Ville Väänänen

Sigma point smoother based expectation maximization in discrete-time state-space models

School of Electrical Engineering

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

Espoo 14.3.2011

Thesis supervisor:

Prof. Jouko Lampinen

Thesis instructor:

D.Sc. (Tech.) Simo Särkkä

Aalto University
School of Electrical
Engineering

Contents

1 Introduction

- 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 (Cappé et al. 2005; Durbin et al. 2012; 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 this thesis we will denote the indexing variable as t and it is assumed to be discrete and nonnegative. We will use the shorthand notation $\mathbf{x}_k$ for the the value of the stochastic process at time t_k, i.e. of $X(t_k)$. Also the shorthand $\mathbf{x}_{1:k}$ will be used to mean $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$.

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 value of the process at the current timestep depends only on the previous one. A further assumption in SSMs is that the process is not directly observed but only through another stochastic process, the *measurement process*. The random variables 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 directed acyclic graph (DAG) (**Bishop2006a**; Pearl 1988). A Bayes network presentation of a SSM is given in figure 1.

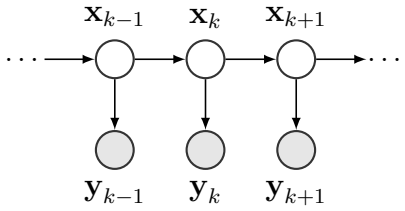

Figure 1: SSM as a graphical model presented with a directed acyclic graph

Taking into account the Markov property of the state process and the conditional indepence property measurement process, the joint distribution of states and measurements factorises as

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

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

Prior distribution $p\left(\mathbf{x}_0\right)$ 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).

Dynamics model $p(\mathbf{x}_k \mid \mathbf{x}_{k-1})$ dictates the time evolution of the states

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

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

$$p(x_k \mid x_{k-1}) = x_{k-1} + q_{k-1}, \tag{2}$$

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

$$p(y_k \mid x_k) = x_k + r_k, \tag{3}$$

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

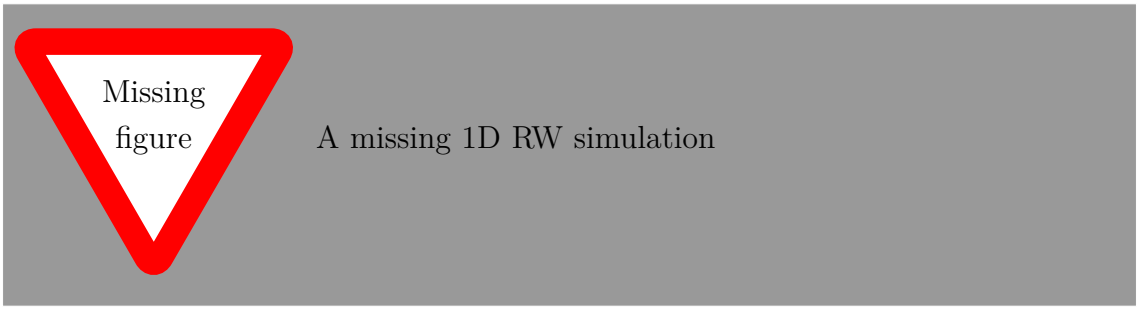

Figure 2: Simulation from the 1D RW model

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

This whole chapter needs to be refactored

stationarity

explain hidden variables

$$\mathbf{x}_k = \mathbf{f}(\mathbf{x}_{k-1}) + \mathbf{q}_{k-1} \tag{4}$$
$$\mathbf{q}_{k-1} \sim \mathrm{N}(0, \mathbf{Q}), \tag{5}$$

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

$$p(\mathbf{x}_k \mid \mathbf{x}_{1:k-1}) = p(\mathbf{x}_k \mid \mathbf{x}_{k-1}) \tag{6}$$

The observations depend on the state through the measurement equation

$$\mathbf{y}_k = \mathbf{h}(\mathbf{x_k}) + \mathbf{r}_k \tag{7}$$
$$\mathbf{r}_k \sim \mathrm{N}(0, \mathbf{R}), \tag{8}$$

where $h : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$. An equivalent way of presenting equations (26a), (26c), (26b) and (26d) is

$$p(\mathbf{x}_k \mid \mathbf{x}_{k-1}) = \mathrm{N}\big(\mathbf{f}(\mathbf{x}_{k-1}), \mathbf{Q}\big) \tag{9}$$
$$p(\mathbf{y}_k \mid \mathbf{x}_k) = \mathrm{N}\big(\mathbf{h}(\mathbf{x}_k), \mathbf{R}\big). \tag{10}$$

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

Smoothing distribution $p(\mathbf{x}_k \mid \mathbf{y}_{1:T})$ 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 $p\left(\mathbf{x}_{k-1} \,\middle|\, \mathbf{y}_{1:k-1}\right)$ be the filtering distribution of the previous step. Then

$$\begin{aligned} p\left(\mathbf{x}_k \,\middle|\, \mathbf{y}_{1:k-1}\right) &= \int p\left(\mathbf{x}_k, \mathbf{x}_{k-1} \,\middle|\, \mathbf{y}_{1:k-1}\right) \, \mathrm{d}\mathbf{x}_{k-1} \\ &= \int p(\mathbf{x}_k \mid \mathbf{x}_{k-1})\, p\left(\mathbf{x}_{k-1} \,\middle|\, \mathbf{y}_{1:k-1}\right) \, \mathrm{d}\mathbf{x}_{k-1}, \end{aligned} \tag{11}$$

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)

$$\begin{aligned}\underbrace{p\left(\mathbf{x}_k \mid \mathbf{y}_{1:k}\right)}_{\text{posterior}} &= \frac{\overbrace{p\left(\mathbf{y}_k \mid \mathbf{x}_k\right)}^{\text{likelihood}} \overbrace{p\left(\mathbf{x}_k \mid \mathbf{y}_{1:k-1}\right)}^{\text{prior}}}{\underbrace{p\left(\mathbf{y}_k \mid \mathbf{y}_{1:k-1}\right)}_{\text{marginal likelihood}}} \\ &= \frac{p\left(\mathbf{y}_k \mid \mathbf{x}_k\right) p\left(\mathbf{x}_k \mid \mathbf{y}_{1:k-1}\right)}{\int p\left(\mathbf{y}_k \mid \mathbf{x}_k\right) p\left(\mathbf{x}_k \mid \mathbf{y}_{1:k-1}\right) \, \mathrm{d}\mathbf{x}_k} \end{aligned} \tag{12}$$

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 $p\left(\mathbf{x}_{k+1} \mid \mathbf{y}_{1:T}\right)$ of the "previous" step are available. Since

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

we get

$$p\left(\mathbf{x}_k \mid \mathbf{y}_{1:T}\right) = p\left(\mathbf{x}_k \mid \mathbf{y}_{1:k}\right) \int \left[\frac{p\left(\mathbf{x}_{k+1} \mid \mathbf{x}_k\right) p\left(\mathbf{x}_{k+1} \mid \mathbf{y}_{1:T}\right)}{p\left(\mathbf{x}_{k+1} \mid \mathbf{y}_{1:k}\right)} \right] \, \mathrm{d}\mathbf{x}_{k+1}, \tag{13}$$

where $p\left(\mathbf{x}_{k+1} \mid \mathbf{y}_{1:k}\right)$ can be computed by equation (11).

