} = \mathbf{A}\mathbf{m}_{k-1|k-1} \tag{63a}$$

$$\mathbf{P}_{k|k-1} = \mathbf{A}\mathbf{P}_{k-1|k-1}\mathbf{A}^T + \mathbf{Q} \tag{63b}$$

update:

$$\mathbf{v}_k = \mathbf{y}_k - \mathbf{H}\mathbf{m}_{k|k-1} \tag{63c}$$
$$\mathbf{S}_k = \mathbf{H}\mathbf{P}_{k|k-1}\mathbf{H}^T + \mathbf{R} \tag{63d}$$
$$\mathbf{K}_k = \mathbf{P}_{k|k-1}\mathbf{H}^T\mathbf{S}_k^{-1} \tag{63e}$$
$$\mathbf{m}_{k|k} = \mathbf{m}_{k|k-1} + \mathbf{K}_k\mathbf{v}_k \tag{63f}$$
$$\mathbf{P}_{k|k} = \mathbf{P}_{k|k-1} - \mathbf{K}_k\mathbf{S}_k\mathbf{K}_k^T \tag{63g}$$

This includes the sufficient statistics for the T joint distributions

$$p\big(\mathbf{x}_k, \mathbf{y}_k \big| \mathbf{y}_1, \ldots, \mathbf{y}_{k-1}, \boldsymbol{\theta}\big) = N\left(\begin{bmatrix}\mathbf{x}_k \\ \mathbf{y}_k\end{bmatrix} \middle| \begin{bmatrix}\mathbf{m}_{k|k-1} \\ \mathbf{H}\mathbf{m}_{k|k-1}\end{bmatrix}, \begin{bmatrix}\mathbf{P}_{k|k-1} & \mathbf{P}_{k|k-1}\mathbf{H}^T \\ \mathbf{H}\mathbf{P}_{k|k-1}^T & \mathbf{S}_k\end{bmatrix}\right) \tag{64}$$
$$\Rightarrow p\big(\mathbf{y}_k \big| \mathbf{y}_1, \ldots, \mathbf{y}_{k-1}, \boldsymbol{\theta}\big) = N\left(\mathbf{y}_k \;\middle|\; \mathbf{H}\mathbf{m}_{k|k-1}, \mathbf{S}_k\right)$$

To see how this enables us to calculate the likelihood, one only needs to note that (it has been assumed that the observations are independent given the states)

$$p(\mathbf{Y}\,|\,\boldsymbol{\theta}) = p(\mathbf{y}_1\,|\,\boldsymbol{\theta}) \prod_{k=2}^{N} p\big(\mathbf{y}_k \big| \mathbf{y}_1, \ldots, \mathbf{y}_{k-1}, \boldsymbol{\theta}\big) \tag{65}$$

Armed with this knowledge, we can write the following expression for the log-likelihood function in this linear-Gaussian case:

$$-2L(\boldsymbol{\theta}) = \sum_{k=1}^{N} \log|\mathbf{S}_k| + \sum_{k=1}^{N} \left(\mathbf{y}_k - \mathbf{H}\mathbf{m}_{k|k-1}\right)^T \mathbf{S}_k^{-1} \left(\mathbf{y}_k - \mathbf{H}\mathbf{m}_{k|k-1}\right) + C, \tag{66}$$

where C is a constant that doesn't depend on $\boldsymbol{\theta}$. Employing an efficient numerical optimization method generally requires that the gradient of the objective function, that is $L(\boldsymbol{\theta})$ in this case, is available. In order to calculate the gradient of $L(\boldsymbol{\theta})$, we need to formally derivate it w.r.t every parameter θ_i in $\boldsymbol{\theta}$:

