Statistics for Data Science

Wintersemester 2024/25

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Thirteenth lecture, January 20, 2025

The linear Gaussian setting

In these notes we study the linear Gaussian setting, where the forward map F is linear and both the prior distribution and the distribution of the observational noise ε are Gaussian.

It arises frequently in applications, either directly or in the form of posterior distributions that are asymptotically Gaussian in the large data limit. It also allows computing explicit solutions which can be used to gain a general understanding. Apart from that, many methods employed in a nonlinear or non-Gaussian setting build on ideas from the linear Gaussian case by performing linearization or Gaussian approximation.

Let us suppose that the unknown $x \in \mathbb{R}^d$ and the data $y \in \mathbb{R}^k$ follow the relation

$$y = Ax + \varepsilon, \tag{1}$$

where

1. The forward model is linear, i.e., $A \in \mathbb{R}^{k \times d}$.
2. The prior distribution is Gaussian: $x \sim \mathcal{N}(x_0, \Gamma_{\mathrm{pr}})$, where $x_0 \in \mathbb{R}^d$ and $\Gamma_{\mathrm{pr}} \in \mathbb{R}^{d \times d}$ is symmetric and positive definite.
3. The noise is Gaussian: $\varepsilon \sim \mathcal{N}(\varepsilon_0, \Gamma_{\mathrm{n}})$, where $\varepsilon_0 \in \mathbb{R}^k$ and $\Gamma_{\mathrm{n}} \in \mathbb{R}^{k \times k}$ is symmetric and positive definite.
4. x and ε are independent.

Theorem

Under assumptions 1–4, the posterior distribution corresponding to (1) is Gaussian with $x|y \sim \mathcal{N}(\mu_{\mathrm{post}}, \Gamma_{\mathrm{post}})$, where we have the posterior mean

$$\mu_{\mathrm{post}} = (\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}} \Gamma_{\mathrm{n}}^{-1} A)^{-1} (A^{\mathrm{T}} \Gamma_{\mathrm{n}}^{-1} (y - \varepsilon_0) + \Gamma_{\mathrm{pr}}^{-1} x_0)$$

and covariance

$$\Gamma_{\mathrm{post}} = (\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}} \Gamma_{\mathrm{n}}^{-1} A)^{-1}.$$

Proof. Noting that $\Gamma_{\mathrm{post}} = (\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}A)^{-1}$ and $\mu_{\mathrm{post}} = \Gamma_{\mathrm{post}}(A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y-\varepsilon_0) + \Gamma_{\mathrm{pr}}^{-1}x_0)$, we obtain

$$
\begin{aligned}
f(x|y) &\propto \exp\Big(-\frac{1}{2}(y - Ax - \varepsilon_0)^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y - Ax - \varepsilon_0)\Big)\exp\Big(-\frac{1}{2}(x - x_0)^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}(x - x_0)\Big) \\
&= \exp\Big(-\frac{1}{2}\big(y^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}y - y^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}Ax - y^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}\varepsilon_0 \\
&\qquad\qquad - x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}y + x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}Ax + x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}\varepsilon_0 \\
&\qquad\qquad - \varepsilon_0^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}y + \varepsilon_0^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}Ax + \varepsilon_0^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}\varepsilon_0 \\
&\qquad\qquad + x^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}x - 2x^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}x_0 + x_0^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}x_0\big)\Big)
\end{aligned}
$$

Proof. Noting that $\Gamma_{\mathrm{post}} = (\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}A)^{-1}$ and $\mu_{\mathrm{post}} = \Gamma_{\mathrm{post}}(A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y - \varepsilon_0) + \Gamma_{\mathrm{pr}}^{-1}x_0)$, we obtain

$$
\begin{aligned}
f(x|y) &\propto \exp\left(-\frac{1}{2}(y - Ax - \varepsilon_0)^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y - Ax - \varepsilon_0)\right) \exp\left(-\frac{1}{2}(x - x_0)^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}(x - x_0)\right) \\
&\propto \exp\Bigg(-\frac{1}{2}\Big(\qquad - x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}y \\
&\qquad\qquad - x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}y + x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}Ax + x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}\varepsilon_0 \\
&\qquad\qquad\qquad + x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}\varepsilon_0 \\
&\qquad\qquad + x^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}x - 2x^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}x_0 \qquad \Big)\Bigg)
\end{aligned}
$$

