Lecture 2: Filtering and Smoothing in Dynamical Systems

Cédric Archambeau

Centre for Computational Statistics and Machine Learning
Department of Computer Science
University College London

c.archambeau@cs.ucl.ac.uk

Advanced Topics in Machine Learning (MSc in Intelligent Systems)
January 2008

Today's plan

- (Hidden) Markov models
- Optimal Bayesian filtering and smoothing
- Linear state space models (Kalman filter/smoother)
- Nonlinear state space models: deterministic versus statistical linearisation (Extended Kalman filter/smoother, Sigma point filters/smoothers.)
- Applications of particle filters

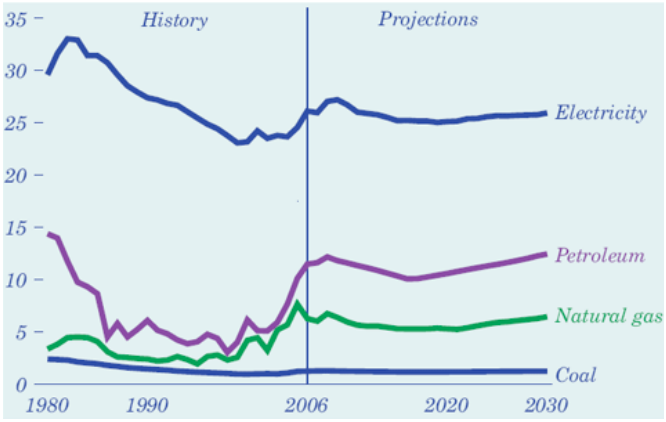
- Guest speakers: **Frank Wood** (Gatsby unit)
  Simon Julier (CS) on 05/02

Sequential data

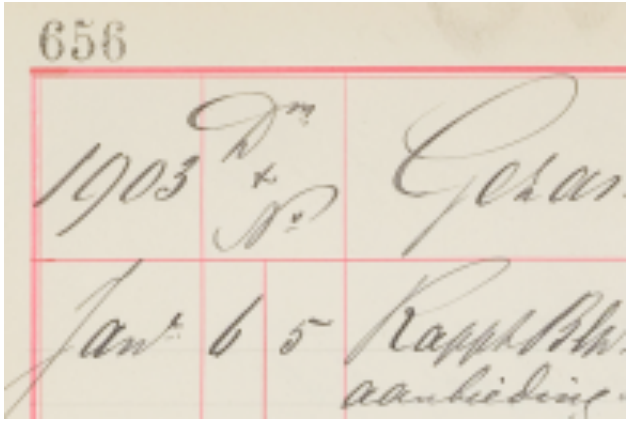
We relax the iid assumption of the data, such that the **likelihood is structured**:

$$p(\mathbf{t}_1, \ldots, \mathbf{t}_N) = \prod_{n=1}^{N} \underbrace{p(\mathbf{t}_n | \mathbf{t}_{n-1} \ldots, \mathbf{t}_1)}_{\textit{Conditionals!}}.$$

- Time series: speech, video, energy consumption, finance, ...
- Spacial sequences: images, DNA sequences, handwritten digits, ...



(a) Energy price.



(b) Text.

Figure: Examples of sequential data.

Markov models

A sequence of state variables $\{\mathbf{t}_n\}_{n>0}$ is an M^{th} order **Markov chain** if, for all n, state $\mathbf{t}_n$ depends only on the values taken by the M previous state variables.

The **Markov property** tells us that $\mathbf{t}_n$ (and in fact the future $\mathbf{t}_{n+1}, \mathbf{t}_{n+2}, \ldots$) is independent of the past when conditioning on $\{\mathbf{t}_{n-1} \ldots, \mathbf{t}_{n-M}\}$.

From the Markov property, the joint distribution of the observed **state variables** up to time τ_N has the following form:

$$P(\mathbf{t}_1, \ldots, \mathbf{t}_N) = P(\mathbf{t}_1)P(\mathbf{t}_2|\mathbf{t}_1) \ldots P(\mathbf{t}_M|\mathbf{t}_{M-1}, \ldots, \mathbf{t}_1) \times \prod_{n=M+1}^{N} P(\mathbf{t}_n|\mathbf{t}_{n-1} \ldots, \mathbf{t}_{n-M}).$$

The Markov model is defined in terms of **transition probabilities** between **discrete states**.

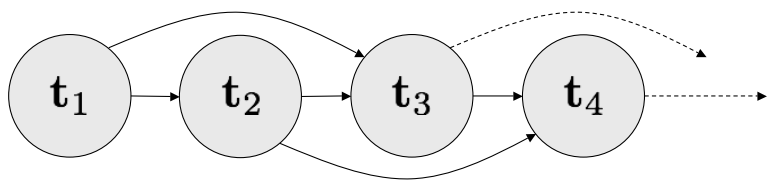


Figure: Graphical model of a 2^{nd} order Markov chain.

Hidden Markov models (HMM)

A more flexible approach is to assume that the observations are noisy realisations of **latent state variables**, which follow a Markov model.

The joint distribution for a (1^{st} order) **hidden Markov model** is given by

$$P(\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{t}_1, \ldots, \mathbf{t}_N) = P(\mathbf{y}_1) \prod_{n=2}^{N} P(\mathbf{y}_n|\mathbf{y}_{n-1}) \prod_{n'=1}^{N'} P(\mathbf{t}_{n'}|\mathbf{y}_{n'}),$$

where $\{\mathbf{t}_n\}_{n>0}$ can be real-valued.

We are not only interested in the conditionals (for making predictions), but also in the **latent causes**.

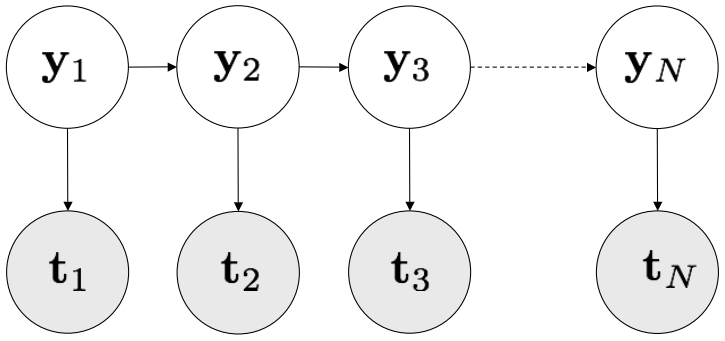


Figure: Graphical model of a 1^{st} order hidden Markov model.

State space models (SMM)

The aim is to model **time varying systems**, which are **indirectly observed** through noisy measurements:

- A set of **dynamic variables** (e.g. position, velocity, acceleration, orientation, etc.) describe the physical state of the system at any time.
- To be realistic, both the time evolution of the system state and the measurements are considered to be **uncertain**.

State space models (and optimal filtering) are commonly used in a wide range of applications:

- Navigation (e.g. GPS)
- (Aero)space engineering
- Remote surveillance
- Telecommunications and signal processing
- Control engineering
- Finance
- ...

State space models (continued)

Let $\mathbf{y}_n \in \mathbb{R}^D$ denote the **continuous state variable** at time τ_n and $\mathbf{t}_n \in \mathbb{R}^d$ the associated observation (not necessarily of the same dimension).

We consider **discrete-time** state space models with **additive noise**:

$$\begin{cases} \mathbf{y}_n &= \mathbf{f}(\mathbf{y}_{n-1}, \mathbf{u}_{n-1}, \tau_{n-1}) + \mathbf{r}_{n-1}, \\ \mathbf{t}_n &= \mathbf{h}(\mathbf{y}_n, \mathbf{u}_n, \tau_n) + \mathbf{q}_n, \end{cases}$$

where n is the time index and $\mathbf{u}_n$ is a deterministic **control variable**.

The noise vectors $\{\mathbf{r}_n\}_{n>0}$ account for the **random perturbations** acting on the system, including the unmodelled dynamics or unmeasured inputs.

The noise vectors $\{\mathbf{q}_n\}_{n>0}$ correspond to **measurement noise**.

The noises at any time are mutually independent.They are also independent from the noises or states at any other times.

For simplicity, we assume that there are **no deterministic inputs** and that the system is **time-invariant**.

