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 12, 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.

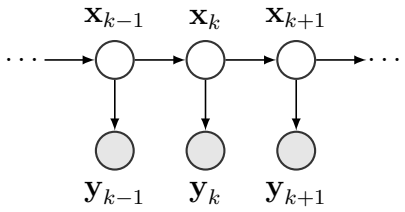

Figure 1: SSM as a graphical model presented with a directed acyclic graph

The value $\mathbf{x}_k \in \mathcal{X} \subset \mathbb{R}^{d_x}$ of the dynamic process at time k is called the *state* at time k. For the measurements we define $\mathbf{y}_k \in \mathcal{Y} \subset \mathbb{R}^{d_y}$. 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 $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^{d_\theta}$ whose value (or distribution) we would like to learn from the measurements.

Traditionally SSMs are specified as a pair of equations specifying the dynamic and measurement models. The most general presentation of discrete-time SSMs is

$$\mathbf{x}_k = \mathbf{f}_{\boldsymbol{\theta},k}\left(\mathbf{x}_{k-1}, \mathbf{q}_{k-1}\right) \tag{4a}$$

$$\mathbf{y}_k = \mathbf{h}_{\boldsymbol{\theta},k}\left(\mathbf{x}_k, \mathbf{r}_k\right). \tag{4b}$$

Here the stochasticity is separated into the noise processes $\mathbf{q}$ and $\mathbf{r}$ which are usually assumed to be zero mean, white and independent of each other. We will restrict ourselves to the case of zero mean, white and additive Gaussian noise and in our case the mappings $\mathbf{f}_{\boldsymbol{\theta},k}$ and $\mathbf{h}_{\boldsymbol{\theta},k}$ and the noise processes $\mathbf{q}$ and $\mathbf{r}$ will be stationary (i.e. independent of k). Thus the SSMs considered in this thesis are of the form

$$\mathbf{x}_k = \mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{x}_{k-1}\right) + \mathbf{q}_{k-1} \qquad \mathbf{q}_{k-1} \sim \mathrm{N}(\mathbf{0}, \mathbf{Q}) \tag{5a}$$

$$\mathbf{y}_k = \mathbf{h}_{\boldsymbol{\theta}}\left(\mathbf{x}_k\right) + \mathbf{r}_k \qquad \mathbf{r}_k \sim \mathrm{N}(\mathbf{0}, \mathbf{R}) \tag{5b}$$

$$\mathbf{x}_0 \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad , \tag{5c}$$

We will assume an implicit dependence of the distributional parameters $\mathbf{Q}$, $\mathbf{R}$,$\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ on $\boldsymbol{\theta}$. Regarding the Gaussian distribution, suppose $\mathbf{x} \in \mathbb{R}^m$ is normally distributed with mean $\mathbf{m}$ and covariance matrix $\mathbf{P}$. We will then use the notation

$$\mathbf{x} \sim \mathrm{N}(\mathbf{m}, \mathbf{P}) \tag{6}$$

$$\Leftrightarrow p(\mathbf{x} \,|\, \mathbf{m}, \mathbf{P}) = \mathrm{N}(\mathbf{x}\,;\, \mathbf{m}, \mathbf{P}) \tag{7}$$

$$= (2\pi)^{-m/2} \left|\mathbf{P}\right|^{-1/2} e^{-1/2(\mathbf{x}-\mathbf{m})^T \mathbf{P}^{-1}(\mathbf{x}-\mathbf{m})} \tag{8}$$

Now clearly the mappings $\mathbf{f}_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{X}$ and $\mathbf{h}_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$ specify the means of the dynamic and the measurement models:

$$p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \boldsymbol{\theta}) = \mathrm{N}\big(\mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{x}_{k-1}\right), \mathbf{Q}\big) \tag{9a}$$

$$p(\mathbf{y}_k \,|\, \mathbf{x}_k, \boldsymbol{\theta}) = \mathrm{N}\big(\mathbf{h}_{\boldsymbol{\theta}}\left(\mathbf{x}_k\right), \mathbf{R}\big) \tag{9b}$$

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

$$p(x_k | x_{k-1}) = x_{k-1} + q_{k-1}, \tag{10}$$

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

$$p(y_k | x_k) = x_k + r_k, \tag{11}$$

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

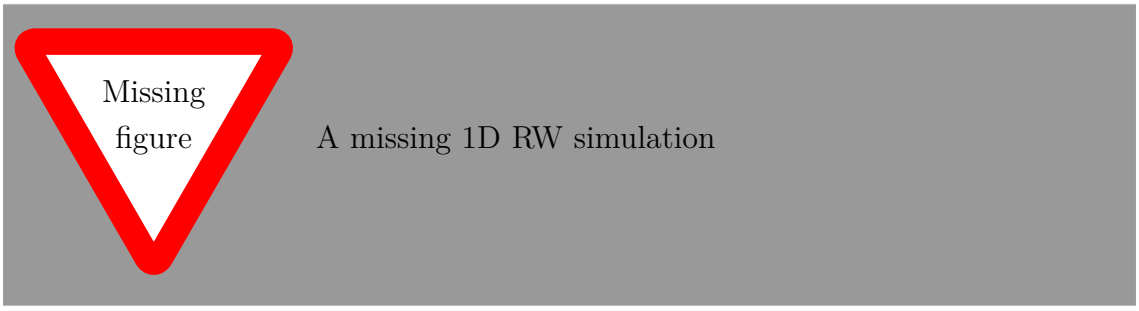

Figure 2: Simulation from the 1D RW model

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

Smoothing distribution $p(\mathbf{x}_k | \mathbf{y}_{1:T}, \boldsymbol{\theta})$ 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

$$\begin{aligned} p\big(\mathbf{x}_k \big| \mathbf{y}_{1:k-1}, \boldsymbol{\theta}\big) &= \int p\big(\mathbf{x}_k, \mathbf{x}_{k-1} \big| \mathbf{y}_{1:k-1}, \boldsymbol{\theta}\big) \ \mathrm{d}\mathbf{x}_{k-1} \\ &= \int p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1})\, p\big(\mathbf{x}_{k-1} \big| \mathbf{y}_{1:k-1}, \boldsymbol{\theta}\big) \ \mathrm{d}\mathbf{x}_{k-1}, \end{aligned} \tag{12}$$

which is known as the *Chapman-Kolmogorov equation* (Särkkä 2006). In this thesis the predictive distributions will be Gaussian or approximated with a Gaussian

$$p\big(\mathbf{x}_k \big| \mathbf{y}_{1:k-1}, \boldsymbol{\theta}\big) = \mathrm{N}\big(\mathbf{x}_k; \mathbf{m}_{k|k-1}, \mathbf{P}_{k|k-1}\big) \tag{13}$$

Filtering distribution

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

check marginal likelihood wording

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

which is called the measurement update equation. In this thesis the filtering distributions will be Gaussian or approximated with a Gaussian

$$p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k}, \boldsymbol{\theta}) = \mathrm{N}\big(\mathbf{x}_k; \mathbf{m}_{k|k}, \mathbf{P}_{k|k}\big) \tag{15}$$

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

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

we get

$$p(\mathbf{x}_k \,|\, \mathbf{y}_{1:T}, \boldsymbol{\theta}) = p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k}, \boldsymbol{\theta}) \int \left[\frac{p(\mathbf{x}_{k+1} \,|\, \mathbf{x}_k, \boldsymbol{\theta})\, p(\mathbf{x}_{k+1} \,|\, \mathbf{y}_{1:T}, \boldsymbol{\theta})}{p(\mathbf{x}_{k+1} \,|\, \mathbf{y}_{1:k}, \boldsymbol{\theta})} \right] \mathrm{d}\mathbf{x}_{k+1}, \tag{16}$$

where $p(\mathbf{x}_{k+1}|\mathbf{y}_{1:k})$ can be computed by equation (12). In this thesis the smoothing distributions will be Gaussian or approximated with a Gaussian

$$p(\mathbf{x}_k|\mathbf{y}_{1:T}) = \mathrm{N}\big(\mathbf{x}_k; \mathbf{m}_{k|T}, \mathbf{P}_{k|T}\big) \tag{17}$$

Marginal likelihood

An important quantity concerning parameter estimation is the marginal likelihood $p(\mathbf{Y}|\boldsymbol{\theta})$. If we're able to compute the distributions

$$p\big(\mathbf{y}_k\big|\mathbf{y}_{1:k-1}, \boldsymbol{\theta}\big) = \int p(\mathbf{y}_k|\mathbf{x}_k, \boldsymbol{\theta})\, p\big(\mathbf{x}_k\big|\mathbf{y}_{1:k-1}, \boldsymbol{\theta}\big)\ \mathrm{d}\mathbf{x}_k \tag{18}$$

then by repeatedly applying the definition of conditional probability we find that the marginal likelihood can be computed from

$$p(\mathbf{Y}|\boldsymbol{\theta}) = p(\mathbf{y}_1|\boldsymbol{\theta}) \prod_{k=2}^{T} p\big(\mathbf{y}_k\big|\mathbf{y}_{1:k-1}, \boldsymbol{\theta}\big) \tag{19}$$

3 State estimation

3.1 Linear-Gaussian State Space Models

Since linear mappings can be described by matrices, stationary linear-Gaussian SSMs are described by the subset of SSMs of the form (5) where

$$\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_{k-1}) = \mathbf{A}\mathbf{x}_{k-1} \tag{20}$$
$$\mathbf{h}_{\boldsymbol{\theta}}(\mathbf{x}_k) = \mathbf{H}\mathbf{x}_k \tag{21}$$

Explain something about linear-Gaussian SSMs

3.1.1 Kalman filter

The recursions are as follows (Kalman et al. 1960; Jazwinski 2007):

Elaborate on the Kalman filter

Predict:

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

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

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

3.1.2 Rauch-Tung-Striebel Smoother

The standard Rauch-Tung-Striebel (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

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

In (Elliott et al. 1999) a new kind of filter is presented that can compute (**??**) with only forward recursions.