$$\begin{aligned}\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_i} = &-\frac{1}{2}\sum_{k=1}^{N} \mathrm{Tr}\left(\mathbf{S}_k^{-1}\frac{\partial \mathbf{S}_k}{\partial \theta_i}\right) \\ &+ \sum_{k=1}^{N}\left(\mathbf{H}_k\frac{\partial \mathbf{m}_{k|k-1}}{\partial \theta_i}\right)^T \mathbf{S}_k^{-1}\left(\mathbf{y}_k - \mathbf{H}\mathbf{m}_{k|k-1}\right) \\ &+ \frac{1}{2}\sum_{k=1}^{N}\left(\mathbf{y}_k - \mathbf{H}\mathbf{m}_{k|k-1}\right)^T \mathbf{S}_k^{-1}\left(\frac{\partial \mathbf{S}_k}{\partial \theta_i}\right)\mathbf{S}_k^{-1}\left(\mathbf{y}_k - \mathbf{H}\mathbf{m}_{k|k-1}\right)\end{aligned} \tag{67}$$

From the Kalman filter recursions we find out that

$$\frac{\partial \mathbf{S}_k}{\partial \theta_i} = \mathbf{H}\frac{\partial \mathbf{P}_{k|k-1}}{\partial \theta_i}\mathbf{H} + \frac{\partial \mathbf{R}}{\partial \theta_i} \tag{68}$$

so that we're left with the task of determining the partial derivatives for $\mathbf{m}_{k|k-1}$ and $\mathbf{P}_{k|k-1}$:

$$\frac{\partial \mathbf{m}_{k|k-1}}{\partial \theta_i} = \frac{\partial \mathbf{A}}{\partial \theta_i}\,\mathbf{m}_{k-1|k-1} + \mathbf{A}\,\frac{\partial \mathbf{m}_{k-1|k-1}}{\partial \theta_i} \tag{69}$$

$$\begin{aligned}\frac{\partial \mathbf{P}_{k|k-1}}{\partial \theta_i} = {} & \frac{\partial \mathbf{A}}{\partial \theta_i}\,\mathbf{P}_{k-1|k-1}\mathbf{A}^T + \mathbf{A}\,\frac{\partial \mathbf{P}_{k-1|k-1}}{\partial \theta_i}\,\mathbf{A}^T \\ & + \mathbf{A}\mathbf{P}_{k-1|k-1}\left(\frac{\partial \mathbf{A}}{\partial \theta_i}\right)^T + \frac{\partial \mathbf{Q}}{\partial \theta_i}\end{aligned} \tag{70}$$

as well as for $\mathbf{m}_{k|k}$ and $\mathbf{P}_{k|k}$:

$$\frac{\partial \mathbf{K}_k}{\partial \theta_i} = \frac{\partial \mathbf{P}_{k|k-1}}{\partial \theta_i}\,\mathbf{H}^T\mathbf{S}_k^{-1} - \mathbf{P}_{k|k-1}\mathbf{H}^T\mathbf{S}_k^{-1}\,\frac{\partial \mathbf{S}_k}{\partial \theta_i}\,\mathbf{S}_k^{-1} \tag{71}$$

$$\frac{\partial \mathbf{m}_{k|k}}{\partial \theta_i} = \frac{\partial \mathbf{m}_{k|k-1}}{\partial \theta_i} + \frac{\partial \mathbf{K}_k}{\partial \theta_i}\left(\mathbf{y}_k - \mathbf{H}\mathbf{m}_{k|k-1}\right) - \mathbf{K}_k\mathbf{H}\,\frac{\partial \mathbf{m}_{k|k-1}}{\partial \theta_i} \tag{72}$$

$$\frac{\partial \mathbf{P}_{k|k}}{\partial \theta_i} = \frac{\partial \mathbf{P}_{k|k-1}}{\partial \theta_i} - \frac{\partial \mathbf{K}_k}{\partial \theta_i}\,\mathbf{S}_k\mathbf{K}_k^T - \mathbf{K}_k\,\frac{\partial \mathbf{S}_k}{\partial \theta_i}\,\mathbf{K}_k^T - \mathbf{K}_k^T\mathbf{S}_k\left(\frac{\partial \mathbf{K}_k}{\partial \theta_i}\right)^T \tag{73}$$

Equations (69), (70), (71), (72) and (73) together specify a recursive algorithm for computing (67) that can be run alongside the Kalman filter recursions.

