Ville Väänänen / 63527M
ville.vaananen@aalto.fi

Parameter estimation in linear-Gaussian state-space models

S-114.4202 Special Course in Computational Engineering II

January 26, 2012

1 Introduction

Parameter estimation refers to a broad class of problems in machine learning. Suppose we have a linear gaussian state-space model (SSM) of the form:

$$\begin{aligned}\mathbf{x}_k &= \mathbf{A}\mathbf{x}_{k-1} + \mathbf{q}_{k-1} \\ \mathbf{y}_k &= \mathbf{H}\mathbf{x}_\mathbf{k} + \mathbf{r}_k \\ \mathbf{q}_{k-1} &\sim \mathrm{N}\left(0, \mathbf{Q}\right) \\ \mathbf{r}_k &\sim \mathrm{N}\left(0, \mathbf{R}\right)\end{aligned}$$

meaning

$$\begin{aligned}\mathbf{x}_k|\mathbf{x}_{k-1} &\sim \mathrm{N}\left(\mathbf{A}\mathbf{x}_{k-1}, \mathbf{Q}\right) \\ \mathbf{y}_k|\mathbf{x}_k &\sim \mathrm{N}\left(\mathbf{H}\mathbf{x}_\mathbf{k}, \mathbf{R}\right)\end{aligned} \tag{1}$$

Let us denote the set of parameters of this model with $\boldsymbol{\theta}$ and we assume an implicit dependance of the matrices $\{\mathbf{A}, \mathbf{H}, \mathbf{Q}, \mathbf{R}\}$ on $\boldsymbol{\theta}$. The matrices of all the states and all the observations are denoted with

$$\begin{aligned}\mathbf{X} = \mathbf{X}_{1:T} &= \begin{bmatrix}\mathbf{x}_1 & \dots & \mathbf{x}_T\end{bmatrix} \\ \mathbf{Y} = \mathbf{Y}_{1:T} &= \begin{bmatrix}\mathbf{y}_1 & \dots & \mathbf{y}_T\end{bmatrix}\end{aligned}$$

respectively.

In the Bayesian sense the complete answer to the parameter estimation problem is the marginal posterior probability

$$\begin{aligned}p\left(\boldsymbol{\theta} \mid \mathbf{Y}\right) &= \frac{p\left(\mathbf{Y} \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right)}{p\left(\mathbf{Y}\right)} \\ \Rightarrow \log p\left(\boldsymbol{\theta} \mid \mathbf{Y}\right) &\propto \log p\left(\mathbf{Y} \mid \boldsymbol{\theta}\right) + \log p\left(\boldsymbol{\theta}\right)\end{aligned} \tag{2}$$

The marginal likelihood $p\left(\mathbf{Y} \mid \boldsymbol{\theta}\right)$ can be obtained by marginalization from the complete-data likelihood. Because of the Markov conditional independence properties of (1), the complete-data likelihood can be written as

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

so that the marginal likelihood is obtained by integration:

$$p\left(\mathbf{Y} \mid \boldsymbol{\theta}\right) = \int_{\mathbf{X}} p\left(\mathbf{X}, \mathbf{Y} \mid \boldsymbol{\theta}\right) \, \mathrm{d}\mathbf{X} \tag{3}$$

Since $\mathbf{Y}$ is observed, (3) is a function of the parameters only.

Maximizing (2) (i.e. finding the MAP estimate) is equal minimizing the so-called energy function

$$\varphi(\boldsymbol{\theta}) = -\log p\left(\mathbf{Y} \mid \boldsymbol{\theta}\right) - \log p\left(\boldsymbol{\theta}\right) \tag{4}$$

The Kalman filter forward recursions give us the means to perform the integration over the states analytically, so that (3) can be evaluated for any given $\boldsymbol{\theta}$.

2 Methods

2.1 Gradient based search

This is the classical way of solving the problem. It consists of computing the gradient of the energy function and using some non-linear optimization method to find its minimum. An effiecient algorithm is the scaled conjugate gradient method. A couple of problems are assoaciated with this approach. Firstly, calculating the gradient of $\varphi(\boldsymbol{\theta})$ at best tedious. And secondly, the result will only be a point estimate to a probability distribution.