Elaborate on the RTS smoother

3.2 Nonlinear-Gaussian SSMs

In the nonlinear case at least one of the mappings $\mathbf{f}_{\boldsymbol{\theta}}$ and $\mathbf{h}_{\boldsymbol{\theta}}$ in (5) is nonlinear. Unfortunately in this case computing the filtering distributions in closed form becomes intractable and one has to resort to approximations. Finding good approximations is however extremely valuable since a great many physically modeled dynamical phenoma could be forumulated as nonlinear-Gaussian SSM.

The types of approximate filtering (and smoothing) solutions can be put into two categories: particle filtering (or sequential Monte Carlo), which is asymptotically exact but computationally demanding and different kinds of deterministic approximation methods that are analytically inexact but less computationally demanding. We will only focus on the latter category.

Mention EKF

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

$$\mathbf{a} \sim \mathrm{N}(\mathbf{m}, \boldsymbol{\Sigma}_a) \tag{25}$$
$$\mathbf{b}|\mathbf{a} \sim \mathrm{N}\big(\mathbf{f}(\mathbf{a}), \boldsymbol{\Sigma}_{b|a}\big) \tag{26}$$

then

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

is only Gaussian if $\mathbf{f}(\mathbf{a})$ is linear. Let the Gaussian approximation to (27) be

$$p\left(\begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix}\right) \approx \mathrm{N}\left(\begin{bmatrix}\boldsymbol{\mu}_a\\ \boldsymbol{\mu}_b\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab}\\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb}\end{bmatrix}\right) \tag{28}$$

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

$$\boldsymbol{\mu}_a = \mathbf{m} \tag{29}$$
$$\boldsymbol{\Sigma}_{aa} = \boldsymbol{\Sigma}_a \tag{30}$$
$$\boldsymbol{\mu}_b = \int \mathbf{b} p(\mathbf{b}) \ \mathrm{d}\mathbf{b} \tag{31}$$
$$\boldsymbol{\Sigma}_{bb} = \int (\mathbf{b} - \boldsymbol{\mu}_b)(\mathbf{b} - \boldsymbol{\mu}_b)^T p(\mathbf{b}) \ \mathrm{d}\mathbf{b} \tag{32}$$

Both (31) and (32) can be written in terms of (25) and (26). To see this, let us rewrite (31) as

$$\begin{aligned}\boldsymbol{\mu}_b &= \int \mathbf{b} p(\mathbf{b}) \ \mathrm{d}\mathbf{b}\\ &= \int \mathbf{b} \int p(\mathbf{b}|\mathbf{a})\, p(\mathbf{a}) \ \mathrm{d}\mathbf{a} \ \mathrm{d}\mathbf{b}\\ &= \int \int \mathbf{b} p(\mathbf{b}|\mathbf{a}) \ \mathrm{d}\mathbf{b} p(\mathbf{a}) \ \mathrm{d}\mathbf{a}\\ &= \int \mathbf{f}(\mathbf{a}) \mathrm{N}(\mathbf{m}, \boldsymbol{\Sigma}_a) \ \mathrm{d}\mathbf{a}\end{aligned} \tag{33}$$

and (32) as

$$\begin{aligned}\boldsymbol{\Sigma}_{bb} &= \int \mathbf{b}\mathbf{b}^T p(\mathbf{b}) \ \mathrm{d}\mathbf{b} - \boldsymbol{\mu}_b \boldsymbol{\mu}_b^T\\ &= \int \mathbf{f}(\mathbf{a})\mathbf{f}(\mathbf{a})^T p(\mathbf{a}) \ \mathrm{d}\mathbf{a} - \boldsymbol{\mu}_b \boldsymbol{\mu}_b^T\\ &\quad + \int \int [(\mathbf{b} - \mathbf{f}(\mathbf{a}))(\mathbf{b} - \mathbf{f}(\mathbf{a}))^T] p(\mathbf{b}|\mathbf{a}) \ \mathrm{d}\mathbf{b} p(\mathbf{a}) \ \mathrm{d}\mathbf{a}\\ &= \int (\mathbf{f}(\mathbf{a}) - \boldsymbol{\mu}_b)(\mathbf{f}(\mathbf{a}) - \boldsymbol{\mu}_b)^T \mathrm{N}(\mathbf{m}, \boldsymbol{\Sigma}_a) \ \mathrm{d}\mathbf{a} + \boldsymbol{\Sigma}_{b|a}.\end{aligned} \tag{34}$$

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

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

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

$$\begin{aligned}
p(\mathbf{x}_{k-1},\mathbf{x}_k\,|\,\mathbf{Y}) &= p(\mathbf{x}_{k-1}\,|\,\mathbf{x}_k,\mathbf{Y}_{1:k-1})\, p(\mathbf{x}_k\,|\,\mathbf{Y}) \\
&= \frac{p(\mathbf{x}_{k-1},\mathbf{x}_k\,|\,\mathbf{Y}_{1:k-1})\, p(\mathbf{x}_k\,|\,\mathbf{Y})}{p(\mathbf{x}_k\,|\,\mathbf{Y}_{1:k-1})},
\end{aligned} \tag{36}$$

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

$$p\Big(\mathbf{x}_{k-1}\,\Big|\,\mathbf{Y}_{1:k-1},\hat{\boldsymbol{\theta}}_j\Big) \approx \mathrm{N}\big(\mathbf{m}_{k-1|k-1},\mathbf{P}_{k-1|k-1}\big)\,. \tag{37}$$

The Gaussian approximation to

$$p(\mathbf{x}_{k-1},\mathbf{x}_k\,|\,\mathbf{Y}_{1:k-1}) = \mathrm{N}\big(\mathbf{x}_k\,\big|\,\mathbf{f}(\mathbf{x}_{k-1}),\mathbf{Q}\big)\, p(\mathbf{x}_{k-1}\,|\,\mathbf{Y}_{1:k-1}) \tag{38}$$

is then given by application of equations (33), (34) and (35)

$$\mathbf{m}_{k|k-1} = \int \mathbf{f}(\mathbf{x}_{k-1})\mathrm{N}\big(\mathbf{x}_{k-1}\,\big|\,\mathbf{m}_{k-1|k-1},\mathbf{P}_{k-1|k-1}\big)\ \mathrm{d}\mathbf{x}_{k-1} \tag{39}$$

$$\begin{aligned}
\mathbf{P}_{k|k-1} = \int &(\mathbf{f}(\mathbf{x}_{k-1})-\mathbf{m}_{k|k-1})(\mathbf{f}(\mathbf{x}_{k-1})-\mathbf{m}_{k|k-1})^T \\
&\mathrm{N}\big(\mathbf{x}_{k-1}\,\big|\,\mathbf{m}_{k-1|k-1},\mathbf{P}_{k-1|k-1}\big)\ \mathrm{d}\mathbf{x}_{k-1} + \mathbf{Q}
\end{aligned} \tag{40}$$

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

so that the approximation is

$$p\Big(\mathbf{x}_{k-1},\mathbf{x}_k\,\Big|\,\mathbf{Y}_{1:k-1},\hat{\boldsymbol{\theta}}_j\Big) \approx \mathrm{N}\left(\begin{bmatrix}\mathbf{m}_{k-1|k-1}\\ \mathbf{m}_{k|k-1}\end{bmatrix}, \begin{bmatrix}\mathbf{P}_{k-1|k-1} & \mathbf{C}_{k^-}\\ \mathbf{C}_{k^-}^T & \mathbf{P}_{k|k-1}\end{bmatrix}\right) \tag{42}$$