Proof. Noting that $\Gamma_{\mathrm{post}} = (\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}A)^{-1}$ and $\mu_{\mathrm{post}} = \Gamma_{\mathrm{post}}(A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y-\varepsilon_0) + \Gamma_{\mathrm{pr}}^{-1}x_0)$, we obtain

$$
\begin{aligned}
f(x|y) &\propto \exp\Big(-\frac{1}{2}(y - Ax - \varepsilon_0)^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y - Ax - \varepsilon_0)\Big)\exp\Big(-\frac{1}{2}(x-x_0)^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}(x-x_0)\Big) \\
&\propto \exp\Big(-\frac{1}{2}\big(\qquad - x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}y \\
&\qquad\qquad - x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}y + x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}Ax + x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}\varepsilon_0 \\
&\qquad\qquad\qquad + x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}\varepsilon_0 \\
&\qquad\qquad + x^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}x - 2x^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}x_0 \qquad \big)\Big) \\
&= \exp\Big(-\frac{1}{2}\big(x^{\mathrm{T}}\underbrace{(\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}A)}_{=\Gamma_{\mathrm{post}}^{-1}}x - 2x^{\mathrm{T}}\underbrace{(A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y-\varepsilon_0) + \Gamma_{\mathrm{pr}}^{-1}x_0)}_{=\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}}}\big)\Big).
\end{aligned}
$$

On the previous slide, we arrived at

$$f(x|y) \propto \exp\left(-\frac{1}{2}\left(x^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}x - 2x^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}}\right)\right).$$

To finish the proof, we "complete the square" by multiplying and dividing by $\exp(-\frac{1}{2}\mu_{\mathrm{post}}^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}})$. Since this term does not depend on x, we can absorb the denominator into the implied coefficient to obtain

$$\begin{aligned} f(x|y) &\propto \exp\left(-\frac{1}{2}\left(x^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}x - 2x^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}}\right)\right)\exp\left(-\frac{1}{2}\mu_{\mathrm{post}}^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}}\right) \\ &= \exp\left(-\frac{1}{2}\left(x^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}x - 2x^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}} + \mu_{\mathrm{post}}^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}}\right)\right) \\ &= \exp\left(-\frac{1}{2}\left((x-\mu_{\mathrm{post}})^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}(x-\mu_{\mathrm{post}}) + 2x^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}} - 2x^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}}\right)\right) \\ &= \exp\left(-\frac{1}{2}\left((x-\mu_{\mathrm{post}})^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}(x-\mu_{\mathrm{post}})\right)\right), \end{aligned}$$

as desired. □

Remark: The previous proof shows that if $x \sim \mathcal{N}(x_0, \Gamma_{\rm pr})$ and $\varepsilon \sim \mathcal{N}(\varepsilon_0, \Gamma_{\rm n})$, then

$$x|y \sim \mathcal{N}(\mu_{\rm post}, \Gamma_{\rm post}),$$

where

$$\Gamma_{\rm post} = (\Gamma_{\rm pr}^{-1} + A^{\rm T}\Gamma_{\rm n}^{-1}A)^{-1} \tag{2}$$

$$\mu_{\rm post} = \Gamma_{\rm post}(A^{\rm T}\Gamma_{\rm n}^{-1}(y - \varepsilon_0) + \Gamma_{\rm pr}^{-1}x_0). \tag{3}$$

One also has the following alternative representations for the posterior mean

$$\mu_{\rm post} = x_0 + \Gamma_{\rm pr}A^{\rm T}(A\Gamma_{\rm pr}A^{\rm T} + \Gamma_{\rm n})^{-1}(y - Ax_0 - \varepsilon_0) \tag{4}$$

and the posterior covariance

$$\Gamma_{\rm post} = \Gamma_{\rm pr} - \Gamma_{\rm pr}A^{\rm T}(A\Gamma_{\rm pr}A^{\rm T} + \Gamma_{\rm n})^{-1}A\Gamma_{\rm pr}. \tag{5}$$

Formula (5) can be proved, e.g., by using the Sherman–Morrison–Woodbury formula on (2). Formula (4) can be proved by plugging the formula (5) into (3) and simplifying the expression.

As the posterior distribution is Gaussian, its mean and its mode coincide. This means that the conditional mean estimator and the MAP estimator coincide in the linear Gaussian setting.