State space models (continued)

SSMs are equivalent to first order HMMs for continuous state variables:

- The **Markov property** still holds and tells us that the state sequence is non-anticipative (cannot look into the future).
- The dynamical model is described by the **transition density**, which is induced by the deterministic, nonlinear output function $\mathbf{f}(\cdot)$ and the process noise $\mathbf{r}$:

  $$p(\mathbf{y}_n|\mathbf{y}_{n-1}, \boldsymbol{\theta}),$$

  where the parameter vector $\boldsymbol{\theta}$ specifies $\mathbf{f}(\cdot)$ and the noise distribution $p(\mathbf{r})$.
- The measurement model is described by the local **likelihood**, which is induced by the observation operator $\mathbf{h}(\cdot)$ and the observation noise $\mathbf{q}$:

  $$p(\mathbf{t}_n|\mathbf{y}_n, \boldsymbol{\vartheta}).$$

  The parameter vector $\boldsymbol{\vartheta}$ specifies $\mathbf{h}(\cdot)$ and the noise distribution $p(\mathbf{q})$.

Note that the prior distribution over the **initial state** needs to be specified.

Example: Gaussian random walk

The transition density and the noise are both Gaussian:

$$p(y_n|y_{n-1}) = \mathcal{N}(y_{n-1}, r^2),$$
$$p(t_n|y_n) = \mathcal{N}(y_n, q^2),$$

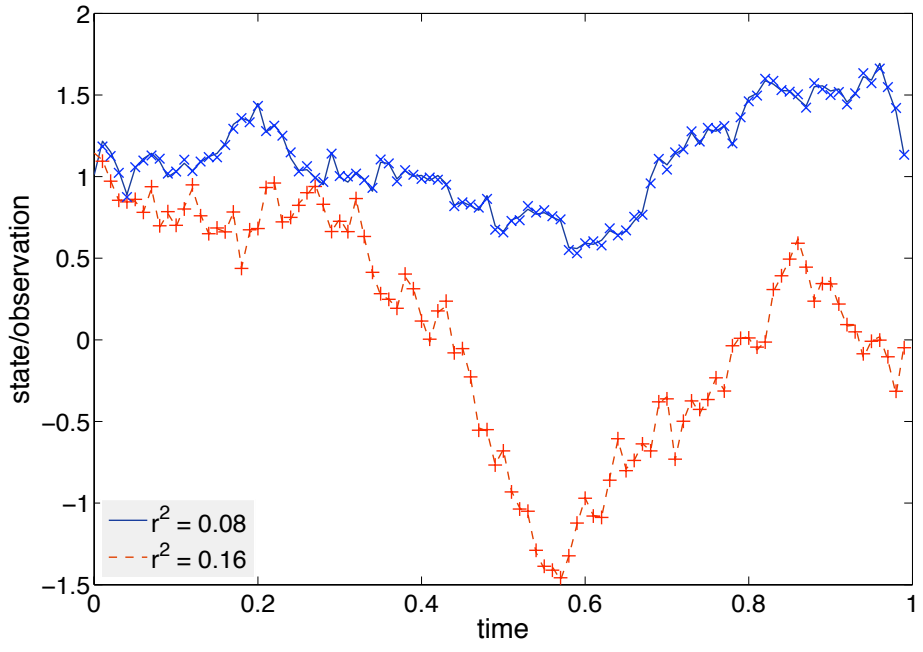


Figure: Two Gaussian random walks starting at $y_1 = 1$. The markers denote the observations (q=0.04).

Estimation problem

Assume $\mathbf{f}(\cdot)$, $\mathbf{h}(\cdot)$ and the noise processes are known. The **conditional mean** is a good candidate for estimating the latent state $\mathbf{y}_n$ given $\mathbf{t}_{1:k} \equiv \{\mathbf{t}_k\}_{k>0}$:

$$\bar{\boldsymbol{\mu}}_n = \langle \mathbf{y}_n | \mathbf{t}_{1:k} \rangle.$$

A suitable measure of the uncertainty is then given by the **conditional covariance**

$$\bar{\boldsymbol{\Sigma}}_n = \langle (\mathbf{y}_n - \bar{\boldsymbol{\mu}}_n)(\mathbf{y}_n - \bar{\boldsymbol{\mu}}_n)^\top | \mathbf{t}_{1:k} \rangle.$$

Depending on the value of k, we call the estimation problem:

- **Prediction** if $k < n$.
- **Filtering** if $k = n$.
- **Smoothing** if $k > n$.

Solving the estimation problem in an optimal way involves a **Bayesian recursion** algorithm.

Optimal (Bayesian) filtering

In order to estimate the conditional mean and the conditional covariance, we need to compute the posterior or **filtering density**:

$$\begin{aligned} p(\mathbf{y}_n|\mathbf{t}_{1:n}) &\propto p(\mathbf{t}_n|\mathbf{y}_n, \mathbf{t}_{1:n-1})p(\mathbf{y}_n|\mathbf{t}_{1:n-1}) \\ &\propto \underbrace{p(\mathbf{t}_n|\mathbf{y}_n)}_{\textit{likelihood}} p(\mathbf{y}_n|\mathbf{t}_{1:n-1}). \qquad (\textit{Markov property}) \end{aligned}$$

The **predictive density** or (projected) prior is given by

$$\begin{aligned} p(\mathbf{y}_n|\mathbf{t}_{1:n-1}) &= \int p(\mathbf{y}_n, \mathbf{y}_{n-1}|\mathbf{t}_{1:n-1})d\mathbf{y}_{n-1} \\ &= \int p(\mathbf{y}_n|\mathbf{y}_{n-1}, \mathbf{t}_{1:n-1})p(\mathbf{y}_{n-1}|\mathbf{t}_{1:n-1})d\mathbf{y}_{n-1} \\ &= \int \underbrace{p(\mathbf{y}_n|\mathbf{y}_{n-1})}_{\textit{transition density}} \underbrace{p(\mathbf{y}_{n-1}|\mathbf{t}_{1:n-1})}_{\textit{filtering density!}} d\mathbf{y}_{n-1}. \qquad (\textit{Markov property}) \end{aligned}$$

This integral is only tractable in the linear-Gaussian case (i.e. Kalman filter).

The extensions of KF propose different ways to approximate this integral in the nonlinear case.

Sequential smoothing

For optimal Bayesian smoothing, we are interested in $p(\mathbf{y}_n|\mathbf{t}_{1:k})$ for $k > n$:

1. The **forward recursion** consists in computing the filtering density up to time τ_k.
2. The **backward recursion** corresponds to propagating back the messages from future observations to time τ_n:

$$\begin{aligned} p(\mathbf{y}_n|\mathbf{t}_{1:k}) &= \int p(\mathbf{y}_n, \mathbf{y}_{n+1}|\mathbf{t}_{1:k}) d\mathbf{y}_{n+1} \\ &= \int p(\mathbf{y}_n|\mathbf{y}_{n+1}, \mathbf{t}_{1:k}) p(\mathbf{y}_{n+1}|\mathbf{t}_{1:k}) d\mathbf{y}_{n+1}, \end{aligned}$$

where $p(\mathbf{y}_n|\mathbf{y}_{n+1}, \mathbf{t}_{1:k}) = p(\mathbf{y}_n|\mathbf{y}_{n+1}, \mathbf{t}_{1:n})$ by the Markov property.

Applying Bayes' rule leads to

$$p(\mathbf{y}_n|\mathbf{y}_{n+1}, \mathbf{t}_{1:n}) = \frac{p(\mathbf{y}_{n+1}|\mathbf{y}_n, \mathbf{t}_{1:n}) \overbrace{p(\mathbf{y}_n|\mathbf{t}_{1:n})}^{\textit{filtering density}}}{\underbrace{p(\mathbf{y}_{n+1}|\mathbf{t}_{1:n})}_{\textit{predictive density}}} = \frac{\overbrace{p(\mathbf{y}_{n+1}|\mathbf{y}_n)}^{\textit{transition density}} p(\mathbf{y}_n|\mathbf{t}_{1:n})}{p(\mathbf{y}_{n+1}|\mathbf{t}_{1:n})}.$$

Usually, the observations taken into account are in a **time window** of fixed size.

Linear dynamical systems

SMM with linear, time-invariant transition and output functions:

$$\mathbf{f}(\mathbf{y}) = \mathbf{F}\mathbf{y} \quad \text{and} \quad \mathbf{h}(\mathbf{t}) = \mathbf{H}\mathbf{y},$$

where $\mathbf{F} \in \mathbb{R}^{D \times D}$ and $\mathbf{H} \in \mathbb{R}^{d \times D}$, i.e. the observations are a linear projection of the latent states.

SMM with additive Gaussian noise distributions:

$$\mathbf{r}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{R}) \quad \text{and} \quad \mathbf{q}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$$

for all n.

The **linear-Gaussian state-space model** can be reformulated as follows:

$$p(\mathbf{y}_n|\mathbf{y}_{n-1}) = \mathcal{N}(\mathbf{F}\mathbf{y}_{n-1}, \mathbf{R}),$$
$$p(\mathbf{t}_n|\mathbf{y}_n) = \mathcal{N}(\mathbf{H}\mathbf{y}_n, \mathbf{Q}).$$

The Kalman filter is **exact** for a linear-Gaussian state-space model.