Marginal likelihood

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

$$p\left(\mathbf{y}_k \mid \mathbf{y}_{1:k-1}\right) = \int p\left(\mathbf{y}_k \mid \mathbf{x}_k\right) p\left(\mathbf{x}_k \mid \mathbf{y}_{1:k-1}\right) \, \mathrm{d}\mathbf{x}_k \tag{14}$$

the marginal likelihood can be computed from

$$p\left(\mathbf{y}_{1:T}\right) = p\left(\mathbf{y}_1\right) \prod_{k=2}^{T} p\left(\mathbf{y}_k \mid \mathbf{y}_{1:k-1}\right) \tag{15}$$

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

$$p\left(\mathbf{x}_{0:T}, \mathbf{y}_{1:T}, \boldsymbol{\theta}\right) = p\left(\boldsymbol{\theta}\right) p\left(\mathbf{x}_0 \mid \boldsymbol{\theta}\right) \prod_{k=1}^{T} p\left(\mathbf{x}_k \mid \mathbf{x}_{k-1}, \boldsymbol{\theta}\right) p\left(\mathbf{y}_k \mid \mathbf{x}_k, \boldsymbol{\theta}\right). \tag{16}$$

3 State estimation

3.1 Linear-Gaussian State Space Models

Linear-Gaussian SSMs can be defined with the following equations

$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} + \mathbf{q}_{k-1} \tag{17a}$$
$$\mathbf{y}_k = \mathbf{H}\mathbf{x}_{\mathbf{k}} + \mathbf{r}_k \tag{17b}$$
$$\mathbf{q}_{k-1} \sim \mathrm{N}(0, \mathbf{Q}) \tag{17c}$$
$$\mathbf{r}_k \sim \mathrm{N}(0, \mathbf{R}) \tag{17d}$$
$$\mathbf{x}_0 \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{17e}$$

Equations (17a), (17b), (17c) and (17d) together specify the following conditional distributions

$$\mathbf{x}_k | \mathbf{x}_{k-1} \sim \mathrm{N}(\mathbf{A}\mathbf{x}_{k-1}, \mathbf{Q}) \tag{18}$$
$$\mathbf{y}_k | \mathbf{x}_k \sim \mathrm{N}(\mathbf{H}\mathbf{x}_{\mathbf{k}}, \mathbf{R}) \tag{19}$$

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 (Mbalawata et al. 2011):

prediction:

$$\mathbf{m}_{k|k-1} = \mathbf{A}\mathbf{m}_{k-1|k-1} \tag{20a}$$
$$\mathbf{P}_{k|k-1} = \mathbf{A}\mathbf{P}_{k-1|k-1}\mathbf{A}^T + \mathbf{Q} \tag{20b}$$

update:

$$\mathbf{v}_k = \mathbf{y}_k - \mathbf{H}\mathbf{m}_{k|k-1} \tag{20c}$$
$$\mathbf{S}_k = \mathbf{H}\mathbf{P}_{k|k-1}\mathbf{H}^T + \mathbf{R} \tag{20d}$$
$$\mathbf{K}_k = \mathbf{P}_{k|k-1}\mathbf{H}^T\mathbf{S}_k^{-1} \tag{20e}$$
$$\mathbf{m}_{k|k} = \mathbf{m}_{k|k-1} + \mathbf{K}_k\mathbf{v}_k \tag{20f}$$
$$\mathbf{P}_{k|k} = \mathbf{P}_{k|k-1} - \mathbf{K}_k\mathbf{S}_k\mathbf{K}_k^T \tag{20g}$$

This includes the sufficient statistics for the T joint distributions

$$\begin{aligned} p\left(\mathbf{x}_k, \mathbf{y}_k \mid \mathbf{y}_1, \ldots, \mathbf{y}_{k-1}, \boldsymbol{\theta}\right) &= N\left(\begin{bmatrix}\mathbf{x}_k \\ \mathbf{y}_k\end{bmatrix} \middle| \begin{bmatrix}\mathbf{m}_{k|k-1} \\ \mathbf{H}\mathbf{m}_{k|k-1}\end{bmatrix}, \begin{bmatrix}\mathbf{P}_{k|k-1} & \mathbf{P}_{k|k-1}\mathbf{H}^T \\ \mathbf{H}\mathbf{P}_{k|k-1}^T & \mathbf{S}_k\end{bmatrix}\right) \\ \Rightarrow p\left(\mathbf{y}_k \mid \mathbf{y}_1, \ldots, \mathbf{y}_{k-1}, \boldsymbol{\theta}\right) &= N\left(\mathbf{y}_k \mid \mathbf{H}\mathbf{m}_{k|k-1}, \mathbf{S}_k\right) \end{aligned} \tag{21}$$

3.1.2 RTS Smoother

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

$$\langle \mathbf{x}_k \rangle = \mathbf{m}_{k|N} \tag{22}$$
$$\left\langle \mathbf{x}_k \mathbf{x}_{k|k-1} T \right\rangle = \mathbf{P}_{k|N} + \mathbf{m}_{k|N}(\mathbf{m}_{k|N})^T \tag{23}$$
$$\left\langle \mathbf{x}_k \mathbf{x}_{k-1}^T \right\rangle = \mathbf{C}_{k|N} + \mathbf{m}_{k|N}(\mathbf{m}_{k-1|N})^T, \tag{24}$$

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 $\mathbf{y}_1, \ldots, \mathbf{y}_N$. The standard RTS smoother gives the statistics $\mathbf{m}_{k|N}$ and $\mathbf{P}_{k|N}$. The cross-timestep variance $\mathbf{C}_{k|N}$ can be computed with an additional recursive formula alongside the usual RTS smoother recursions

$$\mathbf{J}_k = \mathbf{P}_{k|k}\mathbf{A}^T\mathbf{P}_{k|k+1}^{-1} \tag{25a}$$
$$\mathbf{m}_{k|N} = \mathbf{m}_{k|k} + \mathbf{J}_k\left(\mathbf{m}_{k+1|N} - \mathbf{m}_{k+1|k}\right) \tag{25b}$$
$$\mathbf{P}_{k|N} = \mathbf{P}_{k|k} + \mathbf{J}_k\left(\mathbf{P}_{k+1|N} - \mathbf{P}_{k+1|k}\right)\mathbf{J}_k^T \tag{25c}$$
$$\mathbf{C}_{k|N} = \mathbf{P}_{k|k}\mathbf{J}_{k-1}^T + \mathbf{J}_k\left(\mathbf{C}_{k+1|N} - \mathbf{A}\mathbf{P}_{k|k}\right)\mathbf{J}_{k-1}^T \tag{25d}$$
$$\tag{25e}$$