In order to calculate (39) and (40) we also need a Gaussian approximation for the joint distribution of the current state and measurement given the previous measurements

$$\begin{aligned} p(\mathbf{x}_k, \mathbf{y}_k \,|\, \mathbf{Y}_{1:k-1}) &= \mathrm{N}\big(\mathbf{y}_k \,\big|\, \mathbf{h}(\mathbf{x}_k), \mathbf{R}\big)\, p(\mathbf{x}_k \,|\, \mathbf{Y}_{1:k-1}) \\ &\approx \mathrm{N}\left(\begin{bmatrix} \mathbf{m}_{k|k-1} \\ \boldsymbol{\mu}_k \end{bmatrix}, \begin{bmatrix} \mathbf{P}_{k|k-1} & \mathbf{C}_k \\ \mathbf{C}_k^T & \mathbf{S}_k \end{bmatrix} \right). \end{aligned} \tag{43}$$

Applying equations (33), (34) and (35) again, we get

$$\boldsymbol{\mu}_k = \int \mathbf{h}(\mathbf{x}_k) \mathrm{N}\big(\mathbf{x}_k \,\big|\, \mathbf{m}_{k|k-1}, \mathbf{P}_{k|k-1}\big)\ \mathrm{d}\mathbf{x}_k \tag{44}$$

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

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

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

$$\mathbf{m}_{k|k} = \mathbf{m}_{k|k-1} + \mathbf{C}_k \mathbf{S}_k^{-1} \left(\mathbf{y}_k - \boldsymbol{\mu}_k\right) \tag{47}$$

$$\mathbf{P}_{k|k} = \mathbf{P}_{k|k-1} - \mathbf{C}_k \mathbf{S}_k^{-1} \mathbf{C}_k^T. \tag{48}$$

Again using the formula for the conditional of jointly Gaussian variables we get from (42)

$$p(\mathbf{x}_{k-1} \,|\, \mathbf{x}_k, \mathbf{Y}_{1:k-1}) \approx \mathrm{N}(\mathbf{m}_2, \mathbf{P}_2) \tag{49}$$

$$\mathbf{G}_{k-1} = \mathbf{C}_k \mathbf{P}_{k|k-1}^{-1} \tag{50}$$

$$\mathbf{m}_2 = \mathbf{m}_{k-1|k-1} + \mathbf{G}_{k-1}(\mathbf{x}_k - \mathbf{m}_{k|k-1}) \tag{51}$$

$$\mathbf{P}_2 = \mathbf{P}_{k-1|k-1} - \mathbf{G}_{k-1} \mathbf{P}_{k|k-1} \mathbf{G}_{k-1}^T \tag{52}$$

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

$$\begin{aligned} p(\mathbf{x}_{k-1}, \mathbf{x}_k \,|\, \mathbf{Y}) &= p(\mathbf{x}_{k-1} \,|\, \mathbf{x}_k, \mathbf{Y}_{1:k-1})\, p(\mathbf{x}_k \,|\, \mathbf{Y}) \\ &\approx \mathrm{N}\left(\begin{bmatrix} \mathbf{m}_{k-1|N} \\ \mathbf{m}_{k|N} \end{bmatrix}, \begin{bmatrix} \mathbf{P}_{k-1|N} & \mathbf{D}_k \\ \mathbf{D}_k^T & \mathbf{P}_{k|N} \end{bmatrix} \right) \end{aligned} \tag{53}$$

where

$$\mathbf{D}_k = \mathbf{G}_{k-1} \mathbf{P}_{k|N} \tag{54}$$

$$\mathbf{m}_{k-1|N} = \mathbf{m}_{k-1|k-1} + \mathbf{G}_{k-1} \left(\mathbf{m}_{k|N} - \mathbf{m}_{k|k-1}\right) \tag{55}$$

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

3.2.2 Quadrature and cubature

(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, which is given by Bayes' rule as

$$p(\boldsymbol{\theta}\,|\,\mathbf{Y}) = \frac{p(\mathbf{Y}\,|\,\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{p(\mathbf{Y})}. \tag{57}$$

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 $p(\mathbf{Y})$, since it's constant with respect to the parameters. A point estimate that maximizes the posterior distribution is called a *maximum a posteriori* (MAP) estimate. Since the logarithm is a strictly monotonic function, maximizing a function is the same as maximizing its logarithm. Thus the MAP estimate $\boldsymbol{\theta}^*$ is given by

why? examples?

$$\boldsymbol{\theta}_{\mathrm{MAP}} = \operatorname{argmax}_{\boldsymbol{\theta}} \left[\underbrace{\log p(\mathbf{Y}\,|\,\boldsymbol{\theta})}_{\ell(\boldsymbol{\theta})} + \log p(\boldsymbol{\theta}) \right] \tag{58}$$

In the case of a flat (constant and thus improper) prior distribution, $p(\boldsymbol{\theta})$, the MAP estimate converges to the *maximum likelihood* (ML) estimate

$$\boldsymbol{\theta}_{\mathrm{ML}} = \operatorname{argmax}_{\boldsymbol{\theta}} \left[\ell(\boldsymbol{\theta}) \right] \tag{59}$$

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.

Explain MAP in both cases

Among different point estimates, the maximum likelihood estimator has good statistical properties. Let us denote the true parameter value, the value that the data was generated with, with $\boldsymbol{\theta}_\star$ and let T denote the amount of observations. Then provided that some conditions of not very restricting nature hold, we can state the following asymptotic properties for the ML estimate $\boldsymbol{\theta}_{\mathrm{ML}}$:

Modify to reflect p.465 in Cappé

Strong consistency

An important property for an estimator, which says that the estimator tends to the true value as the amount of data tends to infinity:

$$\ell_T\left(\boldsymbol{\theta}_{\mathrm{ML}}\right) \xrightarrow{\text{a.s.}} \ell\left(\boldsymbol{\theta}_\star\right), \quad \text{when } T \to \infty, \tag{60}$$

where ℓ_T is the likelihood function after T measurements and ℓ is a continuous deterministic function with a unique global maximum at $\boldsymbol{\theta}_\star$.

Asymptotic normality

This property gives us the means to compute asymptotic error bounds for the estimate:

$$\sqrt{T}\left(\boldsymbol{\theta}_{\mathrm{ML}} - \boldsymbol{\theta}_\star\right) \xrightarrow{D} \mathrm{N}\big(\mathbf{0}, \mathcal{I}^{-1}\left(\boldsymbol{\theta}_\star\right)\big), \quad \text{when } T \to \infty, \tag{61}$$

where $\mathcal{I}\left(\boldsymbol{\theta}_\star\right)$ is the *Fischer information matrix* evaluated at $\boldsymbol{\theta}_\star$

Efficiency

When the amount of information tends to infinity, the ML-estimate achieves the Cramér-Rao lower bound, i.e. no other consistent estimator has lower asymptotic mean-squared-error.

4.1.1 Identifiability

Intuitively, any parameters $\boldsymbol{\theta}$, $\boldsymbol{\theta}' \in \Theta$ cannot be distinguished from each other with maximum likelihood estimation if

$$p(\mathbf{Y}\,|\,\boldsymbol{\theta}) = p\big(\mathbf{Y}\,\big|\,\boldsymbol{\theta}'\big), \tag{62}$$

i.e., if the same data can arise with two (or more) separate parameter values. (Haykin 2001; Cappé et al. 2005)

Elaborate on identifiability

4.2 Gradient based nonlinear optimization

This is the classical way of solving the parameter estimation problem. It consists of computing the gradient of the log-likelihood function $\ell(\boldsymbol{\theta})$ and then using some non-linear optimization method to find a *local* maximum to it (Mbalawata et al. 2011; Cappé et al. 2005). An efficient non-linear optimization algorithm is the scaled conjugate gradient method (Mbalawata et al. 2011).

