Lecture 6: Bayesian Inference in SDE Models

Bayesian Filtering and Smoothing Point of View

Simo Särkkä

Aalto University

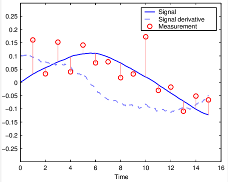
December 2, 2014

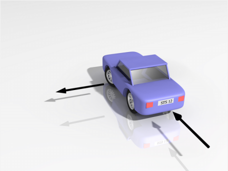
Contents

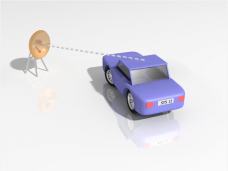
1. Problem Formulation
2. Discrete-Time Bayesian Filtering
3. Discrete-Time Bayesian Smoothing
4. Continuous/Discrete-Time Bayesian Filtering and Smoothing
5. Continuous-Time Bayesian Filtering and Smoothing
6. Related Topics and Summary

The Basic Ideas

- Use SDEs as prior models for the system dynamics.
- We measure the state of the SDE though a measurement model.
- We aim to determine the conditional distribution of the trajectory taken by the SDE given the measurements.
- Because the trajectory is an infinite-dimensional random variable, we do not want to form the full posterior distribution.
- Instead, we target to time-marginals of the posterior – this is the idea of stochastic filtering theory.

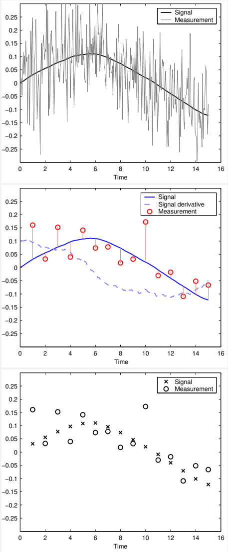




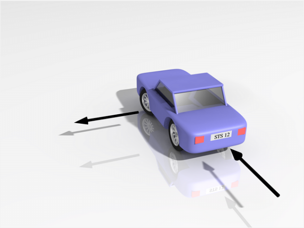


Types of state-estimation problems

- Continuous-time:
  - The dynamics are modeled as continuous-time processes (SDEs).
  - The measurements are modeled as continuous-time processes (SDEs).
- Continuous/discrete-time:
  - The dynamics are modeled as continuous-time processes (SDEs).
  - The measurements are modeled as discrete-time processes.
- Discrete-time:
  - The dynamics are modeled as discrete-time processes.
  - The measurements are modeled as discrete-time processes.



Example: State Space Model for a Car [1/2]



- The dynamics of the car in 2d (x_1, x_2) are given by Newton's law:

$$\mathbf{F}(t) = m\,\mathbf{a}(t),$$

where $\mathbf{a}(t)$ is the acceleration, m is the mass of the car, and $\mathbf{F}(t)$ is a vector of (unknown) forces.

- Let's model $\mathbf{F}(t)/m$ as a 2-dimensional white noise process:

$$\begin{aligned} \mathrm{d}^2 x_1 / \mathrm{d}t^2 &= w_1(t) \\ \mathrm{d}^2 x_2 / \mathrm{d}t^2 &= w_2(t). \end{aligned}$$

Example: State Space Model for a Car [2/2]

- If we define $x_3(t) = dx_1/\mathrm{d}t$, $x_4(t) = dx_2/\mathrm{d}t$, then the model can be written as a first order system of differential equations:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}}_{\mathbf{F}} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} + \underbrace{\begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}}_{\mathbf{L}} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}.$$

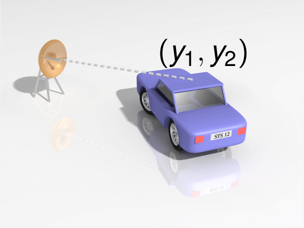
- In shorter matrix form:

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{F}\,\mathbf{x} + \mathbf{L}\,\mathbf{w}.$$

- More rigorously:

$$\mathrm{d}\mathbf{x} = \mathbf{F}\,\mathbf{x}\;\mathrm{d}t + \mathbf{L}\;\mathrm{d}\boldsymbol{\beta}.$$

Continuous-Time Measurement Model for a Car



- Assume that the position of the car (x_1, x_2) is measured and the measurements are corrupted by white noise $e_1(t), e_2(t)$:

$$y_1(t) = x_1(t) + e_1(t)$$
$$y_2(t) = x_2(t) + e_2(t).$$

- The measurement model can be now written as

$$\mathbf{y}(t) = \mathbf{H}\,\mathbf{x}(t) + \mathbf{e}(t), \text{ with } \mathbf{H} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