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

$$\mathbf{x}_k = \mathbf{f}(\mathbf{x}_{k-1}) + \mathbf{q}_{k-1} \tag{26a}$$
$$\mathbf{y}_k = \mathbf{h}(\mathbf{x_k}) + \mathbf{r}_k \tag{26b}$$
$$\mathbf{q}_{k-1} \sim \mathrm{N}(0, \mathbf{Q}) \tag{26c}$$
$$\mathbf{r}_k \sim \mathrm{N}(0, \mathbf{R}) \tag{26d}$$
$$\mathbf{x}_0 \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{26e}$$

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

3.2.1 Assumed density filtering

One approach to forming Gaussian approximations is to assume a Gaussian probability density function with mean and variance that match the actual ones. Let

$$\mathbf{a} \sim \mathrm{N}(\mathbf{m}, \boldsymbol{\Sigma}_a) \tag{27}$$
$$\mathbf{b}|\mathbf{a} \sim \mathrm{N}\big(\mathbf{f}(\mathbf{a}), \boldsymbol{\Sigma}_{b|a}\big) \tag{28}$$

then

$$p\left(\mathbf{a}, \mathbf{b}\right) = \mathrm{N}(\mathbf{m}, \boldsymbol{\Sigma}_a)\, \mathrm{N}\big(\mathbf{f}(\mathbf{a}), \boldsymbol{\Sigma}_{b|a}\big) \tag{29}$$

is only Gaussian if $\mathbf{f}(\mathbf{a})$ is linear. Let the Gaussian approximation to (29) be

$$p\left(\begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix}\right) \approx \mathrm{N}\left(\begin{bmatrix}\boldsymbol{\mu}_a\\ \boldsymbol{\mu}_b\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab}\\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb}\end{bmatrix}\right) \tag{30}$$

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

$$\boldsymbol{\mu}_a = \mathbf{m} \tag{31}$$
$$\boldsymbol{\Sigma}_{aa} = \boldsymbol{\Sigma}_a \tag{32}$$
$$\boldsymbol{\mu}_b = \int \mathbf{b} p\left(\mathbf{b}\right)\, \mathrm{d}\mathbf{b} \tag{33}$$
$$\boldsymbol{\Sigma}_{bb} = \int (\mathbf{b} - \boldsymbol{\mu}_b)(\mathbf{b} - \boldsymbol{\mu}_b)^T p\left(\mathbf{b}\right)\, \mathrm{d}\mathbf{b} \tag{34}$$

Both (33) and (34) can be written in terms of (27) and (28). To see this, let us rewrite (33) as

$$\begin{aligned}\boldsymbol{\mu}_b &= \int \mathbf{b} p\left(\mathbf{b}\right)\ \mathrm{d}\mathbf{b} \\ &= \int \mathbf{b} \int p\left(\mathbf{b} \mid \mathbf{a}\right) p\left(\mathbf{a}\right)\ \mathrm{d}\mathbf{a}\ \mathrm{d}\mathbf{b} \\ &= \int \int \mathbf{b} p\left(\mathbf{b} \mid \mathbf{a}\right)\ \mathrm{d}\mathbf{b} p\left(\mathbf{a}\right)\ \mathrm{d}\mathbf{a} \\ &= \int \mathbf{f}(\mathbf{a}) \mathrm{N}(\mathbf{m}, \boldsymbol{\Sigma}_a)\ \mathrm{d}\mathbf{a} \end{aligned} \tag{35}$$

and (34) as

$$\begin{aligned}\boldsymbol{\Sigma}_{bb} &= \int \mathbf{b}\mathbf{b}^T p\left(\mathbf{b}\right)\ \mathrm{d}\mathbf{b} - \boldsymbol{\mu}_b \boldsymbol{\mu}_b^T \\ &= \int \mathbf{f}(\mathbf{a})\mathbf{f}(\mathbf{a})^T p\left(\mathbf{a}\right)\ \mathrm{d}\mathbf{a} - \boldsymbol{\mu}_b \boldsymbol{\mu}_b^T \\ &\quad + \int \int [(\mathbf{b} - \mathbf{f}(\mathbf{a}))(\mathbf{b} - \mathbf{f}(\mathbf{a}))^T] p\left(\mathbf{b} \mid \mathbf{a}\right)\ \mathrm{d}\mathbf{b} p\left(\mathbf{a}\right)\ \mathrm{d}\mathbf{a} \\ &= \int (\mathbf{f}(\mathbf{a}) - \boldsymbol{\mu}_b)(\mathbf{f}(\mathbf{a}) - \boldsymbol{\mu}_b)^T \mathrm{N}(\mathbf{m}, \boldsymbol{\Sigma}_a)\ \mathrm{d}\mathbf{a} + \boldsymbol{\Sigma}_{b|a}. \end{aligned} \tag{36}$$

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

$$\begin{aligned}\boldsymbol{\Sigma}_{ab} &= \int \int (\mathbf{a} - \boldsymbol{\mu}_a)(\mathbf{b} - \boldsymbol{\mu}_b)^T p\left(\mathbf{a}, \mathbf{b}\right)\ \mathrm{d}\mathbf{a}\ \mathrm{d}\mathbf{b} \\ &= \int \int (\mathbf{a} - \boldsymbol{\mu}_a)(\mathbf{b} - \boldsymbol{\mu}_b)^T p\left(\mathbf{a}\right) p\left(\mathbf{b} \mid \mathbf{a}\right)\ \mathrm{d}\mathbf{a}\ \mathrm{d}\mathbf{b} \\ &= \int (\mathbf{a} - \boldsymbol{\mu}_a)(\int \mathbf{b} p\left(\mathbf{b} \mid \mathbf{a}\right)\ \mathrm{d}\mathbf{b} - \boldsymbol{\mu}_b)^T p\left(\mathbf{a}\right)\ \mathrm{d}\mathbf{a} \\ &= \int (\mathbf{a} - \mathbf{m})(\mathbf{f}(\mathbf{a}) - \boldsymbol{\mu}_b)^T \mathrm{N}(\mathbf{m}, \boldsymbol{\Sigma}_a)\ \mathrm{d}\mathbf{a} \end{aligned} \tag{37}$$

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

$$\begin{aligned} p\left(\mathbf{x}_{k-1}, \mathbf{x}_k \mid \mathbf{Y}\right) &= p\left(\mathbf{x}_{k-1} \mid \mathbf{x}_k, \mathbf{Y}_{1:k-1}\right) p\left(\mathbf{x}_k \mid \mathbf{Y}\right) \\ &= \frac{p\left(\mathbf{x}_{k-1}, \mathbf{x}_k \mid \mathbf{Y}_{1:k-1}\right) p\left(\mathbf{x}_k \mid \mathbf{Y}\right)}{p\left(\mathbf{x}_k \mid \mathbf{Y}_{1:k-1}\right)}, \end{aligned} \tag{38}$$

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

$$p\left(\mathbf{x}_{k-1} \;\middle|\; \mathbf{Y}_{1:k-1}, \hat{\boldsymbol{\theta}}_j\right) \approx \mathrm{N}\big(\mathbf{m}_{k-1|k-1}, \mathbf{P}_{k-1|k-1}\big)\,. \tag{39}$$