Corollary

The conditional mean estimator and the maximum a posteriori estimator coincide in the linear Gaussian setting and are given by $x_{\mathrm{CM}} = x_{\mathrm{MAP}} = \mu_{\mathrm{post}}.$

Example

Let $\Gamma_{\mathrm{n}} = \sigma^2 I$, $\varepsilon_0 = 0$, $\Gamma_{\mathrm{pr}} = \gamma^2 I$, $x_0 = 0$, and set $\lambda = \frac{\sigma}{\gamma}$. Then μ_{post} minimizes

$$J_\lambda(x) := \|y - Ax\|^2 + \lambda^2 \|x\|^2.$$

and therefore satisfies

$$(A^{\mathrm{T}} A + \lambda^2 I)\mu_{\mathrm{post}} = A^{\mathrm{T}} y. \tag{6}$$

This example provides a connection between Bayesian inference and variational regularization: J_λ can be interpreted as the objective functional in a linear regression model with a regularization term $\lambda^2 \|x\|^2$. Equation (6) for μ_{post} is then exactly the normal equation.

In the general case, the formula

$$\mu_{\mathrm{post}} = (\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}} \Gamma_{\mathrm{n}}^{-1} A)^{-1} (A^{\mathrm{T}} \Gamma_{\mathrm{n}}^{-1} (y - \varepsilon_0) + \Gamma_{\mathrm{pr}}^{-1} x_0)$$

can thus be viewed as the solution to a generalized normal equation. This point of view helps to understand the structure of Bayesian inference by linking it to well-understood optimization approaches.

Numerical example: one-dimensional deconvolution

Suppose that we are interested in estimating a signal $g\colon [0,1] \to \mathbb{R}$ from noisy, blurred observations modeled as

$$y_i = y(s_i) = \int_0^1 K(s_i, t) g(t) \, \mathrm{d}t + \varepsilon_i, \quad i \in \{1, \ldots, k\},$$

where the blurring kernel is

$$K(s,t) = \exp\left(-\frac{1}{2\omega^2}(s-t)^2 \right), \quad \omega = 0.5,$$

and we have Gaussian measurement noise $\varepsilon \sim \mathcal{N}(0, \Gamma_{\mathrm{noise}})$ with a symmetric, positive definite covariance matrix Γ_{noise}.

Discrete model

Midpoint rule:

$$y_i = \int_0^1 K(s_i, t) g(t) \, \mathrm{d}t + \varepsilon_i \approx \frac{1}{d} \sum_{j=1}^{d} K(s_i, t_j) x_j + \varepsilon_i,$$

where $t_j = \frac{j}{d} - \frac{1}{2d}$ and $x_j = g(t_j)$ for $j \in \{1, \ldots, d\}$.

If we have $s_i = \frac{i}{k} - \frac{1}{2k}$ for $i \in \{1, \ldots, k\}$, then we have the discrete linear model

$$y = Ax + \varepsilon, \quad \text{where } A_{i,j} = \frac{1}{d} K(s_i, t_j).$$

To employ the Bayesian approach, we treat y, ε, and x as random variables. We assume that ε is Gaussian noise with variance $\sigma^2 I$,

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I), \qquad \nu(\varepsilon) \propto \exp\Big(-\frac{1}{2\sigma^2}\|\varepsilon\|^2\Big).$$

The likelihood is then given by

$$f(y|x) = \nu(y - Ax) \propto \exp\Big(-\frac{1}{2\sigma^2}\|y - Ax\|^2\Big).$$

Next, we have to choose a prior distribution for the unknown. Assume that we know that $g(0) = g(1) = 0$ and that g is quite smooth, that is, the value of $g(t)$ in a point is more or less the same as in its neighbor. We will then model the unknown as

$$x_j = \frac{1}{2}(x_{j-1} + x_{j+1}) + W_j, \quad j = 1, \ldots, k, \tag{7}$$

where the term W_j follows a Gaussian distribution $\mathcal{N}(0, \gamma^2)$.
The variance γ^2 determines how much the reconstructed function x departs from the smoothness model $x_j = \frac{1}{2}(x_{j-1} + x_{j+1})$. We can write (7) as

$$Lx = W, \quad \text{where} \quad L := \frac{1}{2}\begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix}.$$