- Or more rigorously as an SDE

$$\mathrm{d}\mathbf{z} = \mathbf{H}\,\mathbf{x}\;\mathrm{d}t + \mathrm{d}\boldsymbol{\eta}.$$

General Continuous-Time State-Space Models [1/2]

- The resulting model is of the form

$$
\begin{aligned}
\mathrm{d}\mathbf{x} &= \mathbf{F}\,\mathbf{x}\;\mathrm{d}t + \mathbf{L}\;\mathrm{d}\boldsymbol{\beta} \\
\mathrm{d}\mathbf{z} &= \mathbf{H}\,\mathbf{x}\;\mathrm{d}t + \mathrm{d}\boldsymbol{\eta}.
\end{aligned}
$$

- This is a special case of a continuous-time model:

$$
\begin{aligned}
\mathrm{d}\mathbf{x} &= \mathbf{f}(\mathbf{x}, t)\;\mathrm{d}t + \mathbf{L}(\mathbf{x}, t)\;\mathrm{d}\boldsymbol{\beta} \\
\mathrm{d}\mathbf{z} &= \mathbf{h}(\mathbf{x}, t)\;\mathrm{d}t + \mathrm{d}\boldsymbol{\eta}.
\end{aligned}
$$

- The first equation defines dynamics of the system state and the second relates measurements to the state.
- Given that we have observed $\mathbf{z}(t)$ (or $\mathbf{y}(t)$), what can we say (in statistical sense) about the hidden process $\mathbf{x}(t)$?

General Continuous-Time State-Space Models [2/2]

- Bayesian way: what is the *posterior distribution* of $\mathbf{x}(t)$ given the noisy measurements $\mathbf{y}(\tau)$ on $\tau \in [0, T]$?
- This Bayesian solution is surpricingly old, as it dates back to work of Stratonovich around 1950s.
- The aim is usually to compute the filtering (posterior) distribution

$$p(\mathbf{x}(t) \mid \{\mathbf{y}(\tau) \; : \; 0 \leq \tau \leq t\}).$$

- We are also often interested in the smoothing distributions

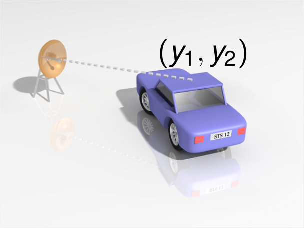
$$p(\mathbf{x}(t^*) \mid \{\mathbf{y}(\tau) \; : \; 0 \leq \tau \leq T\}) \qquad t^* \in [0, T].$$

- Note that we could also attempt to compute the "full" posterior

$$p(\{\mathbf{x}(t^*) \; : \; 0 \leq t^* \leq T\} \mid \{\mathbf{y}(\tau) \; : \; 0 \leq \tau \leq T\}).$$

- The full posterior is not usually feasible nor sensible – we will return to this later.

Discrete-Time Measurement Model for a Car



- Assume that the position of the car (x_1, x_2) is measured at discrete time instants $t_1, t_2, \ldots, t_n$:

$$
\begin{aligned}
y_{1,k} &= x_1(t_k) + e_{1,k} \\
y_{2,k} &= x_2(t_k) + e_{2,k},
\end{aligned}
$$

$(e_{1,k}, e_{2,k}) \sim \mathrm{N}(\mathbf{0}, \mathbf{R})$ are Gaussian.

- The measurement model can be now written as

$$
\mathbf{y}_k = \mathbf{H}\,\mathbf{x}(t_k) + \mathbf{e}_k, \qquad \mathbf{H} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}
$$

- Or in probabilistic notation as

$$
p(\mathbf{y}_k \,|\, \mathbf{x}(t_k)) = \mathrm{N}(\mathbf{y}_k \,|\, \mathbf{H}\,\mathbf{x}(t_k), \mathbf{R}).
$$

General Continuous/Discrete-Time State-Space Models

- The dynamic and measurement models now have the form:

$$\begin{aligned}
\mathrm{d}\mathbf{x} &= \mathbf{F}\,\mathbf{x}\,\mathrm{d}t + \mathbf{L}\,\mathrm{d}\beta \\
\mathbf{y}_k &= \mathbf{H}\,\mathbf{x}(t_k) + \mathbf{r}_k,
\end{aligned}$$