The Gaussian approximation to

$$p\left(\mathbf{x}_{k-1}, \mathbf{x}_k \mid \mathbf{Y}_{1:k-1}\right) = \mathrm{N}\big(\mathbf{x}_k \;\big|\; \mathbf{f}(\mathbf{x}_{k-1}), \mathbf{Q}\big)\, p\left(\mathbf{x}_{k-1} \mid \mathbf{Y}_{1:k-1}\right) \tag{40}$$

is then given by application of equations (35), (36) and (37)

$$\mathbf{m}_{k|k-1} = \int \mathbf{f}(\mathbf{x}_{k-1}) \mathrm{N}\big(\mathbf{x}_{k-1} \;\big|\; \mathbf{m}_{k-1|k-1}, \mathbf{P}_{k-1|k-1}\big)\; \mathrm{d}\mathbf{x}_{k-1} \tag{41}$$

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

$$\begin{aligned} \mathbf{C}_{k^-} = \int & (\mathbf{x}_{k-1} - \mathbf{m}_{k-1|k-1})(\mathbf{f}(\mathbf{x}_{k-1}) - \mathbf{m}_{k|k-1})^T \\ & \mathrm{N}\big(\mathbf{x}_{k-1} \;\big|\; \mathbf{m}_{k-1|N}, \mathbf{P}_{k-1|N}\big)\; \mathrm{d}\mathbf{x}_{k-1} \end{aligned} \tag{43}$$

so that the approximation is

$$p\left(\mathbf{x}_{k-1}, \mathbf{x}_k \;\middle|\; \mathbf{Y}_{1:k-1}, \hat{\boldsymbol{\theta}}_j\right) \approx \mathrm{N}\left(\begin{bmatrix} \mathbf{m}_{k-1|k-1} \\ \mathbf{m}_{k|k-1} \end{bmatrix}, \begin{bmatrix} \mathbf{P}_{k-1|k-1} & \mathbf{C}_{k^-} \\ \mathbf{C}_{k^-}^T & \mathbf{P}_{k|k-1} \end{bmatrix}\right) \tag{44}$$

In order to calculate (41) and (42) we also need a Gaussian approximation for the joint distribution of the current state and measurement given the previous measurements

$$\begin{aligned} p\left(\mathbf{x}_k, \mathbf{y}_k \mid \mathbf{Y}_{1:k-1}\right) &= \mathrm{N}\big(\mathbf{y}_k \;\big|\; \mathbf{h}(\mathbf{x}_k), \mathbf{R}\big)\, p\left(\mathbf{x}_k \mid \mathbf{Y}_{1:k-1}\right) \\ &\approx \mathrm{N}\left(\begin{bmatrix} \mathbf{m}_{k|k-1} \\ \boldsymbol{\mu}_k \end{bmatrix}, \begin{bmatrix} \mathbf{P}_{k|k-1} & \mathbf{C}_k \\ \mathbf{C}_k^T & \mathbf{S}_k \end{bmatrix}\right). \end{aligned} \tag{45}$$

Applying equations (35), (36) and (37) again, we get

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

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

$$\mathbf{C}_k = \int (\mathbf{x}_k - \mathbf{m}_{k|k-1})(\mathbf{h}(\mathbf{x}_k) - \boldsymbol{\mu}_k)^T \mathrm{N}\big(\mathbf{x}_k \;\big|\; \mathbf{m}_{k|k-1}, \mathbf{P}_{k|k-1}\big)\; \mathrm{d}\mathbf{x}_k \tag{48}$$

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

$$\mathbf{m}_{k|k} = \mathbf{m}_{k|k-1} + \mathbf{C}_k \mathbf{S}_k^{-1} \left(\mathbf{y}_k - \boldsymbol{\mu}_k\right) \tag{49}$$

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

Again using the formula for the conditional of jointly Gaussian variables we get from (44)

$$p\left(\mathbf{x}_{k-1} \mid \mathbf{x}_k, \mathbf{Y}_{1:k-1}\right) \approx \mathrm{N}(\mathbf{m}_2, \mathbf{P}_2) \tag{51}$$
$$\mathbf{G}_{k-1} = \mathbf{C}_k \mathbf{P}_{k|k-1}^{-1} \tag{52}$$
$$\mathbf{m}_2 = \mathbf{m}_{k-1|k-1} + \mathbf{G}_{k-1}(\mathbf{x}_k - \mathbf{m}_{k|k-1}) \tag{53}$$
$$\mathbf{P}_2 = \mathbf{P}_{k-1|k-1} - \mathbf{G}_{k-1}\mathbf{P}_{k|k-1}\mathbf{G}_{k-1}^T \tag{54}$$

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

$$\begin{aligned} p\left(\mathbf{x}_{k-1}, \mathbf{x}_k \mid \mathbf{Y}\right) &= p\left(\mathbf{x}_{k-1} \mid \mathbf{x}_k, \mathbf{Y}_{1:k-1}\right) p\left(\mathbf{x}_k \mid \mathbf{Y}\right) \\ &\approx \mathrm{N}\left(\begin{bmatrix} \mathbf{m}_{k-1|N} \\ \mathbf{m}_{k|N} \end{bmatrix}, \begin{bmatrix} \mathbf{P}_{k-1|N} & \mathbf{D}_k \\ \mathbf{D}_k^T & \mathbf{P}_{k|N} \end{bmatrix}\right) \end{aligned} \tag{55}$$

where

$$\mathbf{D}_k = \mathbf{G}_{k-1}\mathbf{P}_{k|N} \tag{56}$$
$$\mathbf{m}_{k-1|N} = \mathbf{m}_{k-1|k-1} + \mathbf{G}_{k-1}\left(\mathbf{m}_{k|N} - \mathbf{m}_{k|k-1}\right) \tag{57}$$
$$\mathbf{P}_{k-1|N} = \mathbf{P}_{k-1|k-1} + \mathbf{G}_{k-1}\left(\mathbf{P}_{k|N} - \mathbf{P}_{k|k-1}\right)\mathbf{G}_{k-1}^T \tag{58}$$

4 Parameter estimation

4.1 Point estimates

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

$$\begin{aligned} p\left(\boldsymbol{\theta} \mid \mathbf{Y}\right) &= \frac{p\left(\mathbf{Y} \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right)}{p\left(\mathbf{Y}\right)} \\ \Rightarrow \log p\left(\boldsymbol{\theta} \mid \mathbf{Y}\right) &\propto \log p\left(\mathbf{Y} \mid \boldsymbol{\theta}\right) + \log p\left(\boldsymbol{\theta}\right) \end{aligned} \tag{59}$$

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

$$\mathbf{X} = \mathbf{X}_{1:N} = \begin{bmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_N \end{bmatrix} \tag{60}$$
$$\mathbf{Y} = \mathbf{Y}_{1:N} = \begin{bmatrix} \mathbf{y}_1 & \dots & \mathbf{y}_N \end{bmatrix} \tag{61}$$

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 $p\left(\mathbf{Y}\right)$. A point estimate that maximizes the posterior distribution is called a *maximum a posteriori* (MAP) estimate. In the case of a flat prior distribution $p\left(\boldsymbol{\theta}\right)$ the MAP

estimate converges to the *maximum likelihood* (ML) estimate, that maximizes the likelihood $p(\mathbf{Y} \mid \boldsymbol{\theta})$ (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.