4.3 Expectation maximization (EM)

The expectation maximization algorithm (Dempster et al. 1977) is a general method for finding ML and MAP estimates in probabilistic models with latent variables (Bishop 2006). As will be seen, instead of maximizing (67) directly, EM maximizes a series of approximations to it. The derivation of the EM algorithm presented here follows along the lines of (Bishop 2006).

In order to formulate the EM algorithm, let us first introduce an arbitrary probability distribution $q(\mathbf{X})$ over the states. We can then decompose the log-likelihood function as follows:

$$L(\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}\left(q \| p\right) \tag{74}$$

where

$$\mathcal{L}(q, \boldsymbol{\theta}) = \int_{\mathbf{X}} q(\mathbf{X}) \log\left(\frac{p(\mathbf{X}, \mathbf{Y} \,|\, \boldsymbol{\theta})}{q(\mathbf{X})}\right) \mathrm{d}\mathbf{X} \tag{75}$$

$$\mathrm{KL}\left(q \| p\right) = -\int_{\mathbf{X}} q(\mathbf{X}) \log\left(\frac{p(\mathbf{X} \,|\, \mathbf{Y}, \boldsymbol{\theta})}{q(\mathbf{X})}\right) \mathrm{d}\mathbf{X} \tag{76}$$

It is easy to verify the decomposition (74) by substituting

$$\log p(\mathbf{X}, \mathbf{Y} \,|\, \boldsymbol{\theta}) = \log p(\mathbf{X} \,|\, \mathbf{Y}, \boldsymbol{\theta}) + L(\boldsymbol{\theta}) \tag{77}$$

Since $\mathrm{KL}\left(q\|p\right)$, the *Kullback-Leibler divergence* between q and p, is always nonnegative, we see that

$$L(\boldsymbol{\theta}) \geq \mathcal{L}(q, \boldsymbol{\theta}) \tag{78}$$

with equality when

$$\begin{aligned} &\mathrm{KL}\left(q\|p\right) = 0 \\ &\Rightarrow q\left(\mathbf{X}\right) = p(\mathbf{X}\,|\,\mathbf{Y}, \boldsymbol{\theta}) \end{aligned} \tag{79}$$

The nonnegativeness of the Kullback-Leibler divergence can be proved by noting that $-\log$ is a convex function and so *Jensen's inequality* can be applied (Bishop 2006). An alternative proof is presented in (Minka 1998).

We are now ready the define the EM algorithm, which produces a series of estimates $\{\hat{\boldsymbol{\theta}}_j\}$ to the parameter $\boldsymbol{\theta}$ starting from an intitial guess $\hat{\boldsymbol{\theta}}_0$. The two alternating steps of the algorithm are:

1. Given a current estimate $\hat{\boldsymbol{\theta}}_j$ of the parameters, maximize $\mathcal{L}(q, \hat{\boldsymbol{\theta}}_j)$ with respect to the distribution q. As shown by equations (78) and (79), the maximum is obtained with $q^*(\mathbf{X}) = p\Big(\mathbf{X}\Big|\,\mathbf{Y}, \hat{\boldsymbol{\theta}}_j\Big)$, the posterior distribution of the states given the current parameter estimate. After the maximization we have

$$L\Big(\hat{\boldsymbol{\theta}}_j\Big) = \mathcal{L}\bigg(p\Big(\mathbf{X}\Big|\,\mathbf{Y}, \hat{\boldsymbol{\theta}}_j\Big), \hat{\boldsymbol{\theta}}_j\bigg)\,. \tag{80}$$

   This is the *E-step*

2. Maximize $\mathcal{L}\bigg(p\Big(\mathbf{X}\Big|\,\mathbf{Y}, \hat{\boldsymbol{\theta}}_j\Big), \boldsymbol{\theta}\bigg)$ with respect to $\boldsymbol{\theta}$ to obtain a new estimate $\hat{\boldsymbol{\theta}}_{j+1}$. This is the M-step.