- Special case of continuous/discrete-time models of the form

$$\begin{aligned}
\mathrm{d}\mathbf{x} &= \mathbf{f}(\mathbf{x}, t)\,\mathrm{d}t + \mathbf{L}(\mathbf{x}, t)\,\mathrm{d}\beta \\
\mathbf{y}_k &\sim p(\mathbf{y}_k \mid \mathbf{x}(t_k)).
\end{aligned}$$

- We are typically interested in the filtering and smoothing (posterior) distributions

$$\begin{aligned}
&p(\mathbf{x}(t_k) \mid \mathbf{y}_1, \ldots, \mathbf{y}_k), \\
&p(\mathbf{x}(t^*) \mid \mathbf{y}_1, \ldots, \mathbf{y}_T), \qquad t^* \in [0, t_T].
\end{aligned}$$

- In principle, the full posterior can also be considered – but we will concetrate on the above.

General Discrete-Time State-Space Models [1/2]

- Recall that the solution to the SDE $\mathrm{d}\mathbf{x} = \mathbf{F}\,\mathbf{x}\,\mathrm{d}t + \mathbf{L}\,\mathrm{d}\beta$ is

$$\mathbf{x}(t) = \exp(\mathbf{F}\,(t - t_0))\,\mathbf{x}(t_0) + \int_{t_0}^{t} \exp(\mathbf{F}\,(t - \tau))\,\mathbf{L}\,\mathrm{d}\beta(\tau).$$

- If we set $t \leftarrow t_k$ and $t_0 \leftarrow t_{k-1}$ we get

$$\mathbf{x}(t_k) = \exp(\mathbf{F}\,(t_k - t_{k-1}))\,\mathbf{x}(t_{k-1}) + \int_{t_{k-1}}^{t_k} \exp(\mathbf{F}\,(t - \tau))\,\mathbf{L}\,\mathrm{d}\beta(\tau).$$

- Thus this is of the form

$$\mathbf{x}(t_k) = \mathbf{A}_{k-1}\,\mathbf{x}(t_{k-1}) + \mathbf{q}_{k-1}$$

  where
  - $\mathbf{A}_{k-1} = \exp(\mathbf{F}\,(t_k - t_{k-1}))$ is a given (deterministic) matrix and
  - $\mathbf{q}_{k-1}$ is zero-mean Gaussian random variable with covariance $\mathbf{Q}_{k-1} = \int_{t_{k-1}}^{t_k} \exp(\mathbf{F}\,(t - \tau))\,\mathbf{L}\,\mathbf{Q}\,\mathbf{L}^\mathsf{T} \exp(\mathbf{F}\,(t - \tau))^\mathsf{T}\,\mathrm{d}\tau$.

General Discrete-Time State-Space Models [2/2]

- Thus we can write the linear state-space model (e.g. the car) equivalently in form such as

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

- This is a special case of discrete-time models of the form

$$
\begin{aligned}
\mathbf{x}(t_k) &\sim p(\mathbf{x}(t_k) \mid \mathbf{x}(t_{k-1})) \\
\mathbf{y}_k &\sim p(\mathbf{y}_k \mid \mathbf{x}(t_k)).
\end{aligned}
$$

- Generally $p(\mathbf{x}(t_k) \mid \mathbf{x}(t_{k-1}))$ is the transition density of the SDE (The Green's function of Fokker–Planck–Kolmogorov)
- We are typically interested in the filtering/smoothing distributions

$$
\begin{aligned}
& p(\mathbf{x}(t_k) \mid \mathbf{y}_1, \ldots, \mathbf{y}_k), \\
& p(\mathbf{x}(t_i) \mid \mathbf{y}_1, \ldots, \mathbf{y}_T), \qquad i = 1, 2, \ldots, T.
\end{aligned}
$$

- Sometimes we can also do the full posterior...

Why Not The Full Posterior?

- Consider a discrete-time state-space model:

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

- Due to Markovianity, the joint prior is now given as

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

- Due to conditional independence of measurements, the joint likelihood is given as

$$
p(\mathbf{y}_{1:T} \mid \mathbf{x}_{0:T}) = \prod_{k=1}^{T} p(\mathbf{y}_k \mid \mathbf{x}_k).
$$

Why Not The Full Posterior? (cont.)