By marginalizing the joint distribution of equation (23) we get

$$p\big(\mathbf{y}_k\,\big|\,\mathbf{y}_{1:k-1}, \boldsymbol{\theta}\big) = \mathrm{N}\big(\mathbf{y}_k\,;\, \mathbf{H}\mathbf{m}_{k|k-1}, \mathbf{S}_k\big). \tag{63}$$

Applying equation (19) and taking the logarithm then gives

$$\ell(\boldsymbol{\theta}) = -\frac{1}{2}\sum_{k=1}^{N}\log|\mathbf{S}_k| - \frac{1}{2}\sum_{k=1}^{N}\left(\mathbf{y}_k - \mathbf{H}\mathbf{m}_{k|k-1}\right)^T \mathbf{S}_k^{-1}\left(\mathbf{y}_k - \mathbf{H}\mathbf{m}_{k|k-1}\right) + C, \quad (64)$$

where C is a constant that doesn't depend on $\boldsymbol{\theta}$ and thus can be ignored in the maximization. Employing an efficient numerical optimization method generally requires that the gradient of the objective function is available. There are at least two seemingly quite different methods for computing the gradient of $\ell(\boldsymbol{\theta})$. The first one proceeds straightforwardly by taking the partial derivatives of $\ell(\boldsymbol{\theta})$. As will soon be demonstrated, this leads to some additional recursive formulas which allow computing the gradient in parallel with the Kalman filter. The second method needs the smoothing distributions with the cross-timestep covariances of equation (24d) and it can be easily computed with the expectation maximization machinery that will be introduced later. These two methods can be proved to compute the exact same quantity. At this point we will focus on the first one.

In order to calculate the gradient of $\ell(\boldsymbol{\theta})$, we can take the partial derivatives of it w.r.t every parameter θ_i in $\boldsymbol{\theta}$:

$$\begin{aligned}\frac{\partial\ell(\boldsymbol{\theta})}{\partial\theta_i} =& -\frac{1}{2}\sum_{k=1}^{N}\operatorname{Tr}\left(\mathbf{S}_k^{-1}\frac{\partial\mathbf{S}_k}{\partial\theta_i}\right) \\ &+ \sum_{k=1}^{N}\left(\mathbf{H}_k\frac{\partial\mathbf{m}_{k|k-1}}{\partial\theta_i}\right)^T\mathbf{S}_k^{-1}\left(\mathbf{y}_k - \mathbf{H}\mathbf{m}_{k|k-1}\right) \\ &+ \frac{1}{2}\sum_{k=1}^{N}\left(\mathbf{y}_k - \mathbf{H}\mathbf{m}_{k|k-1}\right)^T\mathbf{S}_k^{-1}\left(\frac{\partial\mathbf{S}_k}{\partial\theta_i}\right)\mathbf{S}_k^{-1}\left(\mathbf{y}_k - \mathbf{H}\mathbf{m}_{k|k-1}\right)\end{aligned} \quad (65)$$

From the Kalman filter recursions (22) we find out that

$$\frac{\partial\mathbf{S}_k}{\partial\theta_i} = \mathbf{H}\frac{\partial\mathbf{P}_{k|k-1}}{\partial\theta_i}\mathbf{H} + \frac{\partial\mathbf{R}}{\partial\theta_i} \quad (66)$$

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

$$\frac{\partial\mathbf{m}_{k|k-1}}{\partial\theta_i} = \frac{\partial\mathbf{A}}{\partial\theta_i}\mathbf{m}_{k-1|k-1} + \mathbf{A}\frac{\partial\mathbf{m}_{k-1|k-1}}{\partial\theta_i} \quad (67)$$

$$\begin{aligned}\frac{\partial\mathbf{P}_{k|k-1}}{\partial\theta_i} =& \frac{\partial\mathbf{A}}{\partial\theta_i}\mathbf{P}_{k-1|k-1}\mathbf{A}^T + \mathbf{A}\frac{\partial\mathbf{P}_{k-1|k-1}}{\partial\theta_i}\mathbf{A}^T \\ &+ \mathbf{A}\mathbf{P}_{k-1|k-1}\left(\frac{\partial\mathbf{A}}{\partial\theta_i}\right)^T + \frac{\partial\mathbf{Q}}{\partial\theta_i}\end{aligned} \quad (68)$$

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

$$\frac{\partial \mathbf{K}_k}{\partial \theta_i} = \frac{\partial \mathbf{P}_{k|k-1}}{\partial \theta_i}\mathbf{H}^T\mathbf{S}_k^{-1} - \mathbf{P}_{k|k-1}\mathbf{H}^T\mathbf{S}_k^{-1}\frac{\partial \mathbf{S}_k}{\partial \theta_i}\mathbf{S}_k^{-1} \tag{69}$$

$$\frac{\partial \mathbf{m}_{k|k}}{\partial \theta_i} = \frac{\partial \mathbf{m}_{k|k-1}}{\partial \theta_i} + \frac{\partial \mathbf{K}_k}{\partial \theta_i}\left(\mathbf{y}_k - \mathbf{H}\mathbf{m}_{k|k-1}\right) - \mathbf{K}_k\mathbf{H}\frac{\partial \mathbf{m}_{k|k-1}}{\partial \theta_i} \tag{70}$$

$$\frac{\partial \mathbf{P}_{k|k}}{\partial \theta_i} = \frac{\partial \mathbf{P}_{k|k-1}}{\partial \theta_i} - \frac{\partial \mathbf{K}_k}{\partial \theta_i}\mathbf{S}_k\mathbf{K}_k^T - \mathbf{K}_k\frac{\partial \mathbf{S}_k}{\partial \theta_i}\mathbf{K}_k^T - \mathbf{K}_k\mathbf{S}_k\left(\frac{\partial \mathbf{K}_k}{\partial \theta_i}\right)^T \tag{71}$$

Equations (67), (68), (69), (70) and (71) together specify a recursive algorithm for computing (65) that can be run alongside the Kalman filter recursions. As noted in Cappé et al. (2005), these equations are sometimes known as the *sensitivity equations* and they are derived at least in Gupta et al. (1974) and Mbalawata et al. (2011).

Elaborate on nonlinear programming

4.3 Expectation maximization (EM)

The expectation maximization (EM) algorithm (Dempster et al. 1977) is a general method for finding ML and MAP estimates in probabilistic models with missing data or latent variables (Bishop 2006; Barber 2012). As will be seen, instead of maximizing (64) directly, EM alternates between forming a variational lower bound and maximizing it. We shall use $\langle\cdot\rangle_q \equiv \int \cdot\, q(z)\ \mathrm{d}z$ to denote the expectation over some arbitrary distribution $q(z)$. Let us introduce a "variational" distribution $\tilde{p}\big(\mathbf{X}\big|\boldsymbol{\theta}'\big)$ over the states, parameterized with $\boldsymbol{\theta}'$ (not necessarily related to $\boldsymbol{\theta}$). Noting now that $p(\mathbf{X}|\mathbf{Y},\boldsymbol{\theta}) = p(\mathbf{X},\mathbf{Y}|\boldsymbol{\theta})\,/p(\mathbf{Y}|\boldsymbol{\theta})$ and that $\ell(\boldsymbol{\theta}) \equiv \log p(\mathbf{Y}|\boldsymbol{\theta})$ is independent of $\mathbf{X}$ we can then perform the following decomposition on the log likelihood:

$$\begin{aligned}
\ell(\boldsymbol{\theta}) &= \log p(\mathbf{X},\mathbf{Y}|\boldsymbol{\theta}) - \log p(\mathbf{X}|\mathbf{Y},\boldsymbol{\theta}) \\
&= \left\langle \log p(\mathbf{X},\mathbf{Y}|\boldsymbol{\theta})\right\rangle_{\tilde{p}} - \left\langle \log p(\mathbf{X}|\mathbf{Y},\boldsymbol{\theta})\right\rangle_{\tilde{p}} \\
&= \underbrace{\left\langle \log p(\mathbf{X},\mathbf{Y}|\boldsymbol{\theta}) - \log \tilde{p}\big(\mathbf{X}\big|\boldsymbol{\theta}'\big)\right\rangle_{\tilde{p}}}_{\mathcal{L}(\tilde{p},\boldsymbol{\theta})} \underbrace{- \left\langle \log p(\mathbf{X}|\mathbf{Y},\boldsymbol{\theta}) - \log \tilde{p}\big(\mathbf{X}\big|\boldsymbol{\theta}'\big)\right\rangle_{\tilde{p}}}_{\mathrm{KL}\left(\tilde{p}\|p(\mathbf{X}|\mathbf{Y},\boldsymbol{\theta})\right)}
\end{aligned} \tag{72}$$

The important step here is taking the expectation over $\tilde{p}\big(\mathbf{X}\big|\boldsymbol{\theta}'\big)$, since the *complete-data log-likelihood* $\log p(\mathbf{X},\mathbf{Y}|\boldsymbol{\theta})$ cannot be evaluated as $\mathbf{X}$ is unobserved. Since $\mathrm{KL}\left(\tilde{p}\|p(\mathbf{X}|\mathbf{Y},\boldsymbol{\theta})\right)$, the *Kullback-Leibler divergence* between $\tilde{p}\big(\mathbf{X}\big|\boldsymbol{\theta}'\big)$ and $p(\mathbf{X}|\mathbf{Y},\boldsymbol{\theta})$, is always nonnegative, we see that

$$\ell(\boldsymbol{\theta}) \geq \mathcal{L}(\tilde{p},\boldsymbol{\theta}) \tag{73}$$

with equality when

$$\tilde{p}\big(\mathbf{X}\big|\boldsymbol{\theta}'\big) = p(\mathbf{X}|\mathbf{Y},\boldsymbol{\theta})\,, \tag{74}$$

i.e. the posterior distribution of the states with equal parameter value $\boldsymbol{\theta}' = \boldsymbol{\theta}$. Considered as a functional of only $\tilde{p}$, clearly $\mathcal{L}(\tilde{p},\boldsymbol{\theta})$ is maximized and $\mathrm{KL}\left(\tilde{p}\|p(\mathbf{X}|\mathbf{Y},\boldsymbol{\theta})\right)$

vanishes by (74). 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).