We can then formulate a so called *fundamental inequality of EM* (Cappé et al. 2005):

$$L\left(\hat{\boldsymbol{\theta}}_{j+1}\right) - L\left(\hat{\boldsymbol{\theta}}_j\right) \geq \mathcal{L}\bigg(p\Big(\mathbf{X}\Big|\,\mathbf{Y}, \hat{\boldsymbol{\theta}}_j\Big), \hat{\boldsymbol{\theta}}_{j+1}\bigg) - \mathcal{L}\bigg(p\Big(\mathbf{X}\Big|\,\mathbf{Y}, \hat{\boldsymbol{\theta}}_j\Big), \hat{\boldsymbol{\theta}}_j\bigg) \tag{81}$$

which is just the combination of (78) and (80). But it highlights the fact that the likelihood is increased with every new estimate $\hat{\boldsymbol{\theta}}_{j+1}$. Also following from (81) is the fact that if the iterations stop at a certain point, i.e. $\hat{\boldsymbol{\theta}}_{l+1} = \hat{\boldsymbol{\theta}}_l$ at iteration l, then $\mathcal{L}\bigg(p\Big(\mathbf{X}\Big|\,\mathbf{Y}, \hat{\boldsymbol{\theta}}_l\Big), \boldsymbol{\theta}\bigg)$ must be maximal at $\hat{\boldsymbol{\theta}}_l$ and so the gradients of the lower bound and of the likelihood must be zero. Thus $\hat{\boldsymbol{\theta}}_l$ is a *stationary point* of $L(\boldsymbol{\theta})$, i.e a local maximum or a saddle point.

Another property of the lower bound worth stating formally is the following: assume that the likelihood and (76) are continuously differentiable, then

$$\left.\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_i}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_j} = \left.\frac{\partial \mathcal{L}\left(p\left(\mathbf{X}\middle|\mathbf{Y},\hat{\boldsymbol{\theta}}_j\right),\boldsymbol{\theta}\right)}{\partial \theta_i}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_j} \tag{82}$$

This was implicitly clear already from (80) by remembering that $\mathcal{L}$ is a lower bound.

If we substitute $p\left(\mathbf{X}\middle|\mathbf{Y},\hat{\boldsymbol{\theta}}\right)$ for q in (75), we get

$$\begin{aligned}\mathcal{L}(q,\boldsymbol{\theta}) &= \int_{\mathbf{X}} p\left(\mathbf{X}\middle|\mathbf{Y},\hat{\boldsymbol{\theta}}\right)\log p(\mathbf{X},\mathbf{Y}\,|\,\boldsymbol{\theta})\ \mathrm{d}\mathbf{X} - \int_{\mathbf{X}} p\left(\mathbf{X}\middle|\mathbf{Y},\hat{\boldsymbol{\theta}}\right)\log p\left(\mathbf{X}\middle|\mathbf{Y},\hat{\boldsymbol{\theta}}\right)\ \mathrm{d}\mathbf{X} \\ &= \mathfrak{L}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}) + C,\end{aligned} \tag{83}$$

where C is a constant (the differential entropy of $p\left(\mathbf{X}\middle|\mathbf{Y},\hat{\boldsymbol{\theta}}\right)$) and $\mathfrak{L}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})$ can be interpreted as the expectation of the complete-data log-likelihood with respect to the posterior distribution of the states given the current value of the parameter.

4.3.1 EM as a special case of variational Bayes

(Barber 2012; Jordan 1998)