- We can now use Bayes' rule to compute the full posterior

$$
\begin{aligned}
p(\mathbf{x}_{0:T} \mid \mathbf{y}_{1:T}) &= \frac{p(\mathbf{y}_{1:T} \mid \mathbf{x}_{0:T})\, p(\mathbf{x}_{0:T})}{p(\mathbf{y}_{1:T})} \\
&= \frac{\prod_{k=1}^{T} p(\mathbf{y}_k \mid \mathbf{x}_k)\, p(\mathbf{x}_k \mid \mathbf{x}_{k-1})\, p(\mathbf{x}_0)}{\int \prod_{k=1}^{T} p(\mathbf{y}_k \mid \mathbf{x}_k)\, p(\mathbf{x}_k \mid \mathbf{x}_{k-1})\, p(\mathbf{x}_0)\, \mathrm{d}\mathbf{x}_{0:T}}
\end{aligned}
$$

- This is very high dimensional (with SDEs infinite) and hence inefficient to work with – this is why filtering theory was invented.
- We aim to fully utilize the Markovian structure of the model to efficiently compute the following partial posteriors:
  - Filtering distributions

$$p(\mathbf{x}_k \mid \mathbf{y}_{1:k}), \qquad k = 1, \ldots, T.$$

  - Smoothing distributions

$$p(\mathbf{x}_k \mid \mathbf{y}_{1:T}), \qquad k = 1, \ldots, T.$$

Bayesian Optimal Filter: Principle

- Assume that we have been given:
  1. Prior distribution $p(\mathbf{x}_0)$.
  2. State space model:

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

  3. Measurement sequence $\mathbf{y}_{1:k} = \mathbf{y}_1, \ldots, \mathbf{y}_k$.
- We usually have $\mathbf{x}_k \triangleq \mathbf{x}(t_k)$ for some times $t_1, t_2, \ldots$.
- Bayesian optimal filter computes the distribution

$$
p(\mathbf{x}_k \mid \mathbf{y}_{1:k})
$$

- Computation is based on recursion rule for incorporation of the new measurement $\mathbf{y}_k$ into the posterior:

$$
p(\mathbf{x}_{k-1} \mid \mathbf{y}_{1:k-1}) \longrightarrow p(\mathbf{x}_k \mid \mathbf{y}_{1:k})
$$

Bayesian Optimal Filter: Derivation of Prediction Step

- Assume that we know the posterior distribution of previous time step:

$$p(\mathbf{x}_{k-1} \mid \mathbf{y}_{1:k-1}).$$

- The joint distribution of $\mathbf{x}_k$, $\mathbf{x}_{k-1}$ given $\mathbf{y}_{1:k-1}$ can be computed as (recall the Markov property):

$$\begin{aligned} p(\mathbf{x}_k, \mathbf{x}_{k-1} \mid \mathbf{y}_{1:k-1}) &= p(\mathbf{x}_k \mid \mathbf{x}_{k-1}, \mathbf{y}_{1:k-1})\, p(\mathbf{x}_{k-1} \mid \mathbf{y}_{1:k-1}) \\ &= p(\mathbf{x}_k \mid \mathbf{x}_{k-1})\, p(\mathbf{x}_{k-1} \mid \mathbf{y}_{1:k-1}), \end{aligned}$$

- Integrating over $\mathbf{x}_{k-1}$ gives the Chapman-Kolmogorov equation

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

- This is the prediction step of the optimal filter.

Bayesian Optimal Filter: Derivation of Update Step

- Now we have:
  1. Prior distribution from the Chapman-Kolmogorov equation

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

  2. Measurement likelihood from the state space model:

$$p(\mathbf{y}_k \,|\, \mathbf{x}_k)$$

- The posterior distribution can be computed by the Bayes' rule (recall the conditional independence of measurements):

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

- This is the update step of the optimal filter.

Bayesian Optimal Filter: Formal Equations

Optimal filter

- Initialization: The recursion starts from the prior distribution $p(\mathbf{x}_0)$.
- Prediction: by the Chapman-Kolmogorov equation

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

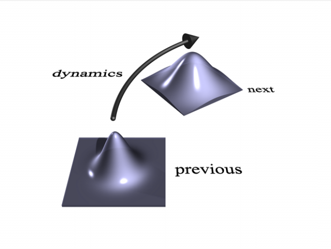
- Update: by the Bayes' rule

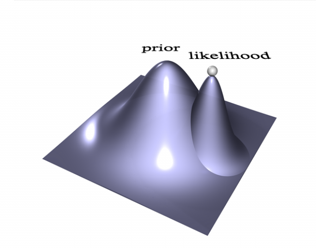
$$p(\mathbf{x}_k \mid \mathbf{y}_{1:k}) = \frac{1}{Z_k} p(\mathbf{y}_k \mid \mathbf{x}_k) \, p(\mathbf{x}_k \mid \mathbf{y}_{1:k-1}).$$

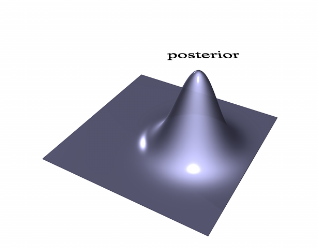
- The normalization constant $Z_k = p(\mathbf{y}_k \mid \mathbf{y}_{1:k-1})$ is given as

$$Z_k = \int p(\mathbf{y}_k \mid \mathbf{x}_k) \, p(\mathbf{x}_k \mid \mathbf{y}_{1:k-1}) \, \mathrm{d}\mathbf{x}_k.$$

Bayesian Optimal Filter: Graphical Explanation







On prediction step the distribution of previous step is propagated through the dynamics.