Kalman filter (KF)

KF is only concerned with propagating the two first moments of the filtering density:

0. Assume the filtering density at τ_{n-1} is given by

$$\mathbf{y}_{n-1}|\mathbf{t}_{1:n-1} \sim \mathcal{N}(\bar{\boldsymbol{\mu}}_{n-1}, \bar{\boldsymbol{\Sigma}}_{n-1}).$$

1. The **predictive density** is then Gaussian:

$$\begin{aligned} p(\mathbf{y}_n|\mathbf{t}_{1:n-1}) &= \int \mathcal{N}(\mathbf{F}\mathbf{y}_{n-1}, \mathbf{R})\, \mathcal{N}(\bar{\boldsymbol{\mu}}_{n-1}, \bar{\boldsymbol{\Sigma}}_{n-1})\, d\mathbf{y}_{n-1} \\ &= \mathcal{N}(\underbrace{\mathbf{F}\bar{\boldsymbol{\mu}}_{n-1}}_{\equiv\hat{\boldsymbol{\mu}}_n}, \underbrace{\mathbf{R} + \mathbf{F}\bar{\boldsymbol{\Sigma}}_{n-1}\mathbf{F}^\top}_{\equiv\widehat{\boldsymbol{\Sigma}}_n}). \end{aligned}$$

2. The new **filtering density** is also Gaussian:

$$\begin{aligned} p(\mathbf{y}_n|\mathbf{t}_{1:n}) &\propto \mathcal{N}(\mathbf{H}\mathbf{y}_n, \mathbf{Q})\, \mathcal{N}(\hat{\boldsymbol{\mu}}_n, \widehat{\boldsymbol{\Sigma}}_n) \\ &= \mathcal{N}(\underbrace{\bar{\boldsymbol{\Sigma}}_n(\widehat{\boldsymbol{\Sigma}}_n^{-1}\hat{\boldsymbol{\mu}}_n + \mathbf{H}^\top\mathbf{Q}^{-1}\mathbf{t}_n)}_{\equiv\bar{\boldsymbol{\mu}}_n}, \underbrace{(\widehat{\boldsymbol{\Sigma}}_n^{-1} + \mathbf{H}^\top\mathbf{Q}^{-1}\mathbf{H})^{-1}}_{\equiv\bar{\boldsymbol{\Sigma}}_n}). \end{aligned}$$

KF is attractive for **online learning** as there is no need to keep track of the past conditional means and covariances.

Traditional view of the Kalman filter

KF is a linear filter which estimates the latent state by a linear combination of the state prediction $\hat{\boldsymbol{\mu}}_n$ and the **innovation vector** $\boldsymbol{\nu}_n \equiv \mathbf{t}_n - \mathbf{H}\hat{\boldsymbol{\mu}}_n$:

$$\bar{\boldsymbol{\mu}}_n = \hat{\boldsymbol{\mu}}_n + \mathbf{K}_n \boldsymbol{\nu}_n,$$

where the weighting matrix $\mathbf{K}_n \in \mathbb{R}^{D \times d}$ is the **Kalman gain** at time τ_n.

1. **Prediction step:**

$$\hat{\boldsymbol{\mu}}_n = \langle \mathbf{y}_n | \mathbf{t}_{1:n-1} \rangle,$$
$$\widehat{\boldsymbol{\Sigma}}_n = \langle (\mathbf{y}_n - \hat{\boldsymbol{\mu}}_n)(\mathbf{y}_n - \hat{\boldsymbol{\mu}}_n)^\top | \mathbf{t}_{1:n-1} \rangle.$$

2. **Correction step:**

$$\bar{\boldsymbol{\mu}}_n = \langle \mathbf{y}_n | \mathbf{t}_{1:n} \rangle = \hat{\boldsymbol{\mu}}_n + \mathbf{K}_n \boldsymbol{\nu}_n,$$
$$\bar{\boldsymbol{\Sigma}}_n = \langle (\mathbf{y}_n - \bar{\boldsymbol{\mu}}_n)(\mathbf{y}_n - \bar{\boldsymbol{\mu}}_n)^\top \rangle = \widehat{\boldsymbol{\Sigma}}_n - \mathbf{K}_n \mathbf{P}_n \mathbf{K}_n^\top,$$

where $\mathbf{P}_n \equiv \langle \boldsymbol{\nu}_n \boldsymbol{\nu}_n^\top \rangle = \mathbf{Q} + \mathbf{H}\widehat{\boldsymbol{\Sigma}}_n \mathbf{H}^\top$

The optimal Kalman gain (i.e. leading to a minimum variance estimator) is given by

$$\mathbf{K}_n = \langle (\mathbf{y}_n - \hat{\boldsymbol{\mu}}_n)\boldsymbol{\nu}_n^\top \rangle \, \langle \boldsymbol{\nu}_n \boldsymbol{\nu}_n^\top \rangle^{-1} = (\widehat{\boldsymbol{\Sigma}}_n \mathbf{H}^\top)\mathbf{P}_n^{-1}.$$

Are the two solutions equivalent?

Invoking the Woodbury identity we get

$$\begin{aligned}\bar{\mathbf{\Sigma}}_n &= (\hat{\mathbf{\Sigma}}_n^{-1} + \mathbf{H}^\top \mathbf{Q}^{-1}\mathbf{H})^{-1} \\ &= \hat{\mathbf{\Sigma}}_n - \hat{\mathbf{\Sigma}}_n\mathbf{H}^\top(\mathbf{Q} + \mathbf{H}\hat{\mathbf{\Sigma}}_n\mathbf{H}^\top)^{-1}\mathbf{H}\hat{\mathbf{\Sigma}}_n \\ &= \hat{\mathbf{\Sigma}}_n - \mathbf{K}_n\mathbf{P}_n \underbrace{\mathbf{P}_n^{-1}\mathbf{H}\hat{\mathbf{\Sigma}}_n}_{=\mathbf{K}_n^\top}.\end{aligned}$$

The posterior mean can also be rewritten in the desired form:

$$\begin{aligned}\bar{\boldsymbol{\mu}}_n &= \bar{\mathbf{\Sigma}}_n(\hat{\mathbf{\Sigma}}_n^{-1}\hat{\boldsymbol{\mu}}_n + \mathbf{H}^\top\mathbf{Q}^{-1}\mathbf{t}_n) \\ &= \hat{\boldsymbol{\mu}}_n - \mathbf{K}_n\mathbf{H}\hat{\boldsymbol{\mu}}_n + \hat{\mathbf{\Sigma}}_n\mathbf{H}^\top\mathbf{Q}^{-1}\mathbf{t}_n - \mathbf{K}_n\mathbf{H}\hat{\mathbf{\Sigma}}_n\mathbf{H}^\top\mathbf{Q}^{-1}\mathbf{t}_n \\ &= \hat{\boldsymbol{\mu}}_n + \mathbf{K}_n\big(-\mathbf{H}\hat{\boldsymbol{\mu}}_n + \mathbf{P}_n\mathbf{Q}^{-1}\mathbf{t}_n - \mathbf{H}\hat{\mathbf{\Sigma}}_n\mathbf{H}^\top\mathbf{Q}^{-1}\mathbf{t}_n\big) \\ &= \hat{\boldsymbol{\mu}}_n + \mathbf{K}_n\big(\underbrace{-\mathbf{H}\hat{\boldsymbol{\mu}}_n + \mathbf{Q}\mathbf{Q}^{-1}\mathbf{t}_n}_{=\boldsymbol{\nu}_n}\big).\end{aligned}$$

Kalman smoother (KS)

Let $p(\mathbf{y}_{n+1}|\mathbf{t}_{1:k})$ be equal to $\mathcal{N}(\mathbf{m}_{n+1}, \mathbf{S}_{n+1})$.

