Lecture 2: From Linear Regression to Kalman Filter and Beyond

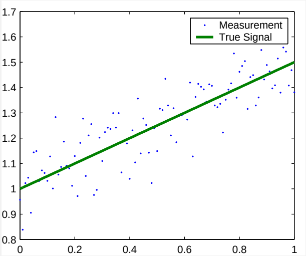
Simo Särkkä

Department of Biomedical Engineering and Computational Science
Aalto University

February 3, 2011

Contents

Batch Linear Regression [1/2]



- Consider the linear regression model

$$y_k = a_1 + a_2\, t_k + \epsilon_k,$$

with $\epsilon_k \sim \mathrm{N}(0, \sigma^2)$ and $\mathbf{a} = (a_1, a_2) \sim \mathrm{N}(\mathbf{m}_0, \mathbf{P}_0)$.

- In probabilistic notation this is:

$$p(y_k \,|\, \mathbf{a}) = \mathrm{N}(y_k \,|\, \mathbf{H}_k\, \mathbf{a}, \sigma^2)$$
$$p(\mathbf{a}) = \mathrm{N}(\mathbf{a} \,|\, \mathbf{m}_0, \mathbf{P}_0),$$

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

Batch Linear Regression [2/2]

- The Bayesian batch solution by the Bayes' rule:

$$p(\mathbf{a} \mid y_{1:N}) \propto p(\mathbf{a}) \prod_{k=1}^{N} p(y_k \mid \mathbf{a})$$
$$= \mathrm{N}(\mathbf{a} \mid \mathbf{m}_0, \mathbf{P}_0) \prod_{k=1}^{N} \mathrm{N}(y_k \mid \mathbf{H}_k \, \mathbf{a}, \sigma^2).$$

- The posterior is Gaussian

$$p(\mathbf{a} \mid y_{1:N}) = \mathrm{N}(\mathbf{a} \mid \mathbf{m}_N, \mathbf{P}_N).$$

- The mean and covariance are given as

$$\mathbf{m}_N = \left[\mathbf{P}_0^{-1} + \frac{1}{\sigma^2}\mathbf{H}^T\mathbf{H}\right]^{-1} \left[\frac{1}{\sigma^2}\mathbf{H}^T\mathbf{y} + \mathbf{P}_0^{-1}\mathbf{m}_0\right]$$
$$\mathbf{P}_N = \left[\mathbf{P}_0^{-1} + \frac{1}{\sigma^2}\mathbf{H}^T\mathbf{H}\right]^{-1},$$

where $\mathbf{H}_k = (1 \;\; t_k)$ and $\mathbf{H} = (\mathbf{H}_1; \mathbf{H}_2; \ldots; \mathbf{H}_N)$, and

Recursive Linear Regression [1/3]

- Assume that we have already computed the posterior distribution, which is conditioned on the measurements up to $k-1$:

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

- Assume that we get the kth measurement y_k. Using the equations from the previous slide we get

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

- The mean and covariance are given as

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

Recursive Linear Regression [2/3]

- By the matrix inversion lemma (or Woodbury identity):

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

- Now the equations for the mean and covariance reduce to

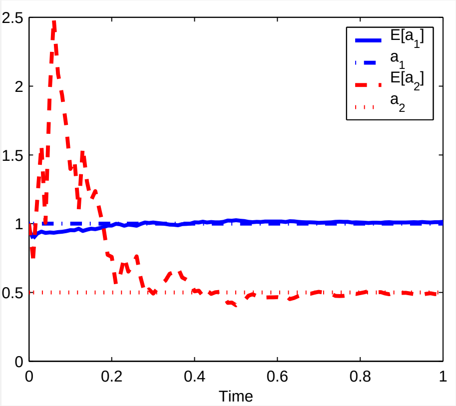
$$\begin{aligned} S_k &= \mathbf{H}_k \mathbf{P}_{k-1} \mathbf{H}_k^T + \sigma^2 \\ \mathbf{K}_k &= \mathbf{P}_{k-1} \mathbf{H}_k^T S_k^{-1} \\ \mathbf{m}_k &= \mathbf{m}_{k-1} + \mathbf{K}_k [y_k - \mathbf{H}_k \mathbf{m}_{k-1}] \\ \mathbf{P}_k &= \mathbf{P}_{k-1} - \mathbf{K}_k S_k \mathbf{K}_k^T.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

- Computing these for $k = 0, \ldots, N$ gives exactly the linear regression solution – but without a matrix inversion[1]!
- A special case of Kalman filter.

[1] Without an *explicit* matrix inversion

Recursive Linear Regression [3/3]

Convergence of the recursive solution to the batch solution – on the last step the solutions are exactly equal:



Batch vs. Recursive Estimation [1/2]

General batch solution:

- Specify the measurement model:

$$p(\mathbf{y}_{1:N} \mid \boldsymbol{\theta}) = \prod_k p(\mathbf{y}_k \mid \boldsymbol{\theta}).$$

- Specify the prior distribution $p(\boldsymbol{\theta})$.
- Compute posterior distribution by the Bayes' rule:

$$p(\boldsymbol{\theta} \mid \mathbf{y}_{1:N}) = \frac{1}{Z} p(\boldsymbol{\theta}) \prod_k p(\mathbf{y}_k \mid \boldsymbol{\theta}).$$

- Compute point estimates, moments, predictive quantities etc. from the posterior distribution.