To derive the expression for the energy function in our case, let us first see what the Kalman filter calculates. Firstly, the recursions are as follows:

prediction:

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

update:

$$\begin{aligned} \mathbf{v}_k &= \mathbf{y}_k - \mathbf{H}\mathbf{m}_k^- \\ \mathbf{S}_k &= \mathbf{H}\mathbf{P}_k^-\mathbf{H} + \mathbf{R} \\ \mathbf{K}_k &= \mathbf{P}_k^-\mathbf{H}^T\mathbf{S}_k^{-1} \\ \mathbf{m}_k &= \mathbf{m}_k^- + \mathbf{K}_k\mathbf{v}_k \\ \mathbf{P}_k &= \mathbf{P}_k^- - \mathbf{K}_k\mathbf{S}_k\mathbf{K}_k^T \end{aligned}$$

This includes the sufficient statistics for the T joint distributions

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

To see how this enables us to calculate (4), one only needs to note that (it has been assumed that the observations are independent given the states)

$$p\left(\mathrm{Y} \mid \boldsymbol{\theta}\right) = p\left(\mathbf{y}_1 \mid \boldsymbol{\theta}\right) \prod_{k=2}^{T} p\left(\mathbf{y}_k \mid \mathrm{Y}_{1:k-1}, \boldsymbol{\theta}\right)$$

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

$$\varphi(\boldsymbol{\theta}) = \frac{1}{2}\sum_{k=1}^{T} \log|\mathrm{S}_k| + \frac{1}{2}\sum_{k=1}^{T} \left(\mathbf{y}_k - \mathbf{H}\mathbf{m}_k^-\right)^T \mathrm{S}_k^{-1} \left(\mathbf{y}_k - \mathbf{H}\mathbf{m}_k^-\right) + \log p\left(\boldsymbol{\theta}\right) + C \tag{5}$$

In order to calculate the gradient, we need to formally derivate (5) w.r.t every parameter θ_i:

$$\begin{aligned}
\frac{\partial\varphi(\boldsymbol{\theta})}{\partial\theta_i} &= \frac{1}{2}\mathrm{Tr}\left(\mathrm{S}_k^{-1}\,\frac{\partial \mathrm{S}_k}{\partial\theta_i}\right) \\
&- \frac{1}{2}\sum_{k=1}^{T}\left(\mathbf{H}_k\,\frac{\partial\mathbf{m}_k^-}{\partial\theta_i}\right)^T \mathrm{S}_k^{-1}\left(\mathbf{y}_k - \mathbf{H}\mathbf{m}_k^-\right) \\
&- \frac{1}{2}\sum_{k=1}^{T}\left(\mathbf{y}_k - \mathbf{H}\mathbf{m}_k^-\right)^T \mathrm{S}_k^{-1}\left(\frac{\partial \mathrm{S}_k}{\partial\theta_i}\right)\mathrm{S}_k^{-1}\left(\mathbf{y}_k - \mathbf{H}\mathbf{m}_k^-\right) \\
&- \frac{1}{2}\sum_{k=1}^{T}\left(\mathbf{y}_k - \mathbf{H}\mathbf{m}_k^-\right)^T \mathrm{S}_k^{-1}\left(\mathbf{H}_k\,\frac{\partial\mathbf{m}_k^-}{\partial\theta_i}\right)
\end{aligned}$$

From Kalman filter recursions we find out that

$$\frac{\partial \mathrm{S}_k}{\partial\theta_i} = \mathbf{H}\,\frac{\partial\mathbf{P}_k^-}{\partial\theta_i}\,\mathbf{H} + \frac{\partial\mathbf{R}}{\partial\theta_i}$$

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

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

as well as for $\mathbf{m}_k$ and $\mathbf{P}_k$:

$$\begin{aligned}
\frac{\partial\mathbf{K}_k}{\partial\theta_i} &= \frac{\partial\mathbf{P}_k^-}{\partial\theta_i}\,\mathbf{H}^T\mathrm{S}_k^{-1} + \mathbf{P}_k^-\mathbf{H}^T\mathrm{S}_k^{-1}\left(\mathbf{H}\,\frac{\partial\mathbf{P}_k^-}{\partial\theta_i}\,\mathbf{H} + \frac{\partial\mathbf{R}}{\partial\theta_i}\right)\mathrm{S}_k^{-1} \\
\frac{\partial\mathbf{m}_k}{\partial\theta_i} &= \frac{\partial\mathbf{m}_k^-}{\partial\theta_i} + \frac{\partial\mathbf{K}_k}{\partial\theta_i}\left(\mathbf{y}_k - \mathbf{H}\mathbf{m}_k^-\right) - \mathbf{K}_k\mathbf{H}\,\frac{\partial\mathbf{m}_k^-}{\partial\theta_i} \\
\frac{\partial\mathbf{P}_k}{\partial\theta_i} &= \frac{\partial\mathbf{P}_k^-}{\partial\theta_i} - \frac{\partial\mathbf{K}_k}{\partial\theta_i}\,\mathrm{S}_k\mathbf{K}_k^T - \mathbf{K}_k\left(\mathbf{H}\,\frac{\partial\mathbf{P}_k^-}{\partial\theta_i}\,\mathbf{H} + \frac{\partial\mathbf{R}}{\partial\theta_i}\right)\mathbf{K}_k^T - \mathbf{K}_k^T\mathrm{S}_k\left(\frac{\partial\mathbf{K}_k}{\partial\theta_i}\right)^T
\end{aligned}$$