1. From the **forward recursion** (i.e. KF), we obtain the predictive and the filltering density.
2. The **backward recursion** leads to

$$
\begin{aligned}
p(\mathbf{y}_n|\mathbf{t}_{1:k}) &= \int p(\mathbf{y}_n|\mathbf{y}_{n+1}, \mathbf{t}_{1:n}) p(\mathbf{y}_{n+1}|\mathbf{t}_{1:k}) d\mathbf{y}_{n+1} \\
&= \int \mathcal{N}(\widetilde{\mathbf{A}}_n \mathbf{y}_{n+1} + \tilde{\mathbf{b}}_n, \widetilde{\mathbf{\Sigma}}_n)\, \mathcal{N}(\mathbf{m}_{n+1}, \mathbf{S}_{n+1}) d\mathbf{y}_{n+1} \\
&= \mathcal{N}(\underbrace{\widetilde{\mathbf{A}}_n \mathbf{m}_{n+1} + \tilde{\mathbf{b}}_n}_{\equiv \mathbf{m}_n}, \underbrace{\widetilde{\mathbf{\Sigma}}_n + \widetilde{\mathbf{A}}_n \mathbf{S}_{n+1} \widetilde{\mathbf{A}}_n^\top}_{\equiv \mathbf{S}_n}),
\end{aligned}
$$

which follows from

$$
\begin{aligned}
p(\mathbf{y}_n|\mathbf{y}_{n+1}, \mathbf{t}_{1:n}) &\propto p(\mathbf{y}_{n+1}|\mathbf{y}_n)\; p(\mathbf{y}_n|\mathbf{t}_{1:n}) \\
&\propto \mathcal{N}(\mathbf{F}\mathbf{y}_n, \mathbf{R})\, \mathcal{N}(\bar{\boldsymbol{\mu}}_n, \bar{\mathbf{\Sigma}}_n) \\
&= \mathcal{N}(\underbrace{\widetilde{\mathbf{A}}_n \mathbf{y}_{n+1} + \tilde{\mathbf{b}}_n}_{\equiv \tilde{\boldsymbol{\mu}}_n}, \underbrace{(\bar{\mathbf{\Sigma}}_n^{-1} + \mathbf{F}^\top \mathbf{R}^{-1} \mathbf{F})^{-1}}_{\equiv \widetilde{\mathbf{\Sigma}}_n}),
\end{aligned}
$$

where $\widetilde{\mathbf{A}}_n \equiv \bar{\mathbf{\Sigma}}_n \mathbf{F}^\top \widehat{\mathbf{\Sigma}}_{n+1}^{-1}$ and $\tilde{\mathbf{b}}_n \equiv \bar{\boldsymbol{\mu}}_n - \widetilde{\mathbf{A}}_n \hat{\boldsymbol{\mu}}_{n+1}$.

The Kalman smoother is also known as the *Rauch-Tung-Striebel smoother*.

The posterior mean $\tilde{\mu}_n$ follows form one of the Gaussian identities discussed in Lecture 1a:

$$
\begin{aligned}
\tilde{\mu}_n &= \widetilde{\Sigma}_n(\bar{\Sigma}_n^{-1}\bar{\mu}_n + \mathbf{F}^\top \mathbf{R}^{-1}\mathbf{y}_{n+1}) && \text{(Woodbury)} \\
&= \bar{\mu}_n - \bar{\Sigma}_n\mathbf{F}^\top(\mathbf{R} + \mathbf{F}\bar{\Sigma}_n\mathbf{F}^\top)^{-1}\mathbf{F}\bar{\mu}_n \\
&\quad + \bar{\Sigma}_n\mathbf{F}^\top\mathbf{R}^{-1}\mathbf{y}_{n+1} - \bar{\Sigma}_n\mathbf{F}^\top(\mathbf{R} + \mathbf{F}\bar{\Sigma}_n\mathbf{F}^\top)^{-1}\mathbf{F}\bar{\Sigma}_n\mathbf{F}^\top\mathbf{R}^{-1}\mathbf{y}_{n+1} \\
&= \bar{\mu}_n - \bar{\Sigma}_n\mathbf{F}^\top(\mathbf{R} + \mathbf{F}\bar{\Sigma}_n\mathbf{F}^\top)^{-1}(\mathbf{F}\bar{\mu}_n - (\mathbf{R} + \mathbf{F}\bar{\Sigma}_n\mathbf{F}^\top)\mathbf{R}^{-1}\mathbf{y}_{n+1} + \mathbf{F}\bar{\Sigma}_n\mathbf{F}^\top\mathbf{R}^{-1}\mathbf{y}_{n+1}) \\
&= \bar{\mu}_n + \bar{\Sigma}_n\mathbf{F}^\top(\mathbf{R} + \mathbf{F}\bar{\Sigma}_n\mathbf{F}^\top)^{-1}(\mathbf{y}_{n+1} - \mathbf{F}\bar{\mu}_n).
\end{aligned}
$$

Linearised dynamical system

When $\mathbf{f}(\cdot)$ is **nonlinear**, the transition probability $p(\mathbf{y}_n|\mathbf{y}_{n-1})$ is non-Gaussian and the predictive distribution $p(\mathbf{y}_n|\mathbf{t}_{1:n-1})$ is in general **intractable**.

A similar problem arises when $\mathbf{h}(\cdot)$ is nonlinear.

A possible approach is to consider a **linearised dynamical system** around the estimate $\bar{\boldsymbol{\mu}}_{n-1}$ (or $\hat{\boldsymbol{\mu}}_n$) of the current state $\mathbf{y}_{n-1}$ (or $\mathbf{y}_n$):

$$\begin{cases} \mathbf{y}_n & \approx & \mathbf{f}\left(\bar{\boldsymbol{\mu}}_{n-1}\right) + \nabla\mathbf{f}\big|_{\bar{\boldsymbol{\mu}}_{n-1}} \left(\mathbf{y}_{n-1} - \bar{\boldsymbol{\mu}}_{n-1}\right) + \ldots + \mathbf{r}_{n-1}, \\ \mathbf{t}_n & \approx & \mathbf{h}\left(\hat{\boldsymbol{\mu}}_n\right) + \nabla\mathbf{h}\big|_{\hat{\boldsymbol{\mu}}_n} \left(\mathbf{y}_n - \hat{\boldsymbol{\mu}}_n\right) + \ldots + \mathbf{q}_n, \end{cases}$$

where $\nabla\mathbf{f} \in \mathbb{R}^{D\times D}$ and $\nabla\mathbf{h} \in \mathbb{R}^{d\times D}$ are the Jacobians of respectively $\mathbf{f}(\cdot)$ and $\mathbf{h}(\cdot)$ wrt $\mathbf{y}$.

Hence, **approximate** transition density and likelihood are again Gaussian:

$$q(\mathbf{y}_n|\mathbf{y}_{n-1}) = \mathcal{N}(\bar{\mathbf{F}}_{n-1}\mathbf{y}_{n-1} + \mathbf{a}_{n-1}, \mathbf{R}),$$
$$q(\mathbf{t}_n|\mathbf{y}_n) = \mathcal{N}(\widehat{\mathbf{H}}_n\mathbf{y}_n + \mathbf{b}_n, \mathbf{Q}),$$

where $\mathbf{a}_{n-1} \equiv \mathbf{f}(\bar{\boldsymbol{\mu}}_{n-1}) - \bar{\mathbf{F}}_{n-1}\bar{\boldsymbol{\mu}}_{n-1}$ and $\mathbf{b}_n \equiv \mathbf{h}\left(\hat{\boldsymbol{\mu}}_n\right) - \widehat{\mathbf{H}}_n\hat{\boldsymbol{\mu}}_n$.

Extended Kalman filter (EKF)

Assume that the filtering density is equal to $\mathcal{N}(\bar{\boldsymbol{\mu}}_{n-1}, \bar{\boldsymbol{\Sigma}}_{n-1})$ at time τ_{n-1}.

1. The **predictive density** is Gaussian:

$$\begin{aligned} p(\mathbf{y}_n|\mathbf{t}_{1:n-1}) &= \int q(\mathbf{y}_n|\mathbf{y}_{n-1}) p(\mathbf{y}_{n-1}|\mathbf{t}_{1:n-1}) d\mathbf{y}_{n-1} \\ &= \mathcal{N}(\underbrace{\mathbf{f}(\bar{\boldsymbol{\mu}}_{n-1})}_{\equiv \hat{\boldsymbol{\mu}}_n}, \underbrace{\mathbf{R} + \bar{\mathbf{F}}_{n-1} \bar{\boldsymbol{\Sigma}}_{n-1} \bar{\mathbf{F}}_{n-1}^\top}_{\equiv \widehat{\boldsymbol{\Sigma}}_n}). \end{aligned}$$

2. The **filtering density** is also Gaussian:

$$\begin{aligned} p(\mathbf{y}_n|\mathbf{t}_{1:n}) &\propto q(\mathbf{t}_n|\mathbf{y}_n) p(\mathbf{y}_n|\mathbf{t}_{1:n-1}) \\ &= \mathcal{N}(\bar{\boldsymbol{\mu}}_n, \bar{\boldsymbol{\Sigma}}_n), \end{aligned}$$

   where

$$\begin{aligned} \bar{\boldsymbol{\mu}}_n &= \bar{\boldsymbol{\Sigma}}_n \left\{ \widehat{\boldsymbol{\Sigma}}_n^{-1} \hat{\boldsymbol{\mu}}_n + \widehat{\mathbf{H}}_n^\top \mathbf{Q}^{-1} (\mathbf{t}_n - \mathbf{b}_n) \right\}, \\ \bar{\boldsymbol{\Sigma}}_n &= (\widehat{\boldsymbol{\Sigma}}_n^{-1} + \widehat{\mathbf{H}}_n^\top \mathbf{Q}^{-1} \widehat{\mathbf{H}}_n)^{-1}. \end{aligned}$$

The filtered state estimate $\bar{\boldsymbol{\mu}}_n$ is also given by $\hat{\boldsymbol{\mu}}_n + \mathbf{K}_n(\mathbf{t}_n - \mathbf{h}(\hat{\boldsymbol{\mu}}_n))$, with $\mathbf{K}_n$ being the **Kalman gain**.