Batch vs. Recursive Estimation [2/2]

General recursive solution:

- Specify the measurement likelihood $p(\mathbf{y}_k \mid \boldsymbol{\theta})$.
- Specify the prior distribution $p(\boldsymbol{\theta})$.
- Process measurements $\mathbf{y}_1, \ldots, \mathbf{y}_N$ one at a time, starting from the prior:

$$
\begin{aligned}
p(\boldsymbol{\theta} \mid \mathbf{y}_1) &= \frac{1}{Z_1} p(\mathbf{y}_1 \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \\
p(\boldsymbol{\theta} \mid \mathbf{y}_{1:2}) &= \frac{1}{Z_2} p(\mathbf{y}_2 \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{y}_1) \\
&\vdots \\
p(\boldsymbol{\theta} \mid \mathbf{y}_{1:N}) &= \frac{1}{Z_N} p(\mathbf{y}_N \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{y}_{1:N-1}).
\end{aligned}
$$

- The posterior at the last step is the same as the batch solution.

Advantages of Recursive Solution

- The recursive solution can be considered as the online learning solution to the Bayesian learning problem.
- Batch Bayesian inference is a special case of recursive Bayesian inference.
- The parameter can be modeled to change between the measurement steps $\Rightarrow$ basis of filtering theory.

Drift Model for Linear Regression [1/3]

- Let assume Gaussian random walk between the measurements in the linear regression model:

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

- Again, assume that we already know

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

- The joint distribution of $\mathbf{a}_k$ and $\mathbf{a}_{k-1}$ is (due to Markovianity of dynamics!):

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

Drift Model for Linear Regression [2/3]

- Integrating over $\mathbf{a}_{k-1}$ gives:

$$p(\mathbf{a}_k \,|\, y_{1:k-1}) = \int p(\mathbf{a}_k \,|\, \mathbf{a}_{k-1})\, p(\mathbf{a}_{k-1} \,|\, y_{1:k-1})\, d\mathbf{a}_{k-1}.$$

- This equation for Markov processes is called the Chapman-Kolmogorov equation.
- Because the distributions are Gaussian, the result is Gaussian

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

where

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

Drift Model for Linear Regression [3/3]

- As in the pure recursive estimation, we get

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

- After applying the matrix inversion lemma, mean and covariance can be written as

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

- Again, we have derived a special case of the Kalman filter.
- The batch version of this solution would be much more complicated.

State Space Notation

- In the previous section we formulated the model as

$$p(\mathbf{a}_k \,|\, \mathbf{a}_{k-1}) = \mathrm{N}(\mathbf{a}_k \,|\, \mathbf{a}_{k-1}, \mathbf{Q})$$
$$p(y_k \,|\, \mathbf{a}_k) = \mathrm{N}(y_k \,|\, \mathbf{H}_k \, \mathbf{a}_k, \sigma^2)$$

- But in Kalman filtering and control theory the vector of parameters $\mathbf{a}_k$ is usually called "state" and denoted as $\mathbf{x}_k$.
- More standard state space notation:

$$p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}) = \mathrm{N}(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \mathbf{Q})$$
$$p(y_k \,|\, \mathbf{x}_k) = \mathrm{N}(y_k \,|\, \mathbf{H}_k \, \mathbf{x}_k, \sigma^2)$$

- Or equivalently

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{q}$$
$$y_k = \mathbf{H}_k \, \mathbf{x}_k + r,$$

where $\mathbf{q} \sim \mathrm{N}(\mathbf{0}, \mathbf{Q})$, $r \sim \mathrm{N}(0, \sigma^2)$.

Kalman Filter [1/2]

- The canonical Kalman filtering model is

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

- More often, this model can be seen in the form

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

- The Kalman filter actually calculates the following distributions:

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

Kalman Filter [2/2]

- Prediction step of the Kalman filter:

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

- Update step of the Kalman filter:

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

- These equations will be derived from the general Bayesian filtering equations in the next lecture.

Probabilistic Non-Linear Filtering [1/2]

- Generic discrete-time state space models

$$\begin{aligned} \mathbf{x}_k &= \mathbf{f}(\mathbf{x}_{k-1}, \mathbf{q}_{k-1}) \\ \mathbf{y}_k &= \mathbf{h}(\mathbf{x}_k, \mathbf{r}_k). \end{aligned}$$

- Generic Markov models

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

- Approximation methods: Extended Kalman filters (EKF), Unscented Kalman filters (UKF), sequential Monte Carlo (SMC) filters a'ka particle filters.

Probabilistic Non-Linear Filtering [2/2]

- In continuous-discrete filtering models, dynamics are modeled in continuous time, measurements at discrete time steps.
- The continuous time versions of Markov models are called as stochastic differential equations:

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}, t) + \mathbf{w}(t)$$

  where $\mathbf{w}(t)$ is a continuous time Gaussian white noise process.
- Approximation methods: Extended Kalman filters, Unscented Kalman filters, sequential Monte Carlo, particle filters.

Summary

- Linear regression problem can be solved as batch problem or recursively – the latter solution is a special case of Kalman filter.
- A generic Bayesian estimation problem can also be solved as batch problem or recursively.
- If we let the linear regression parameter change between the measurements, we get a simple linear state space model – again solvable with Kalman filtering model.
- By generalizing this idea and the solution we get the Kalman filter algorithm.
- By further generalizing to non-Gaussian models results in a generic probabilistic state space model.

Demonstration

Batch and recursive linear regression.