This leads to the so-called *smoothness prior*

$$f(x) \propto \exp\Big(-\frac{1}{2\gamma^2}\|Lx\|^2\Big).$$

Let $L = \mathrm{tridiag}(-1, 2, -1)$ and consider the following priors

$$\begin{aligned} f_{\mathrm{pr},1}(x) &\propto \exp\left(-\frac{1}{2\gamma^2}\|x - x_0\|^2 \right) \quad \text{with covariance } \Gamma_{\mathrm{pr},1} = \gamma^2 I; \\ f_{\mathrm{pr},2}(x) &\propto \exp\left(-\frac{1}{2\gamma^2}\|L(x - x_0)\|^2 \right) \\ &= \exp\left(-\frac{1}{2\gamma^2}(x - x_0)^{\mathrm{T}}(L^{\mathrm{T}}L)(x - x_0) \right) \quad \begin{array}{c} \text{with covariance} \\ \Gamma_{\mathrm{pr},2} = \gamma^2 (L^{\mathrm{T}}L)^{-1}, \end{array} \end{aligned}$$

where $x_0 \in \mathbb{R}^d$ is the prior mean (assumed to be the same in both cases). Hence

$$\begin{aligned} \overline{x_j} &= x_0 + \Gamma_{\mathrm{pr},j} A^{\mathrm{T}} G_j^{-1}(y - Ax_0 - \varepsilon_0), \\ \Gamma_{\mathrm{post},j} &= \Gamma_{\mathrm{pr},j} - \Gamma_{\mathrm{pr},j} A^{\mathrm{T}} G_j^{-1} A \Gamma_{\mathrm{pr},j}, \end{aligned}$$

where $G_j = A\Gamma_{\mathrm{pr},j}A^{\mathrm{T}} + \Gamma_{\mathrm{noise}}$ and $\Gamma_{\mathrm{noise}} = \sigma^2 I$.

For the numerical experiment, we simulate measurements using the (smooth) ground truth signal

$$g(t) = 8t^3 - 16t^2 + 8t,$$

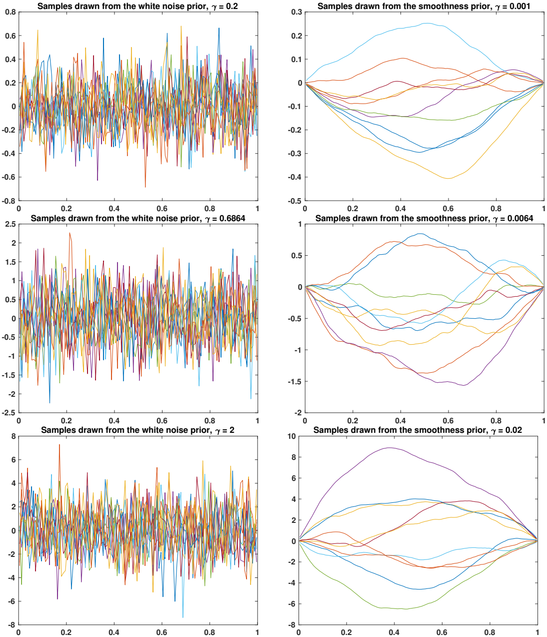
which satisfies $g(0) = g(1) = 0$. The measurements are contaminated with zero-mean 10% *relative* noise ($\sigma \approx 0.0618$) and we set $d = k = 120$.

Remark: We use a higher resolution model to simulate the measurement data. To achieve this, we generate the measurement data using a denser grid and then interpolate the forward solution onto a coarser computational grid, which is actually used to compute the reconstruction.

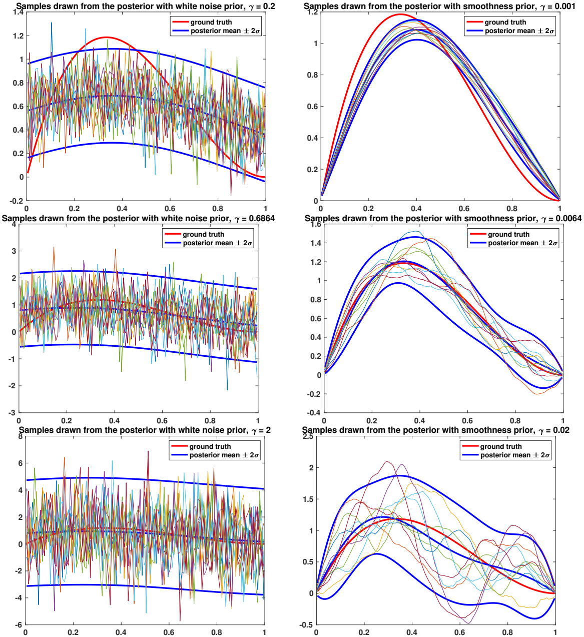
Since both the prior and the posterior are now Gaussian, we can use the coloring transformation to draw samples from the prior and posterior.