Extended Kalman smoother (EKS)

Let $p(\mathbf{y}_{n+1}|\mathbf{t}_{1:k})$ be equal to $\mathcal{N}(\mathbf{m}_{n+1}, \mathbf{S}_{n+1})$.

1. From the **forward recursion** (i.e. EKF), we obtain the predictive and the filltering density.
2. The **backward recursion** is similar to that of KS:

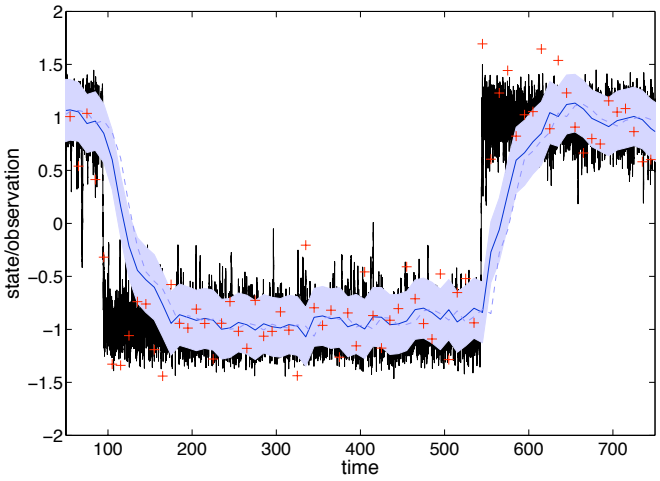
$$p(\mathbf{y}_n|\mathbf{t}_{1:k}) = \mathcal{N}(\underbrace{\widetilde{\mathbf{A}}_n\mathbf{m}_{n+1} + \tilde{\mathbf{b}}_n}_{\equiv\mathbf{m}_n}, \underbrace{\widetilde{\mathbf{\Sigma}}_n + \widetilde{\mathbf{A}}_n\mathbf{S}_{n+1}\widetilde{\mathbf{A}}_n^\top}_{\equiv\mathbf{S}_n}),$$

where

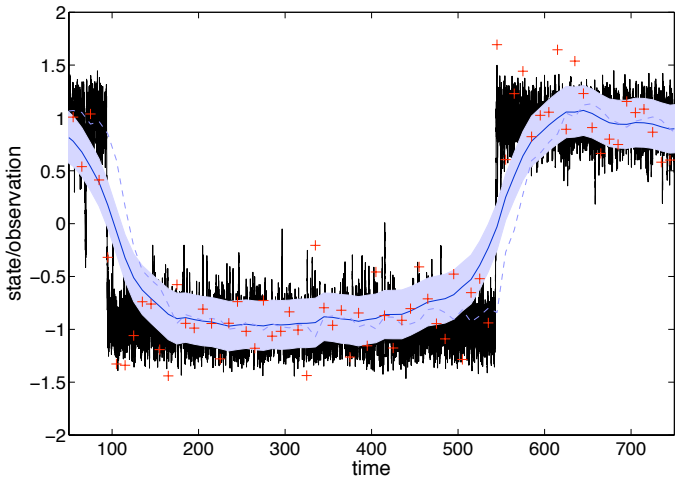
$$\begin{aligned}
\widetilde{\mathbf{A}}_n &= \bar{\mathbf{\Sigma}}_n\bar{\mathbf{F}}_n^\top\widehat{\mathbf{\Sigma}}_{n+1}^{-1}, \\
\tilde{\mathbf{b}}_n &= \bar{\boldsymbol{\mu}}_n - \widetilde{\mathbf{A}}_n\mathbf{f}(\bar{\boldsymbol{\mu}}_n), \\
\widetilde{\mathbf{\Sigma}}_n &= (\bar{\mathbf{\Sigma}}_n^{-1} + \bar{\mathbf{F}}_n^\top\mathbf{R}^{-1}\bar{\mathbf{F}}_n)^{-1}.
\end{aligned}$$

In practice, EKF and EKS work well for quasi linear dynamical systems or when the rate at which observations arrive is sufficiently high.

Example



(a) EKF.



(b) EKS.

Figure: Bi-stable system $f(y_n) = y_n + 4(y_n - y_n^3)\Delta t$, with additive Gaussian noise.

Flaws of EKF/EKS

- The second and higher order terms in the Taylor expansion are not negligible for **highly nonlinear** functions.
- The EKF/EKS do not take the **uncertainty on the latent states** into account when linearising.
- The functions $\mathbf{f}(\cdot)$ and $\mathbf{h}(\cdot)$ are not necessarily differentiable analytically.
- Implementation difficulties may arise when the system is composed of many states.
- Where should we linearise?

(a) Gaussian prior.

(b) Nonlinear transformation.

Figure: First and second order moments when applying a nonlinear transformation to a Gaussian random variable. The blue curve is exact and the red curve is obtained when linearising around the prior mean.

Statistical linearisation

Let Y be a continuous random variable to which we apply a nonlinear transformation $\mathbf{f}(\cdot)$.

We would like to take the **uncertainty** on a specific value $\mathbf{y}$ into account when linearising $\mathbf{f}(\cdot)$ at that point.

Consider the following linear approximation:

$$\mathbf{f}(\mathbf{y}) \approx \mathbf{A}\mathbf{y} + \mathbf{b}.$$

We would like to obtain the best (i.e. in the minimum squared error sense) linearised approximation **on average**:

$$\{\mathbf{A}, \mathbf{b}\} \leftarrow \underset{\mathbf{A},\mathbf{b}}{\operatorname{argmin}} \left\langle (\mathbf{f}(\mathbf{y}) - \mathbf{A}\mathbf{y} - \mathbf{b})^\top (\mathbf{f}(\mathbf{y}) - \mathbf{A}\mathbf{y} - \mathbf{b}) \right\rangle.$$

This leads to

$$\begin{aligned} \mathbf{A} &= \left\langle (\mathbf{f}(\mathbf{y}) - \langle \mathbf{f}(\mathbf{y}) \rangle) \, (\mathbf{y} - \boldsymbol{\mu})^\top \right\rangle \, \boldsymbol{\Sigma}^{-1}, \\ \mathbf{b} &= \langle \mathbf{f}(\mathbf{y}) \rangle - \mathbf{A}\boldsymbol{\mu}, \end{aligned}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and the covariance of Y.

The expected squared error is given by

$$E(\mathbf{A}, \mathbf{b}) = \langle (\mathbf{f}(\mathbf{y}) - \mathbf{A}\mathbf{y} - \mathbf{b})^\top (\mathbf{f}(\mathbf{y}) - \mathbf{A}\mathbf{y} - \mathbf{b}) \rangle.$$

Taking the derivative wrt $\mathbf{b}$ and setting to zero leads to

$$0 = \langle \mathbf{f}(\mathbf{y}) - \mathbf{A}\mathbf{y} - \mathbf{b} \rangle \quad \Leftrightarrow \quad \mathbf{b} = \langle \mathbf{f}(\mathbf{y}) \rangle - \mathbf{A}\langle \mathbf{y} \rangle.$$