4.3.2 EM in linear-Gaussian SSM:s

(Shumway et al. 1982; Ghahramani 1996) Continuing with the application of EM to the linear-Gaussian state space models, the complete-data log-likelihood function is now

$$\begin{aligned}L(\mathbf{X},\boldsymbol{\theta}) = &\frac{1}{2}\left(\mathbf{x}_0-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_0-\boldsymbol{\mu}\right) + \frac{1}{2}\log|\boldsymbol{\Sigma}| \\ &+ \frac{1}{2}\sum_{k=1}^{N}\left(\mathbf{y}_k-\mathbf{H}\mathbf{x}_k\right)^T\mathbf{R}_{k|k-1}{-}1\left(\mathbf{y}_k-\mathbf{H}\mathbf{x}_k\right) + \frac{N}{2}\log|\mathbf{R}| \\ &+ \frac{1}{2}\sum_{k=1}^{N}\left(\mathbf{x}_k-\mathbf{A}\mathbf{x}_{k-1}\right)^T\mathbf{Q}^{-1}\left(\mathbf{x}_k-\mathbf{A}\mathbf{x}_{k-1}\right) + \frac{N}{2}\log|\mathbf{Q}| \\ &+ C\end{aligned} \tag{84}$$

We can then take the expectation of (84):

$$\mathfrak{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_j) = \left\langle L(\mathbf{X}, \boldsymbol{\theta}) \right\rangle_{\hat{\boldsymbol{\theta}}_j} = \tag{85}$$

$$\begin{aligned}
&\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{P}_{0|N} + (\mathbf{m}_{0|N} - \boldsymbol{\mu})(\mathbf{m}_{0|N} - \boldsymbol{\mu})^T\right)\right] + \log|\boldsymbol{\Sigma}| \\
&+\mathrm{Tr}\left[\mathbf{Q}^{-1}\left(\sum_{k=1}^{N}\left\langle \mathbf{x}_k\mathbf{x}_k^T\right\rangle - \mathbf{A}\sum_{k=1}^{N}\left\langle \mathbf{x}_k\mathbf{x}_{k-1}^T\right\rangle^T - \sum_{k=1}^{N}\left\langle \mathbf{x}_k\mathbf{x}_{k-1}^T\right\rangle\mathbf{A}^T + \mathbf{A}\sum_{k=1}^{N}\left\langle \mathbf{x}_{k-1}\mathbf{x}_{k-1}^T\right\rangle\mathbf{A}^T\right)\right] + N\log \\
&+\mathrm{Tr}\left[\mathbf{R}^{-1}\left(\sum_{k=1}^{N}\mathbf{y}_k\mathbf{y}_k^T - \mathbf{H}\sum_{k=1}^{N}\mathbf{y}_k\left\langle \mathbf{x}_k^T\right\rangle^T - \sum_{k=1}^{N}\mathbf{y}_k\left\langle \mathbf{x}_k^T\right\rangle\mathbf{H}^T + \mathbf{H}\sum_{k=1}^{N}\left\langle \mathbf{x}_k\mathbf{x}_k^T\right\rangle\mathbf{A}^T\right)\right] + N\log|\mathbf{R}|
\end{aligned} \tag{86}$$

Finally, the expectations in (**??**) can be calculated with the combined use of the Kalman filter and the RTS smoother:

$$\left\langle \mathbf{x}_k\right\rangle = \mathbf{m}_{k|N} \tag{87}$$

$$\left\langle \mathbf{x}_k\mathbf{x}_{k|k-1}T\right\rangle = \mathbf{P}_{k|N} + \mathbf{m}_{k|N}(\mathbf{m}_{k|N})^T \tag{88}$$

$$\left\langle \mathbf{x}_k\mathbf{x}_{k-1}^T\right\rangle = \mathbf{C}_{k|N} + \mathbf{m}_{k|N}(\mathbf{m}_{k-1|N})^T, \tag{89}$$

where $\mathbf{m}_{k|N}$ is the mean and $\mathbf{P}_{k|N}$ is the variance of the state $\mathbf{x}_k$ given the observations $\mathbf{y}_1, \ldots, \mathbf{y}_N$. For more specific details see (Gibson et al. 2005). All in all, the E-step of the EM algorithm in linear-Gaussian SSM:s corresponds to computing the matrices in (**??**) with the help of the Kalman filter and the RTS smoother. In (Elliott et al. 1999) a new kind of filter is presented that can compute (**??**) with only forward recursions.