Prior distribution from prediction and the likelihood of measurement.

The posterior distribution after combining the prior and likelihood by Bayes' rule.

Filtering Algorithms

- Kalman filter is the classical optimal filter for linear-Gaussian models.
- Extended Kalman filter (EKF) is linearization based extension of Kalman filter to non-linear models.
- Unscented Kalman filter (UKF) is sigma-point transformation based extension of Kalman filter.
- Gauss-Hermite and Cubature Kalman filters (GHKF/CKF) are numerical integration based extensions of Kalman filter.
- Particle filter forms a Monte Carlo representation (particle set) to the distribution of the state estimate.
- Grid based filters approximate the probability distributions on a finite grid.
- Mixture Gaussian approximations are used, for example, in multiple model Kalman filters and Rao-Blackwellized Particle filters.

Kalman Filter: Model

- Gaussian driven linear model, i.e., Gauss-Markov model:

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

- $\mathbf{q}_{k-1} \sim \mathrm{N}(\mathbf{0}, \mathbf{Q}_{k-1})$ white process noise.
- $\mathbf{r}_k \sim \mathrm{N}(\mathbf{0}, \mathbf{R}_k)$ white measurement noise.
- $\mathbf{A}_{k-1}$ is the transition matrix of the dynamic model.
- $\mathbf{H}_k$ is the measurement model matrix.
- In probabilistic terms the model is

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

Kalman Filter: Equations

Kalman Filter

- Initialization: $\mathbf{x}_0 \sim \mathrm{N}(\mathbf{m}_0, \mathbf{P}_0)$
- Prediction step:

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

- Update step:

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

Problem Formulation

- Probabilistic state space model:

$$\text{measurement model: } \mathbf{y}_k \sim p(\mathbf{y}_k \,|\, \mathbf{x}_k)$$
$$\text{dynamic model: } \mathbf{x}_k \sim p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1})$$

- Assume that the filtering distributions $p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k})$ have already been computed for all $k = 0, \ldots, T$.
- We want recursive equations of computing the smoothing distribution for all $k < T$:

$$p(\mathbf{x}_k \,|\, \mathbf{y}_{1:T}).$$

- The recursion will go backwards in time, because on the last step, the filtering and smoothing distributions coincide:

$$p(\mathbf{x}_T \,|\, \mathbf{y}_{1:T}).$$

Derivation of Formal Smoothing Equations [1/2]

- The key: due to the Markov properties of state we have:

$$p(\mathbf{x}_k \mid \mathbf{x}_{k+1}, \mathbf{y}_{1:T}) = p(\mathbf{x}_k \mid \mathbf{x}_{k+1}, \mathbf{y}_{1:k})$$

- Thus we get:

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

Derivation of Formal Smoothing Equations [2/2]

- Assuming that the smoothing distribution of the next step $p(\mathbf{x}_{k+1} \,|\, \mathbf{y}_{1:T})$ is available, we get

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

- Integrating over $\mathbf{x}_{k+1}$ gives

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

Bayesian Optimal Smoothing Equations

Bayesian Optimal Smoothing Equations

The Bayesian optimal smoothing equations consist of prediction step and backward update step:

$$p(\mathbf{x}_{k+1} \mid \mathbf{y}_{1:k}) = \int p(\mathbf{x}_{k+1} \mid \mathbf{x}_k)\, p(\mathbf{x}_k \mid \mathbf{y}_{1:k})\, \mathrm{d}\mathbf{x}_k$$

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

The recursion is started from the filtering (and smoothing) distribution of the last time step $p(\mathbf{x}_T \mid \mathbf{y}_{1:T})$.