Substituting this expression in $E(\mathbf{A}, \mathbf{b})$, taking the derivative wrt $\mathbf{A}$ and setting to zero leads to

$$\begin{aligned}
0 &= \langle (\mathbf{f}(\mathbf{y}) - \langle \mathbf{f}(\mathbf{y}) \rangle - \mathbf{A}(\mathbf{y} - \langle \mathbf{y} \rangle)\ (\mathbf{y} - \langle \mathbf{y} \rangle)^\top \rangle \\
\Leftrightarrow \quad & \mathbf{A}\langle (\mathbf{y} - \langle \mathbf{y} \rangle)(\mathbf{y} - \langle \mathbf{y} \rangle)^\top \rangle = \langle (\mathbf{f}(\mathbf{y}) - \langle \mathbf{f}(\mathbf{y}) \rangle)(\mathbf{y} - \langle \mathbf{y} \rangle)^\top \rangle \\
\Leftrightarrow \quad & \mathbf{A} = \langle (\mathbf{f}(\mathbf{y}) - \langle \mathbf{f}(\mathbf{y}) \rangle)(\mathbf{y} - \langle \mathbf{y} \rangle)^\top \rangle\ \langle (\mathbf{y} - \langle \mathbf{y} \rangle)(\mathbf{y} - \langle \mathbf{y} \rangle)^\top \rangle^{-1}.
\end{aligned}$$

Effect of nonlinear transformations to random variables with Gaussian prior

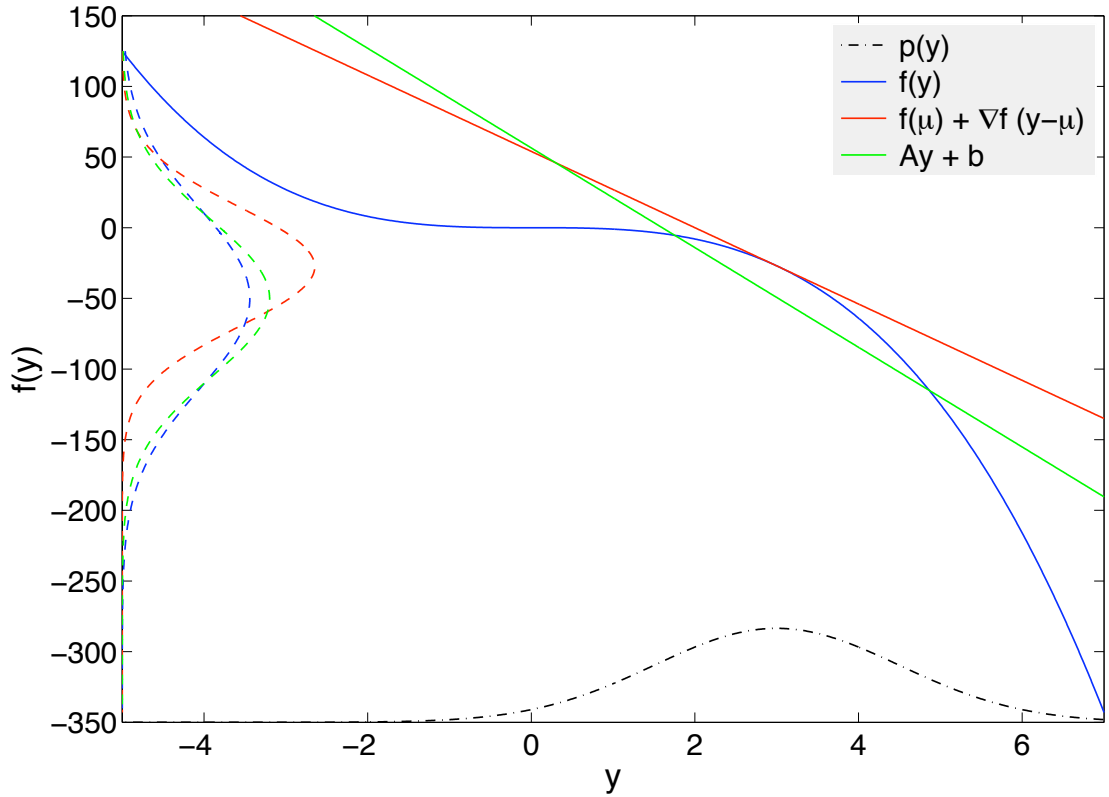


Figure: The first and second order moments are more accurate after statistical linearisation than when using a first order truncated Taylor-series expansion of the function around a single point, e.g. the prior mean.

Statistical linearisation using sigma points

Let us call **sigma points** a set of weighted points $\{\mathbf{y}_l\}_{l=0}^{L}$ chosen deterministically, which capture the mean and covariance[1] of the random variable Y:

$$\boldsymbol{\mu} \approx \sum_{l=0}^{L} w_l \mathbf{y}_l,$$

$$\boldsymbol{\Sigma} \approx \sum_{l=0}^{L} w_l (\mathbf{y}_l - \boldsymbol{\mu}_n)(\mathbf{y}_l - \boldsymbol{\mu}_n)^\top,$$

where $\{w_l\}_{l=0}^{L}$ is the set of **weights**, with $\sum_l w_l = 1$.

The pairs of weights and sigma points allow us to **approximate any expectation** wrt the distribution of Y, such that

$$E(\mathbf{A}, \mathbf{b}) \approx \sum_{l=0}^{L} w_l (\mathbf{f}(\mathbf{y}_l) - \mathbf{A}\mathbf{y}_l - \mathbf{b})^\top (\mathbf{f}(\mathbf{y}_l) - \mathbf{A}\mathbf{y}_l - \mathbf{b})$$

and

$$\mathbf{A} \approx \left(\textstyle\sum_l w_l (\mathbf{f}(\mathbf{y}_l) - \bar{\mathbf{f}})\ (\mathbf{y}_l - \boldsymbol{\mu})^\top \right) \boldsymbol{\Sigma}^{-1},$$

$$\mathbf{b} \approx \bar{\mathbf{f}} - \mathbf{A}\boldsymbol{\mu}, \qquad \bar{\mathbf{f}} \equiv \textstyle\sum_l w_l \mathbf{f}(\mathbf{y}_l).$$

[1] This can be generalised to higher order moments of Y.

Unscented transform (UT)

Let $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ be the mean and the covariance of Y.

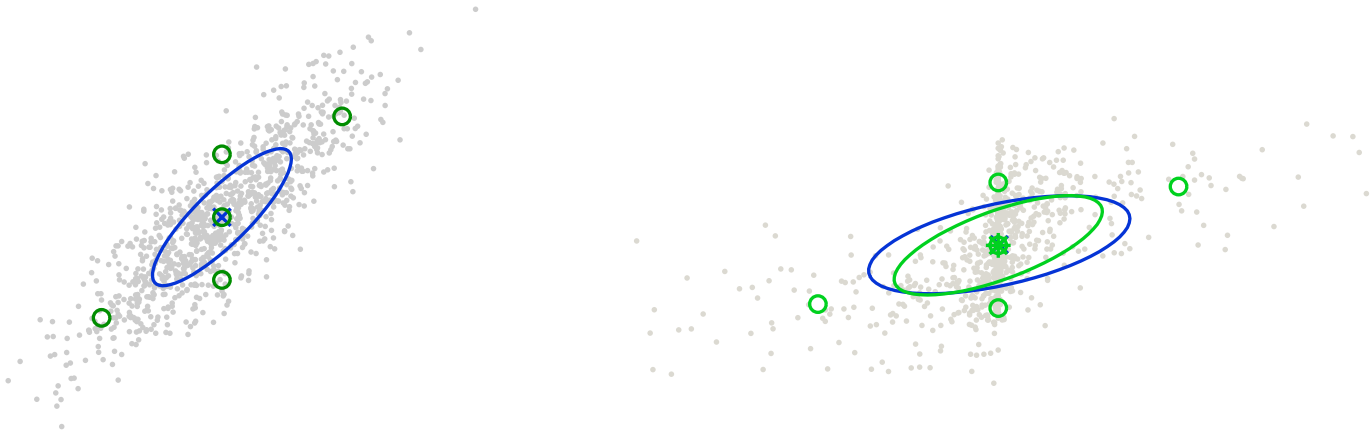
The $2D+1$ sigma points and weights are defined as follows:

$$\begin{aligned}
\mathbf{y}_0 &= \boldsymbol{\mu}, & w_0 &= \tfrac{\kappa}{D+\kappa}, & l &= 0, \\
\mathbf{y}_l &= \boldsymbol{\mu} + \left[\sqrt{(D+\kappa)\boldsymbol{\Sigma}}\right]_l, & w_l &= \tfrac{1}{2(D+\kappa)}, & l &= 1,\ldots,D, \\
\mathbf{y}_l &= \boldsymbol{\mu} - \left[\sqrt{(D+\kappa)\boldsymbol{\Sigma}}\right]_l, & w_l &= \tfrac{1}{2(D+\kappa)}, & l &= D+1,\ldots,2D,
\end{aligned}$$

where $(\cdot)_l$ denotes the l^{th} column of matrix square root $\sqrt{(D+\kappa)\boldsymbol{\Sigma}}$ and κ is a scale parameter (determining the radius of the sigma points from the mean).