We also draw the posterior mean and the 2σ credibility envelopes.

See the script `deconv.py` on the course webpage.



Samples drawn from the white noise prior and the smoothness prior for several values of γ.



Samples drawn from the posterior corresponding to both the white noise prior and the smoothness prior for several values of γ. We also plot the ground truth solution and the posterior mean.

A note on marginal Gaussian distributions

Let

$$f(x) \propto \exp(-\tfrac{1}{2}(x-\mu)^{\mathrm{T}}\Gamma^{-1}(x-\mu))$$

be a multivariate Gaussian PDF with mean μ and positive definite and symmetric covariance matrix Γ.

Q: What is Γ_{ii}?
A: $\sigma_i^2 := \Gamma_{ii}$ is the variance of the marginal distribution with PDF

$$f(x_i) = \int_{\mathbb{R}^{n-1}} f(x_1, \ldots, x_i, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_{i-1}\, \mathrm{d}x_{i+1} \cdots \mathrm{d}x_n,$$

which is itself a (univariate) Gaussian PDF with mean μ_i.

This is why we can obtain the credibility envelopes by taking the square roots of the diagonal values of $\Gamma_{\mathrm{post},j}$.

Relation to conjugate priors

The linear Gaussian setting is a special case of a more general technique, where the prior is chosen in such a way that, together with the likelihood function, the resulting posterior density belongs to the same probability distribution family as the prior. In this case, the prior and posterior are then called *conjugate distributions*, and the prior is called a *conjugate prior* for the likelihood function.

A conjugate prior is an algebraic convenience, giving a closed form expression for the posterior. In consequence, the CM estimator, MAP estimator, and variance typically also have closed form expressions and it is not necessary to use numerical integration or optimization to characterize the statistics of the posterior.

A list of the most commonly used conjugate priors can be found, e.g., at https://en.wikipedia.org/wiki/Conjugate_prior

Kalman filter

So far we have discussed measurement models with a *static target*:

$$y_j = F(x) + \varepsilon_j, \quad \varepsilon_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \gamma^2 I).$$

Examples where the condition may not be valid:

- EEG
- Target tracking
- Weather forecasting

Dynamic observation models

More general observation model:

$$y_j = F(x_j) + \varepsilon_j, \quad j = 1, 2, \ldots, J.$$

The observations cannot be integrated unless we have a *dynamic prior model*.

One of the simplest dynamic prior models is a 1-Markov evolution model

$$x_{j+1} = G(x_j) + \xi_{j+1}, \quad j = 0, 1, \ldots, J-1,$$

where $G : \mathbb{R}^d \to \mathbb{R}^d$ is presumably known and ξ_{j+1} is an *innovation process*.

Examples

- Static measurement: $G(x) = x$, $\xi_{j+1} = 0$.
- Random walk model (often used in lack of anything more sophisticated):

$$x_{j+1} = x_j + \xi_{j+1}, \quad \xi_{j+1} \sim \mathcal{N}(0, \sigma^2 I).$$

- First order differential equation: assume that the unknown is a time-dependent vector $x(t) \in \mathbb{R}^d$ satisfying *ideally* the differential equation

$$x'(t) = f(x(t), t).$$

Time discretization: let $t_j = jh$, $j = 0, 1, \ldots$, and write $x_j = x(t_j)$. Then we can use finite differences, e.g., forward Euler method

$$x_{j+1} = x_j + hf(x_j, t_j) + \xi_{j+1}$$

or backward Euler method

$$x_{j+1} = x_j + hf(x_{j+1}, t_{j+1}) + \xi_{j+1},$$

where ξ_{j+1} accounts for discretization errors as well as possible deviations from the ideal.

Basic form of Bayes filtering

Evolution-observation model:

$$\begin{aligned} x_{j+1} &= G(x_j) + \xi_{j+1}, \quad j = 0, 1, \ldots, J-1, \\ y_{j+1} &= F(x_{j+1}) + \varepsilon_{j+1}, \quad j = 0, 1, \ldots, J-1. \end{aligned}$$

The observations $y_1, \ldots, y_J$ and the prior probability density of x_0 are given.