Since our model contains the latent (unobserved) states, evaluating the likelihood $p(\mathbf{Y} \mid \boldsymbol{\theta})$ 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

$$p(\mathbf{X}, \mathbf{Y} \mid \boldsymbol{\theta}) = p(\mathbf{x}_0) \prod_{k=1}^{N} p(\mathbf{y}_k \mid \mathbf{x}_k)\, p(\mathbf{x}_k \mid \mathbf{x}_{k-1}) \tag{62}$$

so that the marginal likelihood is obtained by integration:

$$p(\mathbf{Y} \mid \boldsymbol{\theta}) = \int_{\mathbf{X}} p(\mathbf{X}, \mathbf{Y} \mid \boldsymbol{\theta})\ \mathrm{d}\mathbf{X} \tag{63}$$

Since $\mathbf{Y}$ is observed, equation (63) 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 (63) 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:

$$\mathbf{m}_{k|k-1} = \mathbf{A}\mathbf{m}_{k-1|k-1} \tag{64a}$$

$$\mathbf{P}_{k|k-1} = \mathbf{A}\mathbf{P}_{k-1|k-1}\mathbf{A}^T + \mathbf{Q} \tag{64b}$$

update:

$$\mathbf{v}_k = \mathbf{y}_k - \mathbf{H}\mathbf{m}_{k|k-1} \tag{64c}$$

$$\mathbf{S}_k = \mathbf{H}\mathbf{P}_{k|k-1}\mathbf{H}^T + \mathbf{R} \tag{64d}$$

$$\mathbf{K}_k = \mathbf{P}_{k|k-1}\mathbf{H}^T\mathbf{S}_k^{-1} \tag{64e}$$

$$\mathbf{m}_{k|k} = \mathbf{m}_{k|k-1} + \mathbf{K}_k\mathbf{v}_k \tag{64f}$$

$$\mathbf{P}_{k|k} = \mathbf{P}_{k|k-1} - \mathbf{K}_k\mathbf{S}_k\mathbf{K}_k^T \tag{64g}$$

This includes the sufficient statistics for the T joint distributions

$$\begin{aligned} p\left(\mathbf{x}_k, \mathbf{y}_k \mid \mathbf{y}_1, \ldots, \mathbf{y}_{k-1}, \boldsymbol{\theta}\right) &= N\left(\begin{bmatrix}\mathbf{x}_k \\ \mathbf{y}_k\end{bmatrix} \middle| \begin{bmatrix}\mathbf{m}_{k|k-1} \\ \mathbf{H}\mathbf{m}_{k|k-1}\end{bmatrix}, \begin{bmatrix}\mathbf{P}_{k|k-1} & \mathbf{P}_{k|k-1}\mathbf{H}^T \\ \mathbf{H}\mathbf{P}_{k|k-1}^T & \mathbf{S}_k\end{bmatrix}\right) \\ \Rightarrow p\left(\mathbf{y}_k \mid \mathbf{y}_1, \ldots, \mathbf{y}_{k-1}, \boldsymbol{\theta}\right) &= N\left(\mathbf{y}_k \mid \mathbf{H}\mathbf{m}_{k|k-1}, \mathbf{S}_k\right) \end{aligned} \tag{65}$$

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)

$$p\left(\mathbf{Y} \mid \boldsymbol{\theta}\right) = p\left(\mathbf{y}_1 \mid \boldsymbol{\theta}\right) \prod_{k=2}^{N} p\left(\mathbf{y}_k \mid \mathbf{y}_1, \ldots, \mathbf{y}_{k-1}, \boldsymbol{\theta}\right) \tag{66}$$

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

$$-2L(\boldsymbol{\theta}) = \sum_{k=1}^{N} \log\left|\mathbf{S}_k\right| + \sum_{k=1}^{N} \left(\mathbf{y}_k - \mathbf{H}\mathbf{m}_{k|k-1}\right)^T \mathbf{S}_k^{-1} \left(\mathbf{y}_k - \mathbf{H}\mathbf{m}_{k|k-1}\right) + C, \tag{67}$$

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 $\boldsymbol{\theta}$:

$$\begin{aligned} \frac{\partial L(\boldsymbol{\theta})}{\partial \theta_i} = &-\frac{1}{2}\sum_{k=1}^{N} \mathrm{Tr}\left(\mathbf{S}_k^{-1} \frac{\partial \mathbf{S}_k}{\partial \theta_i}\right) \\ &+ \sum_{k=1}^{N} \left(\mathbf{H}_k \frac{\partial \mathbf{m}_{k|k-1}}{\partial \theta_i}\right)^T \mathbf{S}_k^{-1} \left(\mathbf{y}_k - \mathbf{H}\mathbf{m}_{k|k-1}\right) \\ &+ \frac{1}{2}\sum_{k=1}^{N} \left(\mathbf{y}_k - \mathbf{H}\mathbf{m}_{k|k-1}\right)^T \mathbf{S}_k^{-1} \left(\frac{\partial \mathbf{S}_k}{\partial \theta_i}\right) \mathbf{S}_k^{-1} \left(\mathbf{y}_k - \mathbf{H}\mathbf{m}_{k|k-1}\right) \end{aligned} \tag{68}$$

From the Kalman filter recursions we find out that

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

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

$$\frac{\partial \mathbf{m}_{k|k-1}}{\partial \theta_i} = \frac{\partial \mathbf{A}}{\partial \theta_i}\,\mathbf{m}_{k-1|k-1} + \mathbf{A}\,\frac{\partial \mathbf{m}_{k-1|k-1}}{\partial \theta_i} \tag{70}$$

$$\begin{aligned}\frac{\partial \mathbf{P}_{k|k-1}}{\partial \theta_i} = {} & \frac{\partial \mathbf{A}}{\partial \theta_i}\,\mathbf{P}_{k-1|k-1}\mathbf{A}^T + \mathbf{A}\,\frac{\partial \mathbf{P}_{k-1|k-1}}{\partial \theta_i}\,\mathbf{A}^T \\ & + \mathbf{A}\mathbf{P}_{k-1|k-1}\left(\frac{\partial \mathbf{A}}{\partial \theta_i}\right)^T + \frac{\partial \mathbf{Q}}{\partial \theta_i}\end{aligned} \tag{71}$$