Let us take a closer at the the first term in (72) with $\tilde{p}\big(\mathbf{X}\big|\boldsymbol{\theta}'\big) = p\big(\mathbf{X},\mathbf{Y}\big|\boldsymbol{\theta}'\big)$:

$$\mathcal{L}\Big(p\big(\mathbf{X},\mathbf{Y}\big|\boldsymbol{\theta}'\big),\boldsymbol{\theta}\Big) = \underbrace{\big\langle \log p(\mathbf{X},\mathbf{Y}\,|\boldsymbol{\theta})\big\rangle_{p\left(\mathbf{X},\mathbf{Y}\,\middle|\,\boldsymbol{\theta}'\right)}}_{\mathfrak{L}\left(\boldsymbol{\theta}',\boldsymbol{\theta}\right)} - \Big\langle \log p\big(\mathbf{X},\mathbf{Y}\big|\boldsymbol{\theta}'\big)\Big\rangle_{p\left(\mathbf{X},\mathbf{Y}\,\middle|\,\boldsymbol{\theta}'\right)}. \quad (75)$$

Clearly the latter term (the differential entropy of $p\big(\mathbf{X},\mathbf{Y}\big|\boldsymbol{\theta}'\big)$) is constant with respect to $\boldsymbol{\theta}$, so that maximizing $\mathcal{L}$ with respect to $\boldsymbol{\theta}$ amounts to maximizing $\mathfrak{L}\big(\boldsymbol{\theta}',\boldsymbol{\theta}\big)$, the *expected complete-data log-likelihood*, with respect to $\boldsymbol{\theta}$.

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

E-step

Given the current estimate $\boldsymbol{\theta}_j$ of the parameters, compute

$$\tilde{p}_{j+1} = \arg\max_{\tilde{p}} \mathcal{L}\big(\tilde{p},\boldsymbol{\theta}_j\big)\,. \quad (76)$$

As stated, the maximum is obtained with $\tilde{p}_{j+1} = p\big(\mathbf{X}\big|\mathbf{Y},\boldsymbol{\theta}_j\big)$, the posterior distribution of the states given the current parameter estimate. After the maximization we have

$$\begin{aligned}\ell\big(\boldsymbol{\theta}_j\big) &= \mathcal{L}\Big(p\big(\mathbf{X}\big|\mathbf{Y},\boldsymbol{\theta}_j\big),\boldsymbol{\theta}_j\Big) \quad (77)\\ &= \mathfrak{L}\big(\boldsymbol{\theta}_j,\boldsymbol{\theta}_j\big) + \texttt{const}\end{aligned}$$

M-step

Set

$$\begin{aligned}\boldsymbol{\theta}_{j+1} &= \arg\max_{\boldsymbol{\theta}} \mathcal{L}\big(\tilde{p}_{j+1},\boldsymbol{\theta}\big) \quad (78)\\ &= \arg\max_{\boldsymbol{\theta}} \mathfrak{L}\big(\boldsymbol{\theta}_j,\boldsymbol{\theta}\big)\,.\end{aligned}$$

We are now in a position to formulate the so called *fundamental inequality of EM* (Cappé et al. 2005):

$$\ell\big(\boldsymbol{\theta}_{j+1}\big) - \ell\big(\boldsymbol{\theta}_j\big) \geq \mathcal{L}\Big(p\big(\mathbf{X}\big|\mathbf{Y},\boldsymbol{\theta}_j\big),\boldsymbol{\theta}_{j+1}\Big) - \mathcal{L}\Big(p\big(\mathbf{X}\big|\mathbf{Y},\boldsymbol{\theta}_j\big),\boldsymbol{\theta}_j\Big) \quad (79)$$

which is just the combination of (73) and (77). But it highlights the fact that *the likelihood is increased or unchanged with every new estimate* $\boldsymbol{\theta}_{j+1}$. Also following from (79) is the fact that if the iterations stop at a certain point, i.e. $\boldsymbol{\theta}_{l+1} = \boldsymbol{\theta}_l$ at iteration l, then $\mathcal{L}\big(p(\mathbf{X}|\mathbf{Y},\boldsymbol{\theta}_l),\boldsymbol{\theta}\big)$ must be maximal at $\boldsymbol{\theta}_l$ and so the gradients of the lower bound and of the likelihood must be zero. Thus $\boldsymbol{\theta}_l$ is a *stationary point* of $\ell(\boldsymbol{\theta})$, i.e a local maximum or a saddle point.

4.3.1 EM as a special case of variational Bayes

Ensure that notation matches prev chapter

(Barber 2012; Jordan 1998) Variational Bayes (VB) is a fully Bayesian methodology where one seeks for an approximation to the parameter posterior

$$p(\boldsymbol{\theta}|\mathbf{Y}) \propto \int_{\mathcal{X}} p(\mathbf{X}, \mathbf{Y}|\boldsymbol{\theta}) \; \mathrm{d}\mathbf{X} p(\boldsymbol{\theta}) \tag{80}$$

As mentioned earlier, finding this distribution is commonly intractable, so in VB we assume a factorized form for the joint posterior of states and parameters

$$p(\mathbf{X}, \boldsymbol{\theta}|\mathbf{Y}) \approx q(\mathbf{X})\, q(\boldsymbol{\theta}) \tag{81}$$

and the task is then to find the best approximation with respect to the KL divergence between the true posterior and the approximation

$$\mathrm{KL}\left(q(\mathbf{X})\, q(\boldsymbol{\theta}) \| p(\mathbf{X}, \boldsymbol{\theta}|\mathbf{Y})\right) = \left\langle q(\mathbf{X})\right\rangle_{q(\mathbf{X})} + \left\langle q(\mathbf{X})\right\rangle_{q(\mathbf{X})} - \left\langle p(\mathbf{X}, \boldsymbol{\theta}|\mathbf{Y})\right\rangle_{q(\mathbf{X})q(\boldsymbol{\theta})}. \tag{82}$$

Using $p(\mathbf{X}, \boldsymbol{\theta}|\mathbf{Y}) = p(\mathbf{X}, \boldsymbol{\theta}, \mathbf{Y})\,/p(\mathbf{Y})$ equation (82) gives

$$\ell(\boldsymbol{\theta}) \geq \left\langle p(\mathbf{X}, \boldsymbol{\theta}, \mathbf{Y})\right\rangle_{q(\mathbf{X})q(\boldsymbol{\theta})} - \left\langle q(\mathbf{X})\right\rangle_{q(\mathbf{X})} - \left\langle q(\mathbf{X})\right\rangle_{q(\mathbf{X})} \tag{83}$$

and thus minimizing the KL divergence is equivalent to finding the tightest lower bound to the log likelihood. Analogously to EM, minimizing the KL divergence is done iteratively keeping $q(\boldsymbol{\theta})$ fixed and minimizing w.r.t $q(\mathbf{X})$ in the "E"-step and vice versa in the "M"-step:

E-step

$$q^{\text{new}}(\mathbf{X}) = \arg\max_{q(\mathbf{X})} \left(\mathrm{KL}\left(q(\mathbf{X})\, q^{\text{old}}(\boldsymbol{\theta}) \| p(\mathbf{X}, \boldsymbol{\theta}|\mathbf{Y})\right)\right) \tag{84}$$

M-step

$$q^{\text{new}}(\boldsymbol{\theta}) = \arg\max_{q(\boldsymbol{\theta})} \left(\mathrm{KL}\left(q^{\text{new}}(\mathbf{X})\, q(\boldsymbol{\theta}) \| p(\mathbf{X}, \boldsymbol{\theta}|\mathbf{Y})\right)\right) \tag{85}$$