After having calculated the statistics (**??**) $\hat{\boldsymbol{\theta}}$, we proceed to estimate the new value $\boldsymbol{\theta}^*$ by finding the maximum of $\mathfrak{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ in the M-step. The complexity of this step depends of the structure in $\boldsymbol{\theta}_M$. In the case of no structure, the M-step reduces to simple linear regression. Let us now derive the M-step maximization formulas for $\mathbf{A}$, $\mathbf{Q}$, $\mathbf{H}$ and $\mathbf{R}$. To do that, we take the partial derivatives of $\mathfrak{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ and set them

to zero. We get (Ghahramani 1996):

$$\mathbf{A}_{j+1} = \left(\sum_{k=1}^{N} \left\langle \mathbf{x}_k \mathbf{x}_{k-1}^T \right\rangle\right) \left(\sum_{k=1}^{N} \left\langle \mathbf{x}_{k-1} \mathbf{x}_{k-1}^T \right\rangle\right)^{-1} \tag{90a}$$

$$\mathbf{Q}_{j+1} = \sum_{k=1}^{N} \left\langle \mathbf{x}_k \mathbf{x}_k^T \right\rangle - \mathbf{A}_{j+1} \left(\sum_{k=1}^{N} \left\langle \mathbf{x}_k \mathbf{x}_{k-1}^T \right\rangle\right)^T \tag{90b}$$

$$\mathbf{H}_{j+1} = \sum_{k=1}^{N} \mathbf{y}_k \left\langle \mathbf{x}_k^T \right\rangle \left(\sum_{k=1}^{N} \left\langle \mathbf{x}_k \mathbf{x}_k^T \right\rangle\right) - 1 \tag{90c}$$

$$\mathbf{R}_{j+1} = \sum_{k=1}^{N} \mathbf{y}_k \mathbf{y}_k^T - \mathbf{H}_{j+1} \left(\sum_{k=1}^{N} \mathbf{y}_k \left\langle \mathbf{x}_k^T \right\rangle\right)^T \tag{90d}$$

4.3.3 EM in nonlinear-Gaussian SSM:s

Wills 2011 When we want to maximize $\mathfrak{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ w.r.t some other parameters than the ones in $\boldsymbol{\theta}_M$, the situation becomes more complicated. In the general case, no analytical formulas can be found. We therefore seek to maximize $\mathfrak{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ numerically, analogously to how $L(\boldsymbol{\theta})$ was maximized in section 4.2.

Fortunately calculating the gradient of $\mathfrak{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ is straightforward:

$$\begin{aligned}
-2\frac{\partial \mathfrak{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}))}{\partial \theta_i} = &\operatorname{Tr}\left[-\mathbf{Q}^{-1}\frac{\partial \mathbf{Q}}{\partial \theta_i}\mathbf{Q}^{-1}\left(\mathbf{B}_1 - \mathbf{A}\mathbf{B}_2^T - \mathbf{B}_2\mathbf{A}^T + \mathbf{A}\mathbf{B}_3\mathbf{A}^T\right)\right] \\
&+ \operatorname{Tr}\left[\mathbf{Q}^{-1}\left(-\frac{\partial \mathbf{A}}{\partial \theta_i}\mathbf{B}_2^T - \mathbf{B}_2\frac{\partial \mathbf{A}}{\partial \theta_i}^T + \frac{\partial \mathbf{A}}{\partial \theta_i}\mathbf{B}_3\mathbf{A}^T + \mathbf{A}\mathbf{B}_3\frac{\partial \mathbf{A}}{\partial \theta_i}^T\right)\right] \\
&+ \operatorname{Tr}\left[-\mathbf{R}^{-1}\frac{\partial \mathbf{R}}{\partial \theta_i}\mathbf{R}^{-1}\left(\mathbf{B}_4 - \mathbf{H}\mathbf{B}_5^T - \mathbf{B}_5\mathbf{H}^T + \mathbf{H}\mathbf{B}_1\mathbf{H}^T\right)\right] \\
&+ N\operatorname{Tr}\left[\mathbf{Q}^{-1}\frac{\partial \mathbf{Q}}{\partial \theta_i}\right] + N\operatorname{Tr}\left[\mathbf{R}^{-1}\frac{\partial \mathbf{R}}{\partial \theta_i}\right]
\end{aligned} \tag{91}$$