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

$$\frac{\partial \mathbf{K}_k}{\partial \theta_i} = \frac{\partial \mathbf{P}_{k|k-1}}{\partial \theta_i}\,\mathbf{H}^T\mathbf{S}_k^{-1} - \mathbf{P}_{k|k-1}\mathbf{H}^T\mathbf{S}_k^{-1}\,\frac{\partial \mathbf{S}_k}{\partial \theta_i}\,\mathbf{S}_k^{-1} \tag{72}$$

$$\frac{\partial \mathbf{m}_{k|k}}{\partial \theta_i} = \frac{\partial \mathbf{m}_{k|k-1}}{\partial \theta_i} + \frac{\partial \mathbf{K}_k}{\partial \theta_i}\left(\mathbf{y}_k - \mathbf{H}\mathbf{m}_{k|k-1}\right) - \mathbf{K}_k\mathbf{H}\,\frac{\partial \mathbf{m}_{k|k-1}}{\partial \theta_i} \tag{73}$$

$$\frac{\partial \mathbf{P}_{k|k}}{\partial \theta_i} = \frac{\partial \mathbf{P}_{k|k-1}}{\partial \theta_i} - \frac{\partial \mathbf{K}_k}{\partial \theta_i}\,\mathbf{S}_k\mathbf{K}_k^T - \mathbf{K}_k\,\frac{\partial \mathbf{S}_k}{\partial \theta_i}\,\mathbf{K}_k^T - \mathbf{K}_k^T\mathbf{S}_k\left(\frac{\partial \mathbf{K}_k}{\partial \theta_i}\right)^T \tag{74}$$

Equations (70), (71), (72), (73) and (74) together specify a recursive algorithm for computing (68) 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 (68) 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 $q\left(\mathbf{X}\right)$ over the states. We can then decompose the log-likelihood function as follows:

$$L(\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}\left(q \| p\right) \tag{75}$$

where

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

$$\mathrm{KL}\left(q \| p\right) = -\int_{\mathbf{X}} q\left(\mathbf{X}\right) \log\left(\frac{p\left(\mathbf{X} \mid \mathbf{Y}, \boldsymbol{\theta}\right)}{q\left(\mathbf{X}\right)}\right)\,\mathrm{d}\mathbf{X} \tag{77}$$

It is easy to verify the decomposition (75) by substituting

$$\log p\left(\mathbf{X}, \mathbf{Y} \mid \boldsymbol{\theta}\right) = \log p\left(\mathbf{X} \mid \mathbf{Y}, \boldsymbol{\theta}\right) + L(\boldsymbol{\theta}) \tag{78}$$

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

$$L(\boldsymbol{\theta}) \geq \mathcal{L}(q, \boldsymbol{\theta}) \tag{79}$$

with equality when

$$\begin{aligned} &\mathrm{KL}\left(q\|p\right) = 0 \\ &\Rightarrow q\left(\mathbf{X}\right) = p\left(\mathbf{X} \mid \mathbf{Y}, \boldsymbol{\theta}\right) \end{aligned} \tag{80}$$

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 (79) and (80), the maximum is obtained with $q^*(\mathbf{X}) = p\left(\mathbf{X} \,\middle|\, \mathbf{Y}, \hat{\boldsymbol{\theta}}_j\right)$, the posterior distribution of the states given the current parameter estimate. After the maximization we have

$$L\left(\hat{\boldsymbol{\theta}}_j\right) = \mathcal{L}\left(p\left(\mathbf{X} \,\middle|\, \mathbf{Y}, \hat{\boldsymbol{\theta}}_j\right), \hat{\boldsymbol{\theta}}_j\right). \tag{81}$$

   This is the *E-step*

2. Maximize $\mathcal{L}\left(p\left(\mathbf{X} \,\middle|\, \mathbf{Y}, \hat{\boldsymbol{\theta}}_j\right), \boldsymbol{\theta}\right)$ 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):

$$L\left(\hat{\boldsymbol{\theta}}_{j+1}\right) - L\left(\hat{\boldsymbol{\theta}}_j\right) \geq \mathcal{L}\left(p\left(\mathbf{X} \,\middle|\, \mathbf{Y}, \hat{\boldsymbol{\theta}}_j\right), \hat{\boldsymbol{\theta}}_{j+1}\right) - \mathcal{L}\left(p\left(\mathbf{X} \,\middle|\, \mathbf{Y}, \hat{\boldsymbol{\theta}}_j\right), \hat{\boldsymbol{\theta}}_j\right) \tag{82}$$

which is just the combination of (79) and (81). But it highlights the fact that the likelihood is increased with every new estimate $\hat{\boldsymbol{\theta}}_{j+1}$. Also following from (82) 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 $\mathcal{L}\left(p\left(\mathbf{X} \,\middle|\, \mathbf{Y}, \hat{\boldsymbol{\theta}}_l\right), \boldsymbol{\theta}\right)$ 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 (77) are continuously differentiable, then

$$\left.\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_i}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_j} = \left.\frac{\partial \mathcal{L}\left(p\left(\mathbf{X} \,\middle|\, \mathbf{Y}, \hat{\boldsymbol{\theta}}_j\right), \boldsymbol{\theta}\right)}{\partial \theta_i}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_j} \tag{83}$$

This was implicitly clear already from (81) 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 (76), we get

$$\begin{aligned}\mathcal{L}(q, \boldsymbol{\theta}) &= \int_{\mathbf{X}} p\left(\mathbf{X} \,\middle|\, \mathbf{Y}, \hat{\boldsymbol{\theta}}\right) \log p\left(\mathbf{X}, \mathbf{Y} \mid \boldsymbol{\theta}\right) \, \mathrm{d}\mathbf{X} - \int_{\mathbf{X}} p\left(\mathbf{X} \,\middle|\, \mathbf{Y}, \hat{\boldsymbol{\theta}}\right) \log p\left(\mathbf{X} \,\middle|\, \mathbf{Y}, \hat{\boldsymbol{\theta}}\right) \, \mathrm{d}\mathbf{X} \\ &= \mathfrak{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) + C,\end{aligned} \tag{84}$$

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.

4.4 EM in linear-Gaussian SSM:s

Continuing with the application of EM to the linear-Gaussian state space models, the complete-data log-likelihood function is now

$$\begin{aligned}L(\mathbf{X}, \boldsymbol{\theta}) = &\frac{1}{2}\left(\mathbf{x}_0 - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_0 - \boldsymbol{\mu}\right) + \frac{1}{2}\log|\boldsymbol{\Sigma}| \\ &+ \frac{1}{2}\sum_{k=1}^{N}\left(\mathbf{y}_k - \mathbf{H}\mathbf{x}_k\right)^T \mathbf{R}_{k|k-1}{-1}\left(\mathbf{y}_k - \mathbf{H}\mathbf{x}_k\right) + \frac{N}{2}\log|\mathbf{R}| \\ &+ \frac{1}{2}\sum_{k=1}^{N}\left(\mathbf{x}_k - \mathbf{A}\mathbf{x}_{k-1}\right)^T \mathbf{Q}^{-1}\left(\mathbf{x}_k - \mathbf{A}\mathbf{x}_{k-1}\right) + \frac{N}{2}\log|\mathbf{Q}| \\ &+ C\end{aligned} \tag{85}$$