Let us then suppose that we only wish to find the MAP point estimate $\boldsymbol{\theta}^*$. This can be accomplished by assuming a delta function form $q(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ for the parameter factor in the joint distribution of states and parameters (81). With this assumption equation (83) becomes

$$p(\mathbf{Y}|\boldsymbol{\theta}^*) \geq \left\langle p(\mathbf{X}, \boldsymbol{\theta}^*, \mathbf{Y})\right\rangle_{q(\mathbf{X})q(\boldsymbol{\theta})} - \left\langle q(\mathbf{X})\right\rangle_{q(\mathbf{X})} + \texttt{const} \tag{86}$$

and the "M"-step (85) can then be written as

$$\boldsymbol{\theta}^* = \mathrm{argmax}_{\boldsymbol{\theta}} \left(\left\langle \log p(\mathbf{X}, \mathbf{Y}|\boldsymbol{\theta})\right\rangle_{q(\mathbf{X})} + \log p(\boldsymbol{\theta})\right). \tag{87}$$

If the point estimate is plugged in the "E"-step equation (84) we have

$$q^{\text{new}}(\mathbf{X}) \propto p(\mathbf{X}, \mathbf{Y}|\boldsymbol{\theta}^*) \propto p(\mathbf{X}|\mathbf{Y}, \boldsymbol{\theta}^*) \tag{88}$$

4.3.2 Partial E and M steps

4.3.3 Gradient computation

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

$$\left.\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_i}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_j} = \left.\frac{\partial \mathcal{L}\Big(p\Big(\mathbf{X}\Big|\mathbf{Y},\widehat{\boldsymbol{\theta}}\Big)_j,\boldsymbol{\theta}\Big)}{\partial \theta_i}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_j} \tag{89}$$

4.4 Applying EM

Let us then look at how to apply EM to a SSM of the form (5). First of all, from the factorization in (3), the complete-data log-likelihood becomes

$$\begin{aligned} p(\mathbf{X},\mathbf{Y}|\boldsymbol{\theta}) = &-\frac{1}{2}\left(\mathbf{x}_0-\boldsymbol{\mu}_0\right)^T\boldsymbol{\Sigma}_0^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_0\right)-\frac{1}{2}\log|\boldsymbol{\Sigma}_0| \\ &-\frac{1}{2}\sum_{k=1}^{T}\big(\mathbf{x}_k-\mathbf{f}(\mathbf{x}_{k-1})\big)^T\mathbf{Q}^{-1}\big(\mathbf{x}_k-\mathbf{f}(\mathbf{x}_{k-1})\big)-\frac{T}{2}\log|\mathbf{Q}| \\ &-\frac{1}{2}\sum_{k=1}^{T}\big(\mathbf{y}_k-\mathbf{h}(\mathbf{x}_k)\big)^T\mathbf{R}^{-1}\big(\mathbf{y}_k-\mathbf{h}(\mathbf{x}_k)\big)-\frac{T}{2}\log|\mathbf{R}| \\ &+\texttt{const} \end{aligned}$$

Taking the expectation w.r.t. $p\big(\mathbf{X}\big|\mathbf{Y},\boldsymbol{\theta}'\big)$ (assumed implicitly in the notation), applying the identity $\mathbf{x}^T\mathbf{X}\mathbf{x} = \mathrm{Tr}\big[\mathbf{x}^T\mathbf{X}\mathbf{x}\big] = \mathrm{Tr}\big[\mathbf{X}\mathbf{x}\mathbf{x}^T\big]$ and using $\mathbf{f}_{k-1}\equiv\mathbf{f}(\mathbf{x}_{k-1})$ and $\mathbf{h}_k\equiv\mathbf{h}(\mathbf{x}_k)$

$$\begin{aligned} \mathfrak{L}\big(\boldsymbol{\theta}',\boldsymbol{\theta}\big) = &-\frac{1}{2}\mathrm{Tr}\Big[\boldsymbol{\Sigma}_0^{-1}\Big\langle\left(\mathbf{x}_0-\boldsymbol{\mu}_0\right)\left(\mathbf{x}_0-\boldsymbol{\mu}_0\right)^T\Big\rangle\Big]-\frac{1}{2}\log|\boldsymbol{\Sigma}_0| \\ &-\frac{1}{2}\sum_{k=1}^{T}\mathrm{Tr}\Big[\mathbf{Q}^{-1}\Big\langle\left(\mathbf{x}_k-\mathbf{f}_{k-1}\right)\left(\mathbf{x}_k-\mathbf{f}_{k-1}\right)^T\Big\rangle\Big]-\frac{T}{2}\log|\mathbf{Q}| \\ &-\frac{1}{2}\sum_{k=1}^{T}\mathrm{Tr}\Big[\mathbf{R}^{-1}\Big\langle\left(\mathbf{y}_k-\mathbf{h}_k\right)\left(\mathbf{y}_k-\mathbf{h}_k\right)^T\Big\rangle\Big]-\frac{T}{2}\log|\mathbf{R}| \\ &+\texttt{const}. \end{aligned} \tag{90}$$

Let us denote the three expectations in equation (90) with

$$\mathbf{I}_1 = \left\langle (\mathbf{x}_0 - \boldsymbol{\mu}_0)(\mathbf{x}_0 - \boldsymbol{\mu}_0)^T \right\rangle \tag{91}$$

$$= \int_{\mathcal{X}\times T} (\mathbf{x}_0 - \boldsymbol{\mu}_0)(\mathbf{x}_0 - \boldsymbol{\mu}_0)^T p\big(\mathbf{X}\,\big|\,\mathbf{Y}, \boldsymbol{\theta}'\big)\ \mathrm{d}\mathbf{X}$$

$$= \int_{\mathcal{X}} (\mathbf{x}_0 - \boldsymbol{\mu}_0)(\mathbf{x}_0 - \boldsymbol{\mu}_0)^T p\big(\mathbf{x}_0\,\big|\,\mathbf{Y}, \boldsymbol{\theta}'\big)\ \mathrm{d}\mathbf{x}_0 \tag{92}$$

$$\mathbf{I}_{2,k} = \int_{\mathcal{X}\times 2} (\mathbf{x}_k - \mathbf{f}_{k-1})(\mathbf{x}_k - \mathbf{f}_{k-1})^T p\left(\begin{bmatrix}\mathbf{x}_k & \mathbf{x}_{k-1}\end{bmatrix}^T \,\middle|\, \mathbf{Y}, \boldsymbol{\theta}'\right)\ \mathrm{d}\begin{bmatrix}\mathbf{x}_k & \mathbf{x}_{k-1}\end{bmatrix}^T \tag{93}$$

$$\mathbf{I}_{3,k} = \int_{\mathcal{X}} (\mathbf{y}_k - \mathbf{h}_k)(\mathbf{y}_k - \mathbf{h}_k)^T p\big(\mathbf{x}_k\,\big|\,\mathbf{Y}, \boldsymbol{\theta}'\big)\ \mathrm{d}\mathbf{x}_k \tag{94}$$

It is clear then that in the E-step one needs to compute the $T+1$ smoothing distributions, including the T cross-timestep distributions, since these will be needed in the expectations. By applying the identity

$$\mathrm{var}[\mathbf{x}] = \left\langle \mathbf{x}\mathbf{x}^T \right\rangle - \langle \mathbf{x} \rangle \langle \mathbf{x} \rangle^T, \tag{95}$$

we can already write the first expectation as

$$\mathbf{I}_1 = \mathbf{P}_{0|T} + (\mathbf{m}_{0|T} - \boldsymbol{\mu}_0)(\mathbf{m}_{0|T} - \boldsymbol{\mu}_0)^T. \tag{96}$$

Thus in the following more specific cases, we will only consider the two remaining expectations and the M-step.

4.4.1 EM in linear-Gaussian SSM:s

(Shumway et al. 1982; Ghahramani 1996) Let us substitute $\mathbf{A}\mathbf{x}_{k-1}$ for $\mathbf{f}_{k-1}$ and $\mathbf{H}\mathbf{x}_k$ for $\mathbf{h}_k$. Let us also denote by

$$p(\mathbf{x}_k, \mathbf{x}_{k-1}\,|\,\mathbf{Y}, \boldsymbol{\theta}) = \mathrm{N}\left(\begin{bmatrix}\mathbf{x}_k \\ \mathbf{x}_{k-1}\end{bmatrix}\ ;\ \mathbf{m}_{k,k-1|T}, \mathbf{P}_{k,k-1|T}\right), \tag{97}$$

the joint smoothing distribution of $\mathbf{x}_k$ and $\mathbf{x}_{k-1}$. Then by applying the manipulation

$$\begin{aligned} &\left\langle (\mathbf{x}_k - \mathbf{A}\mathbf{x}_{k-1})(\mathbf{x}_k - \mathbf{A}\mathbf{x}_{k-1})^T \right\rangle \\ &= \begin{bmatrix}\mathbf{I} & -\mathbf{A}\end{bmatrix} \left\langle \begin{bmatrix}\mathbf{x}_k \\ \mathbf{x}_{k-1}\end{bmatrix} \begin{bmatrix}\mathbf{x}_k^T & \mathbf{x}_{k-1}^T\end{bmatrix} \right\rangle \begin{bmatrix}\mathbf{I} \\ -\mathbf{A}^T\end{bmatrix} \end{aligned} \tag{98}$$

we get

$$\mathbf{I}_{2,k} = \begin{bmatrix}\mathbf{I} & -\mathbf{A}\end{bmatrix} \left(\mathbf{P}_{k,k-1|T} + \mathbf{m}_{k,k-1|T}\mathbf{m}_{k,k-1|T}^T\right) \begin{bmatrix}\mathbf{I} \\ -\mathbf{A}^T\end{bmatrix} \tag{99}$$

$$\mathbf{I}_{3,k} = \mathbf{H}\mathbf{P}_{k|T}\mathbf{H}^T + \big(\mathbf{y}_k - \mathbf{H}\mathbf{m}_{k|T}\big)\big(\mathbf{y}_k - \mathbf{H}\mathbf{m}_{k|T}\big)^T \tag{100}$$

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.