Smoothing Algorithms

- Rauch-Tung-Striebel (RTS) smoother is the closed form smoother for linear Gaussian models.
- Extended, statistically linearized and unscented RTS smoothers are the approximate nonlinear smoothers corresponding to EKF, SLF and UKF.
- Gaussian RTS smoothers: cubature RTS smoother, Gauss-Hermite RTS smoothers and various others
- Particle smoothing is based on approximating the smoothing solutions via Monte Carlo.
- Rao-Blackwellized particle smoother is a combination of particle smoothing and RTS smoothing.

Linear-Gaussian Smoothing Problem

- Gaussian driven linear model, i.e., Gauss-Markov model:

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

- In probabilistic terms the model is

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

- Kalman filter can be used for computing all the Gaussian filtering distributions:

$$p(\mathbf{x}_k \,|\, \mathbf{y}_{1:k}) = \mathrm{N}(\mathbf{x}_k \,|\, \mathbf{m}_k, \mathbf{P}_k).$$

- Rauch–Tung–Striebel smoother then computes the corresponding smoothing distributions

$$p(\mathbf{x}_k \,|\, \mathbf{y}_{1:T}) = \mathrm{N}(\mathbf{x}_k \,|\, \mathbf{m}_k^s, \mathbf{P}_k^s).$$

Rauch-Tung-Striebel Smoother

Rauch-Tung-Striebel Smoother

Backward recursion equations for the smoothed means $\mathbf{m}_k^s$ and covariances $\mathbf{P}_k^s$:

$$
\begin{aligned}
\mathbf{m}_{k+1}^- &= \mathbf{A}_k \, \mathbf{m}_k \\
\mathbf{P}_{k+1}^- &= \mathbf{A}_k \, \mathbf{P}_k \, \mathbf{A}_k^\mathsf{T} + \mathbf{Q}_k \\
\mathbf{G}_k &= \mathbf{P}_k \, \mathbf{A}_k^\mathsf{T} \, [\mathbf{P}_{k+1}^-]^{-1} \\
\mathbf{m}_k^s &= \mathbf{m}_k + \mathbf{G}_k \, [\mathbf{m}_{k+1}^s - \mathbf{m}_{k+1}^-] \\
\mathbf{P}_k^s &= \mathbf{P}_k + \mathbf{G}_k \, [\mathbf{P}_{k+1}^s - \mathbf{P}_{k+1}^-] \, \mathbf{G}_k^\mathsf{T},
\end{aligned}
$$

- $\mathbf{m}_k$ and $\mathbf{P}_k$ are the mean and covariance from Kalman filter.
- The recursion is started from the last time step T, with $\mathbf{m}_T^s = \mathbf{m}_T$ and $\mathbf{P}_T^s = \mathbf{P}_T$.

Continuous/Discrete-Time Bayesian Filtering and Smoothing: Method A

- Consider a continuous-discrete state-space model

$$\begin{aligned} \mathrm{d}\mathbf{x} &= \mathbf{f}(\mathbf{x}, t)\,\mathrm{d}t + \mathbf{L}(\mathbf{x}, t)\,\mathrm{d}\beta \\ \mathbf{y}_k &\sim p(\mathbf{y}_k \mid \mathbf{x}(t_k)). \end{aligned}$$

- We can always convert this into an equivalent discrete-time model

$$\begin{aligned} \mathbf{x}(t_k) &\sim p(\mathbf{x}(t_k) \mid \mathbf{x}(t_{k-1})) \\ \mathbf{y}_k &\sim p(\mathbf{y}_k \mid \mathbf{x}(t_k)). \end{aligned}$$

  by solving the transition density $p(\mathbf{x}(t_k) \mid \mathbf{x}(t_{k-1}))$.
- Then we can simply use the discrete-time filtering and smoothing algorithms.
- With linear SDEs we can discretize exactly; with non-linear SDEs we can use e.g. Itô-Taylor expansions.

Continuous/Discrete-Time Bayesian Filtering and Smoothing: Method B

- Another way is to replace the discrete-time prediction step

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

  with its continuous-time counterpart.
- Generally, we get the Fokker-Planck equation

$$\begin{aligned} \frac{\partial p(\mathbf{x}, t)}{\partial t} &= -\sum_i \frac{\partial}{\partial x_i}[f_i(x, t)\, p(\mathbf{x}, t)] \\ &\quad + \frac{1}{2} \sum_{ij} \frac{\partial^2}{\partial x_i\, \partial x_j} \left\{ [\mathbf{L}(\mathbf{x}, t)\, \mathbf{Q}\, \mathbf{L}^\mathsf{T}(\mathbf{x}, t)]_{ij}\, p(\mathbf{x}, t) \right\}. \end{aligned}$$

  with the initial condition $p(\mathbf{x}, t_{k-1}) \triangleq p(\mathbf{x}_{k-1} \mid \mathbf{y}_{1:k-1})$.