We can then take the expectation of (85):

$$\mathfrak{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_j) = \left\langle L(\mathbf{X}, \boldsymbol{\theta})\right\rangle_{\hat{\boldsymbol{\theta}}_j} = \tag{86}$$

$$\begin{aligned}&\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{P}_{0|N} + (\mathbf{m}_{0|N} - \boldsymbol{\mu})(\mathbf{m}_{0|N} - \boldsymbol{\mu})^T\right)\right] + \log|\boldsymbol{\Sigma}| \\ &+\mathrm{Tr}\left[\mathbf{Q}^{-1}\left(\sum_{k=1}^{N}\left\langle \mathbf{x}_k\mathbf{x}_k^T\right\rangle - \mathbf{A}\sum_{k=1}^{N}\left\langle \mathbf{x}_k\mathbf{x}_{k-1}^T\right\rangle^T - \sum_{k=1}^{N}\left\langle \mathbf{x}_k\mathbf{x}_{k-1}^T\right\rangle\mathbf{A}^T + \mathbf{A}\sum_{k=1}^{N}\left\langle \mathbf{x}_{k-1}\mathbf{x}_{k-1}^T\right\rangle\mathbf{A}^T\right)\right] + N\log \\ &+\mathrm{Tr}\left[\mathbf{R}^{-1}\left(\sum_{k=1}^{N}\mathbf{y}_k\mathbf{y}_k^T - \mathbf{H}\sum_{k=1}^{N}\mathbf{y}_k\left\langle \mathbf{x}_k^T\right\rangle^T - \sum_{k=1}^{N}\mathbf{y}_k\left\langle \mathbf{x}_k^T\right\rangle\mathbf{H}^T + \mathbf{H}\sum_{k=1}^{N}\left\langle \mathbf{x}_k\mathbf{x}_k^T\right\rangle\mathbf{A}^T\right)\right] + N\log|\mathbf{R}|\end{aligned} \tag{87}$$

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

$$\langle \mathbf{x}_k \rangle = \mathbf{m}_{k|N} \tag{88}$$

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

$$\left\langle \mathbf{x}_k \mathbf{x}_{k-1}^T \right\rangle = \mathbf{C}_{k|N} + \mathbf{m}_{k|N} (\mathbf{m}_{k-1|N})^T, \tag{90}$$

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.

4.5 Alternative derivation

It's not necessary to to multiply the terms inside the second order terms in Equation (85). First of all, for the joint distribution of $\mathbf{x}_k, \mathbf{x}_{k-1}$ we get

$$\begin{aligned} p(\mathbf{x}_k, \mathbf{x}_{k-1} \mid \mathbf{Y}, \boldsymbol{\theta}) &= N\left(\begin{bmatrix} \mathbf{x}_k \\ \mathbf{x}_{k-1} \end{bmatrix} \middle| \begin{bmatrix} \mathbf{m}_{k|N} \\ \mathbf{m}_{k-1|N} \end{bmatrix}, \begin{bmatrix} \mathbf{P}_{k|N} & \mathbf{P}_{k|N} \mathbf{J}_k^T \\ \mathbf{J}_k \mathbf{P}_{k|N} & \mathbf{P}_{k-1|N} \end{bmatrix} \right) \\ &= N\left(\begin{bmatrix} \mathbf{x}_k \\ \mathbf{x}_{k-1} \end{bmatrix} \middle| \mathbf{m}_k^{(3)}, \mathbf{P}_k^{(3)} \right), \end{aligned} \tag{91}$$

where we have the identities

$$\left\langle \mathbf{x}_k \mathbf{x}_k^T \right\rangle = \mathbf{P}_{k|N} + \mathbf{m}_{k|N} \mathbf{m}_{k|N}^T \tag{92}$$

$$\left\langle \mathbf{x}_{k-1} \mathbf{x}_{k-1}^T \right\rangle = \mathbf{P}_{k-1|N} + \mathbf{m}_{k-1|N} \mathbf{m}_{k-1|N}^T \tag{93}$$

$$\left\langle \mathbf{x}_k \mathbf{x}_{k-1}^T \right\rangle = \mathbf{P}_{k|N} \mathbf{J}_k^T + \mathbf{m}_{k|N} \mathbf{m}_{k-1|N}^T \tag{94}$$

Instead we get

$$\begin{aligned} &\left\langle (\mathbf{x}_k - \mathbf{A}\mathbf{x}_{k-1})^T \mathbf{Q}^{-1} (\mathbf{x}_k - \mathbf{A}\mathbf{x}_{k-1}) \right\rangle \\ &= \mathrm{Tr}\left[\mathbf{Q}^{-1} \left\langle (\mathbf{x}_k - \mathbf{A}\mathbf{x}_{k-1}) (\mathbf{x}_k - \mathbf{A}\mathbf{x}_{k-1})^T \right\rangle \right] \\ &= \mathrm{Tr}\left[\mathbf{Q}^{-1} \left(\mathrm{Cov}(\mathbf{x}_k - \mathbf{A}\mathbf{x}_{k-1}) + \langle \mathbf{x}_k - \mathbf{A}\mathbf{x}_{k-1} \rangle \langle \mathbf{x}_k - \mathbf{A}\mathbf{x}_{k-1} \rangle^T \right) \right] \\ &= \mathrm{Tr}\left[\mathbf{Q}^{-1} [\mathbf{I}, -\mathbf{A}] \left(\mathbf{P}_k^{(3)} + \mathbf{m}_k^{(3)} (\mathbf{m}_k^{(3)})^T \right) [\mathbf{I}, -\mathbf{A}]^T \right] \end{aligned} \tag{95}$$

and

$$\begin{aligned}
&\left\langle \left(\mathbf{y}_k - \mathbf{H}\mathbf{x}_k\right)^T \mathbf{R}^{-1} \left(\mathbf{y}_k - \mathbf{H}\mathbf{x}_k\right) \right\rangle \\
&= \mathrm{Tr}\left[\mathbf{R}^{-1} \left\langle \left(\mathbf{y}_k - \mathbf{H}\mathbf{x}_k\right) \left(\mathbf{y}_k - \mathbf{H}\mathbf{x}_k\right)^T \right\rangle \right] \\
&= \mathrm{Tr}\left[\mathbf{R}^{-1} \left(\mathbf{H}\mathrm{Var}\left(\mathbf{x}_k\right) \mathbf{H}^T + \left(\mathbf{y}_k - \mathbf{H}\left\langle \mathbf{x}_k \right\rangle\right) \left(\mathbf{y}_k - \mathbf{H}\left\langle \mathbf{x}_k \right\rangle\right)^T \right) \right]
\end{aligned} \tag{96}$$

With these, we have

$$-2\,\frac{\partial \mathfrak{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_j)}{\partial \mathbf{A}} = \sum_{k=1}^{N} \frac{\partial}{\partial \mathbf{A}} \mathrm{Tr}\left[\mathbf{Q}^{-1} \left(\mathbf{I} - \mathbf{A}\right) \left(\mathbf{P}_k^{(3)} + \mathbf{m}_k^{(3)} (\mathbf{m}_k^{(3)})^T \right) \left(\mathbf{I} - \mathbf{A}\right)^T \right] \tag{97}$$