- The matrix square root can be computed by Cholesky factorisation (lower triangular factor).
- The matrix square root is **not unique**, so any orthonormal rotation of the sigma-point set is again a valid set.
- The sigma points capture the true mean and covariance of Y.
- When propagated through any nonlinear system, the transformed sigma points capture the posterior mean and covariance up to the **2$^{\text{nd}}$ order**.
- The errors introduced in the 3^{rd} and higher orders can be minimised by optimising κ.

Example revisited



(a) Gaussian prior.

(b) Nonlinear transformation.

Figure: First and second order moments when applying a nonlinear transformation to a Gaussian random variable. The blue curve is exact and the green curve is obtained when using UT. The sigma points are indicated by small circles.

Comments

When directly linearising the nonlinear transformation (e.g. in EKF), the posterior mean and covariance are only accurate up to the first order.

By contrast, a suitable choice of sigma points results in being **accurate up to higher orders**.

Other methods for determining the sigma points and their weights include:

- The **Scaled UT**, which prevents the covariance to be negative-definite after transformation. This can happen when κ is chosen too negative.
- Sterling's polynomial interpolation formula, which replaces the analytical derivatives in the Taylor series expansion by **central divided differences**.
- **Gauss-Hermite quadrature**, which is specifically designed to approximate Gaussian integrals, but is $\mathcal{O}(L^D)$, the accuracy being dependent on L.
- The **ensemble** heuristic, which picks $L+1$ samples at random and weights them by $1/(L+1)$.

Statistically linearised dynamical system

When the state variable at time τ_n is uncertain, statistical linearisation is more suitable than directly linearising the nonlinear functions $\mathbf{f}(\cdot)$ and $\mathbf{h}(\cdot)$:

$$\begin{cases} \mathbf{y}_n & \approx & \mathbf{A}_{n-1}\mathbf{y}_{n-1} + \mathbf{b}_{n-1} + \mathbf{r}_{n-1}, \\ \mathbf{t}_n & \approx & \mathbf{C}_n\mathbf{y}_n + \mathbf{d}_n + \mathbf{q}_n, \end{cases}$$

where

- $\mathbf{A}_{n-1}$, $\mathbf{b}_{n-1}$ and $\bar{\mathbf{f}}_{n-1}$ are computed with sigma points from the **filtering density** $p(\mathbf{y}_{n-1}|\mathbf{t}_{1:n-1})$.
- $\mathbf{C}_n$, $\mathbf{d}_n$ and $\bar{\mathbf{h}}_n$ use sigma points from the **predictive density** $p(\mathbf{y}_n|\mathbf{t}_{1:n-1})$.

The **approximate** transition density and likelihood are again Gaussian:

$$q(\mathbf{y}_n|\mathbf{y}_{n-1}) = \mathcal{N}(\mathbf{A}_{n-1}\mathbf{y}_{n-1} + \mathbf{b}_{n-1}, \mathbf{R}),$$
$$q(\mathbf{t}_n|\mathbf{y}_n) = \mathcal{N}(\mathbf{C}_n\mathbf{y}_n + \mathbf{d}_n, \mathbf{Q}).$$

Sigma point Kalman filter (SPKF)

The idea is to only propagate **approximate** first and second order **moments**, which are more accurate than the ones obtained for the EKF.

Assume that the filtering density is equal to $\mathcal{N}(\bar{\boldsymbol{\mu}}_{n-1}, \bar{\boldsymbol{\Sigma}}_{n-1})$ at time τ_{n-1}.

1. The **predictive density** is Gaussian:

$$p(\mathbf{y}_n|\mathbf{t}_{1:n-1}) = \mathcal{N}(\underbrace{\bar{\mathbf{f}}_{n-1}}_{\equiv\hat{\boldsymbol{\mu}}_n}, \underbrace{\mathbf{R} + \mathbf{A}_{n-1}\bar{\boldsymbol{\Sigma}}_{n-1}\mathbf{A}_{n-1}^\top}_{\equiv\widehat{\boldsymbol{\Sigma}}_n}).$$

2. The **filtering density** is Gaussian:

$$p(\mathbf{y}_n|\mathbf{t}_{1:n}) = \mathcal{N}(\bar{\boldsymbol{\mu}}_n, \bar{\boldsymbol{\Sigma}}_n),$$

where

$$\bar{\boldsymbol{\mu}}_n = \bar{\boldsymbol{\Sigma}}_n \left\{ \widehat{\boldsymbol{\Sigma}}_n^{-1} \hat{\boldsymbol{\mu}}_n + \mathbf{C}_n^\top \mathbf{Q}^{-1}(\mathbf{t}_n - \mathbf{d}_n) \right\},$$

$$\bar{\boldsymbol{\Sigma}}_n = (\widehat{\boldsymbol{\Sigma}}_n^{-1} + \mathbf{C}_n^\top \mathbf{Q}^{-1}\mathbf{C}_n)^{-1}.$$

Sigma point Kalman filter (continued)

SPKF is attractive in practice:

- Derivativeless
- Based on deterministic sampling
- Gaussian approximate filter with **exact** nonlinear models:

1. **Prediction:**

$$\hat{\boldsymbol{\mu}}_n = \sum_l v_l \mathbf{f}(\mathbf{y}_{n-1}^{(l)}),$$
$$\widehat{\boldsymbol{\Sigma}}_n = \mathbf{R} + \sum_l v_l (\mathbf{f}(\mathbf{y}_{n-1}^{(l)}) - \hat{\boldsymbol{\mu}}_n)(\mathbf{f}(\mathbf{y}_{n-1}^{(l)}) - \hat{\boldsymbol{\mu}}_n)^\top,$$

where $\{\mathbf{y}_{n-1}^{(l)}\}_{l=0}^L$ are sigma points of the filtering density $p(\mathbf{y}_{n-1}|\mathbf{t}_{1:n-1})$ and $\{v_l\}_{l=0}^L$ the corresponding weights.

2. **Correction:**

$$\bar{\boldsymbol{\mu}}_n = \hat{\boldsymbol{\mu}}_n + \mathbf{K}_n(\mathbf{t}_n - \bar{\mathbf{h}}_n), \quad \mathbf{K}_n = (\sum_l w_l (\mathbf{h}(\tilde{\mathbf{y}}_n^{(l)}) - \bar{\mathbf{h}}_n)(\tilde{\mathbf{y}}_n^{(l)} - \hat{\boldsymbol{\mu}}_n)^\top)\mathbf{P}_n^{-1},$$
$$\bar{\boldsymbol{\Sigma}}_n = \widehat{\boldsymbol{\Sigma}}_n - \mathbf{K}_n \mathbf{P}_n \mathbf{K}_n^\top, \quad \mathbf{P}_n = \mathbf{Q} + \sum_l w_l (\mathbf{h}(\tilde{\mathbf{y}}_{n-1}^{(l)}) - \bar{\mathbf{h}}_n)(\mathbf{h}(\tilde{\mathbf{y}}_{n-1}^{(l)}) - \bar{\mathbf{h}}_n)^\top,$$

where $\{\tilde{\mathbf{y}}_n^{(l)}\}_{l=0}^L$ are sigma points of the predictive density $p(\mathbf{y}_n|\mathbf{t}_{1:n-1})$ and $\{w_l\}_{l=0}^L$ the corresponding weights.

Consider the sigma points $\{\mathbf{y}_{n-1}^{(l)}\}_{l=0}^{L}$ and the weights $\{v_l\}_{l=0}^{L}$. The covariance of the filtering density at time τ_{n-1} is approximated by

$$\bar{\mathbf{\Sigma}}_{n-1} \approx \sum_{l=0}^{L} v_l (\mathbf{y}_{n-1}^{(l)} - \bar{\boldsymbol{\mu}}_{n-1})(\mathbf{y}_{n-1}^{(l)} - \bar{\boldsymbol{\mu}}_{n-1})^\top.$$

Hence, the covariance of the predictive density at time τ_{n-1} is given by

$$\begin{aligned}
\widehat{\mathbf{\Sigma}}_n &= \mathbf{R} + \mathbf{A}_{n-1} \bar{\mathbf{\Sigma}}_{n-1} \mathbf{A}_{n-1}^\top \\
&= \mathbf{R} + \left(\textstyle\sum_l v_l (\mathbf{f}(\mathbf{y}_{n-1}^{(l)}) - \hat{\boldsymbol{\mu}}_n)(\mathbf{y}_{n-1}^{(l)} - \bar{\boldsymbol{\mu}}_{n-1})^\top \right) \bar{\mathbf{\Sigma}}_{n-1}^{-1} \bar{\mathbf{\Sigma}}_{n-1} \mathbf{A}_{n-1}^\top \\
&= \mathbf{R} + \textstyle\sum_l v_l (\mathbf{f}(\mathbf{y}_{n-1}^{(l)}) - \hat{\boldsymbol{\mu}}_n)(\mathbf{f}(\mathbf{y}_{n-1}^{(l)}) - \hat{\boldsymbol{\mu}}_n)^\top.
\end{aligned}$$

The proof is analogue for $\mathbf{P}_n$ and the Kalman gain is given by

$$\mathbf{K}_n \approx \left(\sum_{l=0}^{L} w_l (\mathbf{h}(\tilde{\mathbf{y}}_n^{(l)}) - \bar{\mathbf{h}}_n)(\tilde{\mathbf{y}}_n^{(l)} - \hat{\boldsymbol{\mu}}_n)^\top \right) \mathbf{P}_n^{-1},$$

where $\{\tilde{\mathbf{y}}_n^{(l)}\}_{l=0}^{L}$ and $\{w_l\}_{l=0}^{L}$ are the sigma points and the weights.