Continuous/Discrete-Time Bayesian Filtering and Smoothing: Method B (cont.)

Continuous-Discrete Bayesian Optimal filter

1. Prediction step: Solve the Fokker-Planck-Kolmogorov PDE

$$\frac{\partial p}{\partial t} = -\sum_i \frac{\partial}{\partial x_i} (f_i \, p) + \frac{1}{2} \sum_{ij} \frac{\partial^2}{\partial x_i \partial x_j} \left([\mathbf{L} \, \mathbf{Q} \, \mathbf{L}^\mathsf{T}]_{ij} \, p \right)$$

2. Update step: Apply the Bayes' rule.

$$p(\mathbf{x}(t_k) \mid \mathbf{y}_{1:k}) = \frac{p(\mathbf{y}_k \mid \mathbf{x}(t_k)) \, p(\mathbf{x}(t_k) \mid \mathbf{y}_{1:k-1})}{\int p(\mathbf{y}_k \mid \mathbf{x}(t_k)) \, p(\mathbf{x}(t_k) \mid \mathbf{y}_{1:k-1}) \, \mathrm{d}\mathbf{x}(t_k)}$$

- In linear models we can use the mean and covariance equations.
- Approximate mean/covariance equations are used in EKF/UKF/...
- The smoother consists of partial/ordinary differential equations.

Continuous-Time Stochastic Filtering Theory [3/3]

- Continuous-time state-space model

$$
\begin{aligned}
\mathrm{d}\mathbf{x} &= \mathbf{f}(\mathbf{x}, t)\,\mathrm{d}t + \mathbf{L}(\mathbf{x}, t)\,\mathrm{d}\boldsymbol{\beta} \\
\mathrm{d}\mathbf{z} &= \mathbf{h}(\mathbf{x}, t)\,\mathrm{d}t + \mathrm{d}\boldsymbol{\eta}.
\end{aligned}
$$

- To ease notation, let's define a linear operator $\mathscr{A}^*$:

$$
\begin{aligned}
\mathscr{A}^*(\bullet) = &-\sum_i \frac{\partial}{\partial x_i}[f_i(x, t)\,(\bullet)] \\
&+ \frac{1}{2}\sum_{ij}\frac{\partial^2}{\partial x_i\,\partial x_j}\{[\mathbf{L}(\mathbf{x}, t)\,\mathbf{Q}\,\mathbf{L}^\mathsf{T}(\mathbf{x}, t)]_{ij}\,(\bullet)\}.
\end{aligned}
$$

- The Fokker–Planck–Kolmogorov equation can then be written compactly as

$$
\frac{\partial p}{\partial t} = \mathscr{A}^* p.
$$

Kushner–Stratonovich equation

By taking the continuous-time limit of the discrete-time Bayesian filtering equations we get the following:

Kushner–Stratonovich equation

The stochastic partial differential equation for the filtering density $p(\mathbf{x}, t \mid \mathscr{Y}_t) \triangleq p(\mathbf{x}(t) \mid \mathscr{Y}_t)$ is

$$\begin{aligned} \mathrm{d}p(\mathbf{x}, t \mid \mathscr{Y}_t) &= \mathscr{A}^* \, p(\mathbf{x}, t \mid \mathscr{Y}_t) \, \mathrm{d}t \\ &+ (\mathbf{h}(\mathbf{x}, t) - \mathrm{E}[\mathbf{h}(\mathbf{x}, t) \mid \mathscr{Y}_t])^\mathsf{T} \, \mathbf{R}^{-1} (\mathrm{d}\mathbf{z} - \mathrm{E}[\mathbf{h}(\mathbf{x}, t) \mid \mathscr{Y}_t] \, \mathrm{d}t) \, p(\mathbf{x}, t \mid \mathscr{Y}_t), \end{aligned}$$

where $\mathrm{d}p(\mathbf{x}, t \mid \mathscr{Y}_t) = p(\mathbf{x}, t + \mathrm{d}t \mid \mathscr{Y}_{t+\mathrm{d}t}) - p(\mathbf{x}, t \mid \mathscr{Y}_t)$ and

$$\mathrm{E}[\mathbf{h}(\mathbf{x}, t) \mid \mathscr{Y}_t] = \int \mathbf{h}(\mathbf{x}, t) \, p(\mathbf{x}, t \mid \mathscr{Y}_t) \, \mathrm{d}\mathbf{x}.$$

Zakai equation

We can get rid of the non-linearity by using an unnormalized equation:

Zakai equation

Let $q(\mathbf{x}, t \mid \mathscr{Y}_t) \triangleq q(\mathbf{x}(t) \mid \mathscr{Y}_t)$ be the solution to Zakai's stochastic partial differential equation

$$\mathrm{d}q(\mathbf{x}, t \mid \mathscr{Y}_t) = \mathscr{A}^* \, q(\mathbf{x}, t \mid \mathscr{Y}_t) \, \mathrm{d}t + \mathbf{h}^\mathsf{T}(\mathbf{x}, t) \, \mathbf{R}^{-1} \, \mathrm{d}\mathbf{z} \, q(\mathbf{x}, t \mid \mathscr{Y}_t),$$

where $\mathrm{d}q(\mathbf{x}, t \mid \mathscr{Y}_t) = q(\mathbf{x}, t + \mathrm{d}t \mid \mathscr{Y}_{t+\mathrm{d}t}) - q(\mathbf{x}, t \mid \mathscr{Y}_t)$. Then we have

$$p(\mathbf{x}(t) \mid \mathscr{Y}_t) = \frac{q(\mathbf{x}(t) \mid \mathscr{Y}_t)}{\int q(\mathbf{x}(t) \mid \mathscr{Y}_t) \, \mathrm{d}\mathbf{x}(t)}.$$

Kalman–Bucy filter

The Kalman–Bucy filter is the exact solution to the linear Gaussian filtering problem

$$
\begin{aligned}
\mathrm{d}\mathbf{x} &= \mathbf{F}(t)\,\mathbf{x}\;\mathrm{d}t + \mathbf{L}(t)\;\mathrm{d}\boldsymbol{\beta} \\
\mathrm{d}\mathbf{z} &= \mathbf{H}(t)\,\mathbf{x}\;\mathrm{d}t + \mathrm{d}\boldsymbol{\eta}.
\end{aligned}
$$

Kalman–Bucy filter

The Bayesian filter, which computes the posterior distribution $p(\mathbf{x}(t) \mid \mathscr{Y}_t) = \mathrm{N}(\mathbf{x}(t) \mid \mathbf{m}(t), \mathbf{P}(t))$ for the above system is

$$
\begin{aligned}
\mathbf{K}(t) &= \mathbf{P}(t)\,\mathbf{H}^{\mathsf{T}}(t)\,\mathbf{R}^{-1} \\
\mathrm{d}\mathbf{m}(t) &= \mathbf{F}(t)\,\mathbf{m}(t)\;\mathrm{d}t + \mathbf{K}(t)\;[\mathrm{d}\mathbf{z}(t) - \mathbf{H}(t)\,\mathbf{m}(t)\;\mathrm{d}t] \\
\frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} &= \mathbf{F}(t)\,\mathbf{P}(t) + \mathbf{P}(t)\,\mathbf{F}^{\mathsf{T}}(t) + \mathbf{L}(t)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(t) - \mathbf{K}(t)\,\mathbf{R}\,\mathbf{K}^{\mathsf{T}}(t).
\end{aligned}
$$

Related Topics

- We can also estimate parameters θ in SDEs/state-spate models:

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t; \theta)\, \mathrm{d}t + \mathbf{L}(\mathbf{x}, t; \theta)\, \mathrm{d}\beta$$

- The filtering theory provides the means to compute the required marginal likelihoods and parameter posteriors.
- It is also possible estimate $\mathbf{f}(\mathbf{x}, t)$ non-parametrically, that is, using Gaussian process (GP) regression.
- Model selection, Bayesian model averaging, and other advanced concepts can also be combined with state-space inference.
- Stochastic control theory is related to optimal control design for SDE models.
- GP-regression can also be sometimes converted to inference on SDE models.

Summary

- We can use SDEs to model dynamics in Bayesian models.
- Dynamic (state-) estimation problems can be divided into continuous-time, continuous/discrete-time, and discrete-time problems – the continuous models are SDEs.
- The full posterior of state trajectory is usually intractable – therefore we compute filtering and smoothing distributions:

$$\begin{aligned}
&p(\mathbf{x}(t_k) \mid \mathbf{y}_1, \ldots, \mathbf{y}_k), \\
&p(\mathbf{x}(t^*) \mid \mathbf{y}_1, \ldots, \mathbf{y}_T), \qquad t^* \in [0, t_T].
\end{aligned}$$

- The Bayesian filtering and smoothing equations also often need to be approximated.
- Methods: Kalman filters, extended Kalman filters (EKF/UKF/...), particle filters – and the related smoothers.