4.3.4 EM in linear-in-the-parameters SSM:s

4.4 Comparisons

4.4.1 Convergence

4.4.2 Computational complexity

(Harvey 1990; Watson 1983; Cappé et al. 2005; Saatci 2011; Olsson et al. 2007; Salakhutdinov et al. 2003)

5 Results

5.1 Tracking a ballistic object on reentry

(Ristic et al. 2004)

5.2 fMRI signal component analysis

(Särkkä et al. 2012)

6 Discussion

6.1 Identifiability

(Haykin 2001; Cappé et al. 2005)

6.2 Stability

(Haykin 2001)

6.3 Dual and joint filtering

(Haykin 2001)

6.4 Particle filtering approaches

(Kantas et al. 2009; Doucet et al. 2001; Lindsten 2010)

*Todo list

References

Arasaratnam, Ienkaran and Simon Haykin (June 2009). "Cubature Kalman Filters". In: *IEEE Transactions on Automatic Control* 54.6, pp. 1254–1269. ISSN: 0018-9286. DOI: 10.1109/TAC.2009.2019800.

— (Aug. 2011). "Cubature Kalman smoothers". In: *Automatica* 47.10, pp. 2245–2250. ISSN: 00051098. DOI: 10.1016/j.automatica.2011.08.005.

Barber, David (2012). *Bayesian Reasoning and Machine Learning*. Cambridge University Press. ISBN: 9780521518147.

Barber, David, A T Cemgil, and S Chiappa (2011). *Bayesian Time Series Models*. Cambridge University Press. ISBN: 9780521196765.

Bishop, C.M. (2006). *Pattern recognition and machine learning*. Springer Verlag. ISBN: 9780387310732.

Cappé, Olivier, Éric Moulines, and T. Rydén (2005). *Inference in hidden Markov models*. Springer Verlag. ISBN: 9780387402642.

Dempster, AP and NM Laird (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society*. 39.1, pp. 1–38.

Doucet, Arnaud, N De Freitas, and N Gordon (2001). *Sequential Monte Carlo Methods in Practice*. Statistics for Engineering and Information Science. Springer. ISBN: 9780387951461.

Durbin, J. and S. J. Koopman (2012). *Time Series Analysis by State Space Methods: Second Edition*. Oxford Statistical Science Series. OUP Oxford. ISBN: 9780199641178.

Elliott, R.J. and Vikram Krishnamurthy (1999). "New finite-dimensional filters for parameter estimation of discrete-time linear Gaussian models". In: *Automatic Control, IEEE Transactions on* 44.5, pp. 938–951.

Gelman, A et al. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC. ISBN: 9781584883883.

Ghahramani, Zoubin (1996). "Parameter estimation for linear dynamical systems". In: *University of Toronto technical report CRG-TR*, pp. 1–6.

Gibson, Stuart and Brett Ninness (Oct. 2005). "Robust maximum-likelihood estimation of multivariable dynamic systems". In: *Automatica* 41.10, pp. 1667–1682. ISSN: 00051098. DOI: 10.1016/j.automatica.2005.05.008.

Harvey, AC (1990). "Estimation procedures for structural time series models". In: *Journal of Forecasting* 9.June 1988, pp. 89–108.

Haykin, Simon (2001). *Kalman filtering and neural networks*. Wiley Online Library. ISBN: 0471221546.

Ito, Kazufumi (2000). "Gaussian filters for nonlinear filtering problems". In: *Automatic Control, IEEE Transactions on* 45.5, pp. 910–927.

Jazwinski, A H (2007). *Stochastic Processes and Filtering Theory*. Dover Books on Electrical Engineering Series. Dover Publications. ISBN: 9780486462745.