With these, we have

$$\begin{aligned}
-2\,\frac{\partial \mathfrak{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_j)}{\partial \mathbf{Q}^{-\mathbf{1}}} = \sum_{k=1}^{N} \frac{\partial}{\partial \mathbf{Q}^{-\mathbf{1}}} \mathrm{Tr}\left[\mathbf{Q}^{-1} \left(\mathbf{I} - \mathbf{A}\right) \left(\mathbf{P}_k^{(3)} + \mathbf{m}_k^{(3)} (\mathbf{m}_k^{(3)})^T \right) \left(\mathbf{I} - \mathbf{A}\right)^T \right] \\
+ N \frac{\partial}{\partial \mathbf{Q}^{-\mathbf{1}}} \log |\mathbf{Q}|
\end{aligned} \tag{98}$$

Using formula 92 in (Petersen et al. 2008) for the first derivative and formula 51 for the second, we get

$$-2\,\frac{\partial \mathfrak{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_j)}{\partial \mathbf{Q}^{-\mathbf{1}}} = \left(\mathbf{I} - \mathbf{A}\right) \sum_{k=1}^{N} \left(\mathbf{P}_k^{(3)} + \mathbf{m}_k^{(3)} (\mathbf{m}_k^{(3)})^T \right) \left(\mathbf{I} - \mathbf{A}\right)^T - N\mathbf{Q} \tag{99}$$

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

$$\mathbf{Q}_{j+1} = \frac{1}{N} \left(\mathbf{I} - \mathbf{A}\right) \sum_{k=1}^{N} \left(\mathbf{P}_k^{(3)} + \mathbf{m}_k^{(3)} (\mathbf{m}_k^{(3)})^T \right) \left(\mathbf{I} - \mathbf{A}\right)^T \tag{100}$$

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):

$$\mathbf{A}_{j+1} = \left(\sum_{k=1}^{N} \left\langle \mathbf{x}_k \mathbf{x}_{k-1}^T \right\rangle\right) \left(\sum_{k=1}^{N} \left\langle \mathbf{x}_{k-1} \mathbf{x}_{k-1}^T \right\rangle\right)^{-1} \tag{101a}$$

$$\mathbf{Q}_{j+1} = \sum_{k=1}^{N} \left\langle \mathbf{x}_k \mathbf{x}_k^T \right\rangle - \mathbf{A}_{j+1} \left(\sum_{k=1}^{N} \left\langle \mathbf{x}_k \mathbf{x}_{k-1}^T \right\rangle\right)^T \tag{101b}$$

$$\mathbf{H}_{j+1} = \sum_{k=1}^{N} \mathbf{y}_k \left\langle \mathbf{x}_k^T \right\rangle \left(\sum_{k=1}^{N} \left\langle \mathbf{x}_k \mathbf{x}_k^T \right\rangle\right) - 1 \tag{101c}$$

$$\mathbf{R}_{j+1} = \sum_{k=1}^{N} \mathbf{y}_k \mathbf{y}_k^T - \mathbf{H}_{j+1} \left(\sum_{k=1}^{N} \mathbf{y}_k \left\langle \mathbf{x}_k^T \right\rangle\right)^T \tag{101d}$$

4.6 M-step with structured matrices

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:

$$\begin{aligned}
-2\,\frac{\partial \mathfrak{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}))}{\partial \theta_i} &= \mathrm{Tr}\left[-\mathbf{Q}^{-1}\,\frac{\partial \mathbf{Q}}{\partial \theta_i}\,\mathbf{Q}^{-1}\left(\mathbf{B}_1 - \mathbf{A}\mathbf{B}_2^T - \mathbf{B}_2\mathbf{A}^T + \mathbf{A}\mathbf{B}_3\mathbf{A}^T\right)\right] \\
&\quad + \mathrm{Tr}\left[\mathbf{Q}^{-1}\left(-\frac{\partial \mathbf{A}}{\partial \theta_i}\,\mathbf{B}_2^T - \mathbf{B}_2\,\frac{\partial \mathbf{A}}{\partial \theta_i}^T + \frac{\partial \mathbf{A}}{\partial \theta_i}\,\mathbf{B}_3\mathbf{A}^T + \mathbf{A}\mathbf{B}_3\,\frac{\partial \mathbf{A}}{\partial \theta_i}^T\right)\right] \\
&\quad + \mathrm{Tr}\left[-\mathbf{R}^{-1}\,\frac{\partial \mathbf{R}}{\partial \theta_i}\,\mathbf{R}^{-1}\left(\mathbf{B}_4 - \mathbf{H}\mathbf{B}_5^T - \mathbf{B}_5\mathbf{H}^T + \mathbf{H}\mathbf{B}_1\mathbf{H}^T\right)\right] \\
&\quad + N\mathrm{Tr}\left[\mathbf{Q}^{-1}\,\frac{\partial \mathbf{Q}}{\partial \theta_i}\right] + N\mathrm{Tr}\left[\mathbf{R}^{-1}\,\frac{\partial \mathbf{R}}{\partial \theta_i}\right]
\end{aligned} \tag{102}$$

*Todo list

References

Barber, D, A T Cemgil, and S Chiappa (2011). *Bayesian Time Series Models*. Cambridge University Press. ISBN: 9780521196765. URL: http://books.google.fi/books?id=k4z6m0FsEv8C.

Bishop, C.M. (2006). *Pattern recognition and machine learning*. Springer Verlag. ISBN: 9780387310732. URL: http://www.library.wisc.edu/selectedtocs/bg0137.pdf.

Cappé, O., E. Moulines, and T. Rydén (2005). *Inference in hidden Markov models*. Springer Verlag. ISBN: 9780387402642. URL: http://books.google.com/books?hl=en\&lr=\&id=4d_oEYn8Fl0C\&oi=fnd\&pg=PR5\&dq=Inference+in+Hidden+Markov+Models\&ots=tima6AR1qw\&sig=nY0OHyJotdhsNdhPkJcCLsiFbGc.

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. URL: http://www.jstor.org/stable/10.2307/2984875.

Durbin, J and S J Koopman (2012). *Time Series Analysis by State Space Methods: Second Edition*. Oxford Statistical Science Series. OUP Oxford. ISBN: 9780199641178. URL: http://books.google.fi/books?id=f0q39Zh0olQC.

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. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=763210.

Gelman, A et al. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC. ISBN: 9781584883883. URL: http://books.google.fi/books?id=TNYhnkXQSjAC.

Ghahramani, Zoubin (1996). "Parameter estimation for linear dynamical systems". In: *University of Toronto technical report CRG-TR*, pp. 1–6. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.55.5997\&rep=rep1\&type=pdf.

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. URL: http://linkinghub.elsevier.com/retrieve/pii/S0005109805001810.

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*.

Minka, T (1998). "Expectation-Maximization as lower bound maximization". In: *Tutorial published on the web at http://www-white.* 1977, pp. 1–8. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.29.8562\&rep=rep1\&type=pdf.

Pearl, J (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann Publishers. ISBN: 9781558604797. URL: http://books.google.fi/books?id=AvNID7LyMusC.

Petersen, K.B. and M.S. Pedersen (2008). "The matrix cookbook". In: *Technical University of Denmark*, pp. 7–15. URL: http://gugaguigui.mouciel.com/etudes/cours/archives-web-cours/IFT6266/web/hebdomadaire/3/matrix_cookbook.pdf.

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