2.2 Expectation maximization (EM)

In order to derive the EM algorithm, let us imagine temporarily that also the states $\mathbf{X}$ have been observed. In this case the likelihood doesn't have to be marginalized, since we already know everything to calculate it:

$$\log p\left(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{Y}\right) \propto \log p\left(\mathbf{X}, \mathbf{Y} \mid \boldsymbol{\theta}\right) + \log p\left(\boldsymbol{\theta}\right)$$

and so the energy function is now

$$\begin{aligned}
\varphi(\boldsymbol{\theta}) = & \frac{1}{2}\sum_{k=1}^{T} \log\left|\mathbf{R}_k\right| + \frac{T}{2}\log\left|\mathbf{Q}\right| \\
& + \frac{1}{2}\sum_{k=1}^{T}\left(\mathbf{y}_k - \mathbf{H}\mathbf{x}_k\right)^T \mathbf{R}_k^{-1}\left(\mathbf{y}_k - \mathbf{H}\mathbf{x}_k\right) \\
& + \frac{1}{2}\sum_{k=1}^{T}\left(\mathbf{x}_k - \mathbf{A}\mathbf{x}_{k-1}\right)^T \mathbf{Q}^{-1}\left(\mathbf{x}_k - \mathbf{A}\mathbf{x}_{k-1}\right) \\
& + \log p\left(\boldsymbol{\theta}\right) + C
\end{aligned}$$

EM is an iterative algorithm, where we have to start from some initial parameter value $\boldsymbol{\theta}_{\mathbf{0}}$. In the j:th iteration of the algorithm, we first form the posterior distribution of the latent variables given the previous parameter values:

$$p\left(\mathbf{X} \,\middle|\, \mathbf{Y}, \boldsymbol{\theta}_{j-1}\right) = \frac{p\left(\mathbf{x}_0 \,\middle|\, \boldsymbol{\theta}_{j-1}\right)\prod_{k=1}^{T} p\left(\mathbf{y}_k \,\middle|\, \mathbf{x}_k, \boldsymbol{\theta}_{j-1}\right) p\left(\mathbf{x}_k \,\middle|\, \mathbf{x}_{k-1}, \boldsymbol{\theta}_{j-1}\right)}{p\left(\mathbf{y}_1 \,\middle|\, \boldsymbol{\theta}_{\mathbf{j-1}}\right)\prod_{k=2}^{T} p\left(\mathbf{y}_k \,\middle|\, \mathbf{Y}_{1:k-1}, \boldsymbol{\theta}_{j-1}\right)}$$

This is called the E-step. In the subsequent M-step, we obtain the new estimate $\boldsymbol{\theta}_k$ by maximizing the expectation of the energy function, where the expectation is calculated over $p\left(\mathbf{X} \,\middle|\, \mathbf{Y}, \boldsymbol{\theta}_{j-1}\right)$:

$$\begin{aligned}
\boldsymbol{\theta}_j &= \arg\max_{\boldsymbol{\theta}}\left(\int_{\mathbf{X}} p\left(\mathbf{X} \,\middle|\, \mathbf{Y}, \boldsymbol{\theta}_{j-1}\right)\varphi(\boldsymbol{\theta})\ \mathrm{d}\mathbf{X}\right) \\
&= \arg\max_{\boldsymbol{\theta}}\left(L\left(\boldsymbol{\theta}, \boldsymbol{\theta}_{j-1}\right)\right)
\end{aligned}$$

In our case we get the following form for the funtion to be maximized in the M-step on iteration j:

$$
\begin{aligned}
L\left(\boldsymbol{\theta}, \boldsymbol{\theta}_{j-1}\right) = & \frac{1}{2} \sum_{k=1}^{T} \log |\mathbf{R}_k| + \frac{T}{2} \log |\mathbf{Q}| \\
& + \frac{1}{2} \sum_{k=1}^{T} \mathbf{y}_k^T \mathbf{R}^{-1} \mathbf{y}_k - \sum_{k=1}^{T} \mathbf{y}_k^T \mathbf{R}^{-1} \mathbf{H} \left\langle \mathbf{x}_k \right\rangle + \frac{1}{2} \sum_{k=1}^{T} \operatorname{tr} \left(\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} \left\langle \mathbf{x}_k \mathbf{x}_k^T \right\rangle \right) \\
& + \frac{1}{2} \sum_{k=1}^{T} \operatorname{tr} \left(\mathbf{Q}^{-1} \left\langle \mathbf{x}_k \mathbf{x}_k^T \right\rangle \right) - \sum_{k=1}^{T} \operatorname{tr} \left(\mathbf{Q}^{-1} \mathbf{A} \left\langle \mathbf{x}_k \mathbf{x}_{k-1}^T \right\rangle \right) \\
& + \frac{1}{2} \sum_{k=1}^{T} \operatorname{tr} \left(\mathbf{A}^T \mathbf{Q}^{-1} \mathbf{A} \left\langle \mathbf{x}_{k-1} \mathbf{x}_{k-1}^T \right\rangle \right) + \log p\left(\boldsymbol{\theta}\right) + C \qquad (6)
\end{aligned}
$$