Adaptive algorithm

The goal is an algorithm which works as follows:

- Given the density of x_0, *predict* the density of x_1 using the prior evolution model.
- Using the predicted density of x_1 as *prior*, calculate the posterior density of $x_1|y_1$.
- Using the posterior density of $x_1|y_1$, predict the density of x_2.
- Using the predicted density of x_2 as *prior*, calculate the posterior density of $x_2|y_1, y_2$.
- Continue similarly.

$$\begin{tikzcd}
 & f(x_0) \arrow[d] & \\
f(x_1|x_0) \arrow[r] & \text{evolution updating} \arrow[d] & \\
 & f(x_1) \arrow[d] & \\
 & \text{observation updating} \arrow[d] & f(y_1|x_1) \arrow[l] \\
 & f(x_1|y_1) \arrow[d] & \\
f(x_2|x_1) \arrow[r] & \text{evolution updating} \arrow[d] & \\
 & f(x_2|y_1) \arrow[d] & \\
 & \text{observation updating} \arrow[d] & f(y_2|x_2) \arrow[l] \\
 & f(x_2|y_1, y_2) \arrow[d] & \\
 & \vdots &
\end{tikzcd}$$

- **Prediction step:** Given the density of x_j, calculate the density of x_{j+1} from

$$x_{j+1} = G(x_j) + \xi_{j+1}. \qquad \text{(propagation problem)}$$

- **Correction step:** Given the prior density of x_{j+1}, calculate the posterior density of $x_{j+1}|y_{j+1}$ using the observational model

$$y_{j+1} = F(x_{j+1}) + \varepsilon_{j+1}. \qquad \text{(inference problem)}$$

Particular approaches

- Linear model, Gaussian innovation and error: classical Kalman filtering.
- Linearization of non-linear evolution (or observation) model: extended Kalman filtering.
- Nonlinear and/or non-Gaussian models: particle filtering.

Kalman filter

Consider the linear ($G(\cdot) = M\cdot$, $F(\cdot) = H\cdot$) evolution-observation system

$$\begin{aligned} x_{j+1} &= Mx_j + \xi_{j+1}, \quad \xi_{j+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma), \\ y_{j+1} &= Hx_{j+1} + \varepsilon_{j+1}, \quad \varepsilon_{j+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Gamma). \end{aligned}$$

Prediction: Suppose $x_j \sim \mathcal{N}(m_j, C_j)$. Then

$$x_{j+1} = Mx_j + \xi_{j+1} \sim \mathcal{N}(\widehat{m}_{j+1}, \widehat{C}_{j+1}),$$

where $\widehat{m}_{j+1} = Mm_j$ and $\widehat{C}_{j+1} = MC_jM^{\mathrm{T}} + \Sigma$.

Correction: Linear Gaussian setting implies $x_{j+1}|y_{j+1} \sim \mathcal{N}(m_{j+1}, C_{j+1})$, where

$$\begin{aligned} m_{j+1} &= \widehat{m}_{j+1} + \widehat{C}_{j+1}H^{\mathrm{T}}(H\widehat{C}_{j+1}H^{\mathrm{T}} + \Gamma)^{-1}(y_{j+1} - H\widehat{m}_{j+1}), \\ C_{j+1} &= \widehat{C}_{j+1} - \widehat{C}_{j+1}H^{\mathrm{T}}(H\widehat{C}_{j+1}H^{\mathrm{T}} + \Gamma)^{-1}H\widehat{C}_{j+1}. \end{aligned}$$

Remark: The expensive step in Kalman filtering is the computation of the so-called *Kalman gain* matrix:

$$K_{j+1} = \widehat{C}_{j+1}H^{\mathrm{T}}(H\widehat{C}_{j+1}H^{\mathrm{T}} + \Gamma)^{-1}.$$

Kalman filter algorithm

```
Given: Initial distribution for $x_0 \sim \mathcal{N}(m_0, C_0)$, where $m_0 \in \mathbb{R}^d$ and
$C_0 \in \mathbb{R}^{d \times d}$ is symmetric and positive definite.
for $j = 0, 1, 2, \ldots, J-1$, do
    Prediction step:
        $\widehat{m}_{j+1} = M m_j$
        $\widehat{C}_{j+1} = M C_j M^{\mathrm{T}} + \Sigma$
    Correction step:
        $K_{j+1} = \widehat{C}_{j+1} H^{\mathrm{T}} (H \widehat{C}_{j+1} H^{\mathrm{T}} + \Gamma)^{-1}$
        $m_{j+1} = \widehat{m}_{j+1} + K_{j+1}(y_{j+1} - H \widehat{m}_{j+1})$
        $C_{j+1} = \widehat{C}_{j+1} - K_{j+1} H \widehat{C}_{j+1}$
end for
Output: Predicted distributions $\mathcal{N}(\widehat{m}_{j+1}, \widehat{C}_{j+1})$ and filtering distributions
for $x_{j+1} | y_1, \ldots, y_{j+1} \sim \mathcal{N}(m_{j+1}, C_{j+1})$, $j = 0, \ldots, J-1$.
```