Types of SPKFs

The (scaled) UT and the central difference approximation lead respectively to the **unscented KF** (UKF) and the **central difference KF** (CDKF):

- The sigma points of both filters have a similar form.
- They perform equally well (i.e. with negligible difference) in estimation accuracy, but CDKF only needs to set a single parameter.
- Inaccuracies arise when the posterior is multi-modal.

SPKFs are different from particle filters as they work with a small number of particles, which are chosen deterministically.

The ensemble approach leads to the **Ensemble KF** (EnsKF):

- Degenerate solution
- Able to deal with very high dimensional state space, popular in Data Assimilation (e.g. numerical weather prediction).
- Use of heuristics to deal with practical problems such as rank deficiency, ensemble update, ensemble perturbation, etc.

EnsKF is different from particle filters as it works with a small number particles with same weight and assumes everything is Gaussian.

Sigma point Kalman smoother (SPKS)

Let $p(\mathbf{y}_{n+1}|\mathbf{t}_{1:k})$ be equal to $\mathcal{N}(\mathbf{m}_{n+1}, \mathbf{S}_{n+1})$.

1. From the **forward recursion** (i.e. SPKF), we obtain the predictive and the filltering density.
2. The **backward recursion** is similar to that of KS:

$$p(\mathbf{y}_n|\mathbf{t}_{1:k}) = \mathcal{N}(\underbrace{\widetilde{\mathbf{A}}_n\mathbf{m}_{n+1} + \tilde{\mathbf{b}}_n}_{\equiv \mathbf{m}_n}, \underbrace{\widetilde{\mathbf{\Sigma}}_n + \widetilde{\mathbf{A}}_n\mathbf{S}_{n+1}\widetilde{\mathbf{A}}_n^\top}_{\equiv \mathbf{S}_n}),$$

where

$$\begin{aligned}
\widetilde{\mathbf{A}}_n &= \bar{\mathbf{\Sigma}}_n\mathbf{A}_n^\top\widehat{\mathbf{\Sigma}}_{n+1}^{-1}, \\
\tilde{\mathbf{b}}_n &= \bar{\boldsymbol{\mu}}_n - \widetilde{\mathbf{A}}_n\bar{\mathbf{f}}_n, \\
\widetilde{\mathbf{\Sigma}}_n &= (\bar{\mathbf{\Sigma}}_n^{-1} + \mathbf{A}_n^\top\mathbf{R}^{-1}\mathbf{A}_n)^{-1}.
\end{aligned}$$

These quantities can be reformulated in terms of the weighted sigma points.

Parameter inference by latent state augmentation

Let us denote the parameters of $\mathbf{f}(\cdot)$ and $\mathbf{h}(\cdot)$ by $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$.

1. **Joint filtering (or smoothing):**
   - Consider the **augmented** state variable:

$$\begin{bmatrix} \mathbf{y}_n \\ \boldsymbol{\theta}_n \\ \boldsymbol{\vartheta}_n \end{bmatrix} = \begin{bmatrix} \mathbf{f}(\mathbf{y}_{n-1}, \boldsymbol{\theta}_{n-1}) + \mathbf{r}_{n-1} \\ \boldsymbol{\theta}_{n-1} + \boldsymbol{\sigma}_{n-1} \\ \boldsymbol{\vartheta}_{n-1} + \boldsymbol{\varsigma}_{n-1} \end{bmatrix}.$$

     where $\{\boldsymbol{\sigma}_n, \boldsymbol{\varsigma}_n\}_{n>0}$ reflects the (prior) uncertainty on the parameters.
   - Use EKF/S or SPKF/S to **jointly estimate** the latent states and the parameters.
2. **Dual filtering:**
   - Alternate between a filter to estimate the latent states for fixed parameters:

$$\begin{aligned} \mathbf{y}_n &= \mathbf{f}(\mathbf{y}_{n-1}, \bar{\boldsymbol{\theta}}_{n-1}) + \mathbf{r}_{n-1}, \\ \mathbf{t}_n &= \mathbf{h}(\mathbf{y}_n, \bar{\boldsymbol{\vartheta}}_{n-1}) + \mathbf{q}_n, \end{aligned}$$

   - And a filter to estimate the parameters for fixed latent states:

$$\begin{aligned} \boldsymbol{\theta}_n &= \boldsymbol{\theta}_{n-1} + \boldsymbol{\sigma}_{n-1}, \\ \boldsymbol{\vartheta}_n &= \boldsymbol{\vartheta}_{n-1} + \boldsymbol{\varsigma}_{n-1}, \\ \mathbf{t}_n &= \mathbf{h}(\bar{\boldsymbol{\mu}}_n, \boldsymbol{\vartheta}_n) + \mathbf{q}_n. \end{aligned}$$

A priori, the joint approach is to be preferred in practice as it models the **cross-correlations** between $\mathbf{y}$ and $\boldsymbol{\theta}$. However, it can lead to slow convergence.

Parameter inference by EM

The lower bound to the marginal log-likelihood is given by

$$\begin{aligned} -\mathcal{F}(q, \boldsymbol{\theta}, \boldsymbol{\vartheta}) &= \ln p(\mathbf{t}_{1:N}|\boldsymbol{\theta}, \boldsymbol{\vartheta}) - \mathrm{KL}[q(\mathbf{y}_{1:N}) \| p(\mathbf{y}_{1:N}|\mathbf{t}_{1:N}, \boldsymbol{\theta}, \boldsymbol{\vartheta})], \\ &= \left\langle \ln p(\mathbf{t}_{1:N}, \mathbf{y}_{1:N}|\boldsymbol{\theta}, \boldsymbol{\vartheta}) \right\rangle_{q(\mathbf{y}_{1:N})} + \mathrm{H}[q(\mathbf{y}_{1:N})]. \end{aligned}$$

- The **E step** consists in estimating the latent states for fixed $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$:

$$q(\mathbf{y}_{1:N}) = \prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{t}_{1:N}),$$

  where $p(\mathbf{y}_n|\mathbf{t}_{1:N})$ are the **local posterior marginals** obtained by the smoothing algorithm.

- The **M step** maximises the complete log-likelihood for fixed q:

$$\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\} \leftarrow \operatorname*{argmax}_{\boldsymbol{\theta}, \boldsymbol{\vartheta}} \left\langle \ln p(\mathbf{t}_{1:N}, \mathbf{y}_{1:N}|\boldsymbol{\theta}, \boldsymbol{\vartheta}) \right\rangle_{q(\mathbf{y}_{1:N})}.$$

- The **M step** for initial state distribution is ok as well.

For nonlinear state space models, the M step is performed by gradient ascent techniques (see e.g. Nocedal and Wright, 2000).

References

- Christopher M. Bishop: Pattern Recognition and Machine Learning. Springer, 2006.
- Geir Evensen: Data assimilation : The ensemble Kalman filter. Springer, 2007.
- Simon J. Julier and Jeffrey K. Uhlmann, A general method of approximating nonlinear transformations of probability distributions. Technical report, Department of Engineering Science (Robotics Research Group), University of Oxford, 1995.
- Rudolph E. Kalman, A new approach to linear filtering and prediction problems, Transactions of the ASME, Journal of Basic Engineering, 82, 34-45, 1960.
- Simo Särkkä, Recursive Bayesian inference on stochastic differential equations, Doctoral Dissertation, Helsinki University of Technology, 2006.
- Rudolph van der Merwe, Sigma-point Kalman filters for Ppobabilistic inference in dynamic state-space models, Doctoral Dissertation, Oregon Health & Science University, 2004.

- The Matrix Cookbook by Kaare B. Petersen and Michael S. Pedersen.