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:

$$\begin{aligned}
-2\frac{\partial \mathfrak{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}))}{\partial \theta_i} =& \mathrm{Tr}\left[-\mathbf{Q}^{-1}\frac{\partial \mathbf{Q}}{\partial \theta_i}\mathbf{Q}^{-1}\left(\mathbf{B}_1 - \mathbf{A}\mathbf{B}_2^T - \mathbf{B}_2\mathbf{A}^T + \mathbf{A}\mathbf{B}_3\mathbf{A}^T\right)\right] \\
&+ \mathrm{Tr}\left[\mathbf{Q}^{-1}\left(-\frac{\partial \mathbf{A}}{\partial \theta_i}\mathbf{B}_2^T - \mathbf{B}_2\frac{\partial \mathbf{A}}{\partial \theta_i}^T + \frac{\partial \mathbf{A}}{\partial \theta_i}\mathbf{B}_3\mathbf{A}^T + \mathbf{A}\mathbf{B}_3\frac{\partial \mathbf{A}}{\partial \theta_i}^T\right)\right] \\
&+ \mathrm{Tr}\left[-\mathbf{R}^{-1}\frac{\partial \mathbf{R}}{\partial \theta_i}\mathbf{R}^{-1}\left(\mathbf{B}_4 - \mathbf{H}\mathbf{B}_5^T - \mathbf{B}_5\mathbf{H}^T + \mathbf{H}\mathbf{B}_1\mathbf{H}^T\right)\right] \\
&+ N\mathrm{Tr}\left[\mathbf{Q}^{-1}\frac{\partial \mathbf{Q}}{\partial \theta_i}\right] + N\mathrm{Tr}\left[\mathbf{R}^{-1}\frac{\partial \mathbf{R}}{\partial \theta_i}\right]
\end{aligned} \tag{101}$$

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

Suppose the function $\mathbf{f}$ is linear in the parameters and the dimension of the state $\mathbf{x}$ is d. Then in the most general case $\mathbf{f}(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^d$ is a linear combination of vector valued functions $\boldsymbol{\rho}_k(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^{d_k}$ and the parameters are matrices $\boldsymbol{\Phi}_k \in \mathbb{R}^{d \times d_k}$. More specifically, we have

$$\begin{aligned}
\mathbf{f}(\mathbf{x}) &= \boldsymbol{\Phi}\boldsymbol{\rho}_1(\mathbf{x}) + \cdots + \boldsymbol{\Phi}_m\boldsymbol{\rho}_m(\mathbf{x}) \\
&= \begin{bmatrix}\boldsymbol{\Phi}_1 & \dots & \boldsymbol{\Phi}_m\end{bmatrix}\begin{bmatrix}\boldsymbol{\rho}_1(\mathbf{x}) \\ \vdots \\ \boldsymbol{\rho}_m(\mathbf{x})\end{bmatrix} \\
&= \mathbf{A}\mathbf{g}(\mathbf{x}),
\end{aligned} \tag{102}$$

where $\mathbf{A} \in \mathbb{R}^{d \times \sum_{k=1}^m d_k}$ and $\mathbf{g}(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^{\sum_{k=1}^m d_k}$. For example, in case of the function

$$f(x, t) = ax + b\frac{x}{1 + x^2} + c\cos(1.2t) \tag{103}$$

we would have

$$f(x,t) = \begin{bmatrix} a & b & c \end{bmatrix} \begin{bmatrix} x \\ \frac{x}{1+x^2} \\ \cos(1.2t) \end{bmatrix} \tag{104}$$

Suppose now, that the matrix A depends on parameters $\mathbf{s}$. Let us then deduce the maximization equation for the parameter s_i:

$$-2\,\frac{\partial \mathfrak{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_j)}{\partial s_i} = \sum_{k=1}^{N} \left\langle \frac{\partial}{\partial s_i} \left(\mathbf{x}_k - \mathbf{A}\mathbf{g}(\mathbf{x}_{k-1})\right)^T \mathbf{Q}^{-1} \left(\mathbf{x}_k - \mathbf{A}\mathbf{g}(\mathbf{x}_{k-1})\right) \right\rangle_{\hat{\boldsymbol{\theta}}_j} \tag{105}$$

and

$$\frac{\partial}{\partial s_i} \left(\mathbf{x}_k - \mathbf{A}\mathbf{g}(\mathbf{x}_{k-1})\right)^T \mathbf{Q}^{-1} \left(\mathbf{x}_k - \mathbf{A}\mathbf{g}(\mathbf{x}_{k-1})\right) \tag{106}$$

$$= -\,2\mathbf{g}(\mathbf{x}_{k-1})^T \frac{\partial \mathbf{A}}{\partial s_i}^T \mathbf{Q}^{-1} \left(\mathbf{x}_k - \mathbf{A}\mathbf{g}(\mathbf{x}_{k-1})\right) \tag{107}$$

$$= -\,2\mathrm{Tr}\left[\frac{\partial \mathbf{A}}{\partial s_i}^T \mathbf{Q}^{-1} \mathbf{x}_k \mathbf{g}(\mathbf{x}_{k-1})^T\right] - 2\mathrm{Tr}\left[\frac{\partial \mathbf{A}}{\partial s_i}^T \mathbf{Q}^{-1} \mathbf{A}\mathbf{g}(\mathbf{x}_{k-1})\mathbf{g}(\mathbf{x}_{k-1})^T\right] \tag{108}$$

$$= -\,2\mathrm{Tr}\left[\frac{\partial \mathbf{A}}{\partial s_i}^T \mathbf{Q}^{-1} \left(\mathbf{x}_k \mathbf{g}(\mathbf{x}_{k-1})^T - \mathbf{A}\mathbf{g}(\mathbf{x}_{k-1})\mathbf{g}(\mathbf{x}_{k-1})^T\right)\right]. \tag{109}$$

Combining equations (109) and (113) then gives

$$\frac{\partial \mathfrak{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_j)}{\partial s_i} = \\ -\,2\mathrm{Tr}\left[\frac{\partial \mathbf{A}}{\partial s_i}^T \mathbf{Q}^{-1} \left(\sum_{k=1}^{N} \left\langle \mathbf{x}_k \mathbf{g}(\mathbf{x}_{k-1})^T \right\rangle - \mathbf{A} \sum_{k=1}^{N} \left\langle \mathbf{g}(\mathbf{x}_{k-1})\mathbf{g}(\mathbf{x}_{k-1})^T \right\rangle\right)\right] \tag{110}$$

If we then set (110) to zero, we have

$$\mathbf{A}_{j+1} = \left(\sum_{k=1}^{N} \left\langle \mathbf{x}_k \mathbf{g}(\mathbf{x}_{k-1})^T \right\rangle\right) \left(\sum_{k=1}^{N} \left\langle \mathbf{g}(\mathbf{x}_{k-1})\mathbf{g}(\mathbf{x}_{k-1})^T \right\rangle\right)^{-1} \tag{111}$$

which is similar to (**??**)

Suppose the function $\mathbf{h}$ is linear in the parameters and the dimension of the state $\mathbf{x}$ is d^x and of the measurements d^y. Then in the most general case $\mathbf{h}(\mathbf{x}) : \mathbb{R}^{d^x} \to \mathbb{R}^{d^y}$ is a linear combination of vector valued functions $\boldsymbol{\pi}_k(\mathbf{x}) : \mathbb{R}^{d^x} \to \mathbb{R}^{d_k}$ and the parameters are matrices $\boldsymbol{\Upsilon}_k \in \mathbb{R}^{d^x \times d_k}$. More specifically, we have

$$\begin{aligned}\mathbf{h}(\mathbf{x}) &= \boldsymbol{\Upsilon}\boldsymbol{\pi}_1(\mathbf{x}) + \cdots + \boldsymbol{\Upsilon}_m\boldsymbol{\pi}_m(\mathbf{x}) \\ &= \begin{bmatrix}\boldsymbol{\Upsilon}_1 & \dots & \boldsymbol{\Upsilon}_m\end{bmatrix} \begin{bmatrix}\boldsymbol{\pi}_1(\mathbf{x}) \\ \vdots \\ \boldsymbol{\pi}_m(\mathbf{x})\end{bmatrix} \\ &= \mathbf{H}\mathbf{l}(\mathbf{x}),\end{aligned} \tag{112}$$