Jia, Bin, Ming Xin, and Yang Cheng (Feb. 2012). "Sparse-grid quadrature nonlinear filtering". In: *Automatica* 48.2, pp. 327–341. ISSN: 00051098. DOI: 10.1016/j.automatica.2011.08.057.

Jordan, M I (1998). *Learning in Graphical Models*. Adaptive Computation and Machine Learning. Mit Press. ISBN: 9780262600323.

Julier, Simon and Jeffrey Uhlmann (1997). "A new extension of the Kalman filter to nonlinear systems". In: *Int. Symp. Aerospace/Defense Sensing, Simul. and Controls*. Vol. 3. Spie Bellingham, WA, p. 26.

Kalman, R.E. et al. (1960). "A new approach to linear filtering and prediction problems". In: *Journal of basic Engineering* 82.1, pp. 35–45.

Kantas, N, Arnaud Doucet, and SS Singh (2009). "An overview of sequential Monte Carlo methods for parameter estimation in general state-space models". In: *Proceedings of the IFAC* Ml.

Lindsten, Fredrik (2010). "Identification of mixed linear/nonlinear state-space models". In: *and Control (CDC), 2010 49th IEEE*.

Ljung, L and Torkel Glad (1994). *Modeling of Dynamic Systems*. Prentice Hall Information and System Sciences Series. PTR Prentice Hall. ISBN: 9780135970973.

Mbalawata, Isambi S., Simo Särkkä, and Heikki Haario (2011). "Parameter Estimation in Stochastic Differential Equations with Markov Chain Monte Carlo and Non-Linear Kalman Filtering". In: *Computational Statistics*.

Merwe, Rudolph Van Der (2004). "Sigma-point Kalman filters for probabilistic inference in dynamic state-space models". PhD thesis. Oregon Health & Science University.

Minka, T (1998). "Expectation-Maximization as lower bound maximization". In: *Tutorial published on the web at http://www-white.* 1977, pp. 1–8.

Murphy, KP (2002). "Dynamic Bayesian Networks: Representation, Inference and Learning". In:

Olsson, Rasmus Kongsgaard, Kaare Brandt Petersen, and Tue Lehn-Schiøler (Apr. 2007). "State-Space Models: From the EM Algorithm to a Gradient Approach".

In: *Neural Computation* 19.4, pp. 1097–1111. ISSN: 0899-7667. DOI: 10.1162/neco.2007.19.4.1097.

Pearl, J (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann Publishers. ISBN: 9781558604797.

Rauch, H E, F Tung, and C T Striebel (1965). "Maximum likelihood estimates of linear dynamic systems". In: *AIAA Journal* 3.8, pp. 1445–1450. ISSN: 00011452. DOI: 10.2514/3.3166.

Ristic, B, S Arulampalam, and N Gordon (2004). *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House Radar Library. Artech House. ISBN: 9781580536318.

Saatci, Yunus (2011). "Scalable Inference for Structured Gaussian Process Models". PhD thesis. University of Cambridge.

Salakhutdinov, R and Sam Roweis (2003). "Optimization with EM and expectation-conjugate-gradient". In: *MACHINE LEARNING-*.

Särkkä, Simo (2006). "Recursive bayesian inference on stochastic differential equations". PhD thesis. Helsinki University of Technology. ISBN: 9512281279.

Särkkä, Simo et al. (Jan. 2012). "Dynamic retrospective filtering of physiological noise in BOLD fMRI: DRIFTER." In: *NeuroImage* 60.2, pp. 1517–1527. ISSN: 1095-9572. DOI: 10.1016/j.neuroimage.2012.01.067.

Shumway, R H and D S Stoffer (1982). "An approach to time series smoothing and forecasting using the EM algorithm". In: *Journal of time series analysis* 3.4, pp. 253–264.

Watson, MW (1983). "Alternative Algorithms For The Estimation Of Dynamic Factor, Mimic And Varying Coefficient Regression Models". In: *Journal of Econometrics* 23.

Wills, Adrian (2011). "System identification of nonlinear state-space models". In: *Automatica* November.