The expectations that are left in (6) can be calculated with the Kalman smoother:

$$
\begin{aligned}
\left\langle \mathbf{x}_k \right\rangle &= \mathbf{m}_k^S \\
\left\langle \mathbf{x}_k \mathbf{x}_k^T \right\rangle &= \mathbf{P}_k^S + \mathbf{m}_k^S (\mathbf{m}_k^S)^T \\
\left\langle \mathbf{x}_k \mathbf{x}_{k-1}^T \right\rangle &= \mathbf{P}_k^S + \mathbf{m}_k^S (\mathbf{m}_k^S)^T
\end{aligned}
$$

After we have calculated the sufficient statistics with the Kalman smoother given $\boldsymbol{\theta}_{j-1}$, we proceed to estimate the new value $\boldsymbol{\theta}_{j-1}$ by finding the maximum of $L\left(\boldsymbol{\theta}, \boldsymbol{\theta}_{j-1}\right)$ in the M-step. For that, we take the derivatives of (6) w.r.t each of the parameters. Let us first proceed by taking the derivatives w.r.t $\{\mathbf{A}, \mathbf{H}, \mathbf{Q}, \mathbf{R}\}$ after which we'll apply the chain rule

$$
\frac{\partial L\left(\boldsymbol{\theta}, \boldsymbol{\theta}_{j-1}\right)}{\partial \theta_i} = \operatorname{tr} \left[\left(\frac{\partial L\left(\boldsymbol{\theta}, \boldsymbol{\theta}_{j-1}\right)}{\partial \mathbf{A}} \right)^T \frac{\partial \mathbf{A}}{\partial \theta_i} \right].
$$

We get

$$\frac{\partial L\left(\boldsymbol{\theta},\boldsymbol{\theta}_{j-1}\right)}{\partial \mathbf{A}} = \mathbf{Q}^{-1}\mathbf{A}\sum_{k=1}^{T}\left\langle \mathbf{x}_k\mathbf{x}_{k-1}^T\right\rangle - \mathbf{Q}^{-1}\sum_{k=1}^{T}\left\langle \mathbf{x}_{k-1}\mathbf{x}_{k-1}^T\right\rangle = 0$$

$$\hat{\mathbf{A}} = \sum_{k=1}^{T}\left\langle \mathbf{x}_k\mathbf{x}_{k-1}^T\right\rangle\left(\sum_{k=1}^{T}\left\langle \mathbf{x}_{k-1}\mathbf{x}_{k-1}^T\right\rangle\right)^{-1}$$

$$\frac{\partial L\left(\boldsymbol{\theta},\boldsymbol{\theta}_{j-1}\right)}{\partial \mathbf{H}} = \mathbf{R}^{-1}\sum_{k=1}^{T}\mathbf{y}_k\left\langle \mathbf{x}_k^T\right\rangle - \mathbf{R}^{-1}\mathbf{H}\sum_{k=1}^{T}\left\langle \mathbf{x}_k\mathbf{x}_k^T\right\rangle = 0$$

$$\hat{\mathbf{H}} = \sum_{k=1}^{T}\left\langle \mathbf{y}_k\mathbf{x}_k^T\right\rangle\left(\sum_{k=1}^{T}\left\langle \mathbf{x}_k\mathbf{x}_k^T\right\rangle\right)^{-1}$$

$$\hat{\mathbf{Q}} = \sum_{k=1}^{T}\left\langle \mathbf{x}_k\mathbf{x}_k^T\right\rangle - \hat{\mathbf{A}}\left(\sum_{k=1}^{T}\left\langle \mathbf{x}_k\mathbf{x}_{k-1}^T\right\rangle\right)^{T}$$

$$\hat{\mathbf{R}} = \sum_{k=1}^{T}\mathbf{y}_k\mathbf{y}_k^T - \hat{\mathbf{H}}\left(\sum_{k=1}^{T}\left\langle \mathbf{y}_k\mathbf{x}_k^T\right\rangle\right)^{T}$$

3 Problem

4 Results