Suppose now, that the matrix $\mathbf{H}$ depends on parameters $\mathbf{t}$. Let us then deduce the maximization equation for the parameter t_i:

$$-2\,\frac{\partial \mathfrak{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_j)}{\partial t_i} = \frac{1}{2}\mathrm{Tr}\left[\mathbf{R}^{-1}\sum_{k=1}^{N}\left\langle \frac{\partial}{\partial t_i}\left(\mathbf{y}_k - \mathbf{H}\mathbf{l}(\mathbf{x}_k)\right)\left(\mathbf{y}_k - \mathbf{H}\mathbf{l}(\mathbf{x}_k)\right)^T\right\rangle_{\hat{\boldsymbol{\theta}}_j}\right] \tag{113}$$

and

If we then set (110) to zero, we have

$$\mathbf{H}_{j+1} = \left(\sum_{k=1}^{N}\mathbf{y}_k\left\langle \mathbf{l}(\mathbf{x}_k)^T\right\rangle\right)\left(\sum_{k=1}^{N}\left\langle \mathbf{l}(\mathbf{x}_k)\mathbf{l}(\mathbf{x}_k)^T\right\rangle\right)^{-1} \tag{114}$$

which is similar to (**??**)

4.4.3 EM in nonlinear-Gaussian SSM:s

As explained in section 3.2, in then nonlinear case the filtering and smoothing distributions cannot be computed exactly. Thus the E-step solution is also only approximate and the convergence guarantees of EM won't apply anymore. The proposed methods can nevertheless provide good results most of the time. In the fortunate case that the model is linear-in-the-parameters the M-step can be solved in closed form. This situation will be covered later in section 4.4.2.

$$\frac{\partial \mathfrak{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}))}{\partial \theta_i} = \frac{\partial I_1}{\partial \theta_i} + \frac{\partial I_3}{\partial \theta_i} + \frac{\partial I_3}{\partial \theta_i} \tag{115}$$

and

$$
\begin{aligned}
-2\,\frac{\partial I_2}{\partial \theta_i} =& -\operatorname{Tr}\left[\mathbf{Q}^{-1}\,\frac{\partial \mathbf{Q}}{\partial \theta_i}\,\mathbf{Q}^{-1}\sum_{k=1}^{N}\left\langle \left(\mathbf{x}_k-\mathbf{f}(\mathbf{x}_{k-1})\right)\left(\mathbf{x}_k-\mathbf{f}(\mathbf{x}_{k-1})\right)^T\right\rangle_{\hat{\boldsymbol{\theta}}_j}\right]\\
&-\operatorname{Tr}\left[\mathbf{Q}^{-1}\sum_{k=1}^{N}\left\langle \frac{\partial \mathbf{f}(\mathbf{x}_{k-1})}{\partial \theta_i}\left(\mathbf{x}_k-\mathbf{f}(\mathbf{x}_{k-1})\right)^T+\left(\mathbf{x}_k-\mathbf{f}(\mathbf{x}_{k-1})\right)\frac{\partial \mathbf{f}^T(\mathbf{x}_{k-1})}{\partial \theta_i}\right\rangle_{\hat{\boldsymbol{\theta}}_j}\right]\\
&+N\operatorname{Tr}\left[\mathbf{Q}^{-1}\,\frac{\partial \mathbf{Q}}{\partial \theta_i}\right]\\
\frac{\partial I_3}{\partial \theta_i} =& -\operatorname{Tr}\left[\mathbf{R}^{-1}\,\frac{\partial \mathbf{R}}{\partial \theta_i}\,\mathbf{R}^{-1}\sum_{k=1}^{N}\left\langle \left(\mathbf{y}_k-\mathbf{h}(\mathbf{x}_k)\right)\left(\mathbf{y}_k-\mathbf{h}(\mathbf{x}_k)\right)^T\right\rangle_{\hat{\boldsymbol{\theta}}_j}\right]\\
&-\operatorname{Tr}\left[\mathbf{R}^{-1}\sum_{k=1}^{N}\left\langle \frac{\partial \mathbf{h}(\mathbf{x}_k)}{\partial \theta_i}\left(\mathbf{y}_k-\mathbf{h}(\mathbf{x}_k)\right)^T+\left(\mathbf{y}_k-\mathbf{h}(\mathbf{x}_k)\right)\frac{\partial \mathbf{h}^T(\mathbf{x}_k)}{\partial \theta_i}\right\rangle_{\hat{\boldsymbol{\theta}}_j}\right]\\
&+N\operatorname{Tr}\left[\mathbf{R}^{-1}\,\frac{\partial \mathbf{R}}{\partial \theta_i}\right]
\end{aligned}
\tag{116}
$$

Let us separate the set of all parameters $\boldsymbol{\theta}$ into subsets

$$
\boldsymbol{\theta}=\{\boldsymbol{\psi},\boldsymbol{\omega},\mathbf{Q},\mathbf{R}\}\,, \tag{117}
$$

where $\boldsymbol{\psi}$ are the parameters of f and $\boldsymbol{\omega}$ are the parameters of h (these sets can intersect). Let us first consider the maximization with respect to $\mathbf{Q}$. We have

$$
\begin{aligned}
-\,2\,\frac{\partial \mathfrak{L}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}_j)}{\partial \mathbf{Q}^{-\mathbf{1}}} =& \sum_{k=1}^{N}\frac{\partial}{\partial \mathbf{Q}^{-\mathbf{1}}}\operatorname{Tr}\left[\mathbf{Q}^{-1}\left\langle \left(\mathbf{x}_k-\mathbf{f}(\mathbf{x}_{k-1})\right)\left(\mathbf{x}_k-\mathbf{f}(\mathbf{x}_{k-1})\right)^T\right\rangle_{\hat{\boldsymbol{\theta}}_j}\right]\\
&+N\,\frac{\partial}{\partial \mathbf{Q}^{-\mathbf{1}}}\log|\mathbf{Q}|
\end{aligned}
\tag{118}
$$

Using formula 92 in **Petersen2008** for the first derivative and formula 51 for the second, we get

$$
-2\,\frac{\partial \mathfrak{L}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}_j)}{\partial \mathbf{Q}^{-\mathbf{1}}} = \sum_{k=1}^{N}\left\langle \left(\mathbf{x}_k-\mathbf{f}(\mathbf{x}_{k-1})\right)\left(\mathbf{x}_k-\mathbf{f}(\mathbf{x}_{k-1})\right)^T\right\rangle_{\hat{\boldsymbol{\theta}}_j}-N\mathbf{Q} \tag{119}
$$

and setting this to zero we get the update equation for the next estimate of $\mathbf{Q}$

$$
\mathbf{Q}_{j+1}=\frac{1}{N}\sum_{k=1}^{N}\left\langle \left(\mathbf{x}_k-\mathbf{f}(\mathbf{x}_{k-1})\right)\left(\mathbf{x}_k-\mathbf{f}(\mathbf{x}_{k-1})\right)^T\right\rangle_{\hat{\boldsymbol{\theta}}_j} \tag{120}
$$

The derivation of the update equation for the next estimate of $\mathbf{R}$ is exactly analogous, giving

$$\mathbf{R}_{j+1} = \frac{1}{N} \sum_{k=1}^{N} \left\langle \left(\mathbf{y}_k - \mathbf{h}(\mathbf{x}_k)\right) \left(\mathbf{y}_k - \mathbf{h}(\mathbf{x}_k)\right)^T \right\rangle_{\hat{\boldsymbol{\theta}}_j} \tag{121}$$

4.5 Comparisons

4.5.1 Convergence

4.5.2 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 Stability

(Haykin 2001)

6.2 Dual and joint filtering

(Haykin 2001)

6.3 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.

Gupta, N. and R. Mehra (Dec. 1974). "Computational aspects of maximum likelihood estimation and reduction in sensitivity function calculations". In: *IEEE Transactions on Automatic Control* 19.6, pp. 774–783. ISSN: 0018-9286. DOI: 10.1109/TAC.1974.1100714.

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.

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: *Proceedings of the Twentieth International Conference on Machine Learning*. Washington DC.

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.