Extended Kalman filter (non-linear evolution model)

Consider non-linear $G \colon \mathbb{R}^d \to \mathbb{R}^d$ and linear $F(\cdot) = H\cdot$ with

$$\begin{aligned} x_{j+1} &= G(x_j) + \xi_{j+1}, \quad \xi_{j+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma), \\ y_{j+1} &= Hx_{j+1} + \varepsilon_{j+1}, \quad \varepsilon_{j+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Gamma), \end{aligned}$$

with $x_0 \sim \mathcal{N}(m_0, C_0)$.

Prediction: Suppose $x_j \sim \mathcal{N}(m_j, C_j)$. We can linearize

$$x_{j+1} = G(x_j) + \xi_{j+1} \approx G(m_j) + DG(m_j)(x_j - m_j) + \xi_{j+1}.$$

An affine transformation is still Gaussian, so we obtain the approximations

$$\widehat{m}_{j+1} = G(m_j), \; \widehat{C}_{j+1} = DG(m_j)C_jDG(m_j)^{\mathrm{T}} + \Sigma.$$

Correction: Now that $x_{j+1} \sim \mathcal{N}(\widehat{m}_{j+1}, \widehat{C}_{j+1})$, we can use the linear Gaussian setting to obtain $x_{j+1}|y_{j+1} \sim \mathcal{N}(m_{j+1}, C_{j+1})$ with

$$\begin{aligned} m_{j+1} &= \widehat{m}_{j+1} + \widehat{C}_{j+1}H^{\mathrm{T}}(H\widehat{C}_{j+1}H^{\mathrm{T}} + \Gamma)^{-1}(y_{j+1} - B\widehat{m}_{j+1}), \\ C_{j+1} &= \widehat{C}_{j+1} - \widehat{C}_{j+1}H^{\mathrm{T}}(H\widehat{C}_{j+1}H^{\mathrm{T}} + \Gamma)^{-1}H\widehat{C}_{j+1}. \end{aligned}$$

Ensemble Kalman filter (non-linear evolution model)

Consider

$$x_{j+1} = G(x_j) + \xi_{j+1}, \quad \xi_{j+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma),$$
$$y_{j+1} = Hx_{j+1} + \varepsilon_{j+1}, \quad \varepsilon_{j+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Gamma),$$

with $x_0 \sim \mathcal{N}(m_0, C_0)$.

The computation of the analytical predictive covariances and (in the non-linear setting) the Jacobi matrix become computationally inefficient and expensive for high-dimensional systems. The basic idea of ensemble Kalman filter is as follows:

1. Draw a sample from the initial distribution of x_0 ("initial ensemble")
2. Replace the predictive mean $\widehat{m}_{j+1}$ and covariance $\widehat{C}_{j+1}$ as well as the filtering mean m_{j+1} and covariance C_{j+1} with their corresponding sample means and covariances by propagating the initial ensemble through the evolution-observation model.

Ensemble Kalman filter algorithm

```
Given: Ensemble size $N$. Initial ensemble $\{x_0^{(i)}\}_{i=1}^N$ drawn from the initial distribution of
$x_0 \sim \mathcal{N}(m_0, C_0)$, where $m_0 \in \mathbb{R}^d$ and $C_0 \in \mathbb{R}^{d \times d}$ is symmetric and positive definite.
Parameter $s \in \{0, 1\}$.
for $j = 0, 1, 2, \dots, J-1$, do
    Prediction step:
        draw $\xi_{j+1}^{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma)$,  $i = 1, \dots, N$,
        $\widehat{x}_{j+1}^{(i)} = G(x_j^{(i)}) + \xi_{j+1}^{(i)}$,    $i = 1, \dots, N$,
        $\widehat{m}_{j+1} = \frac{1}{N} \sum_{i=1}^N \widehat{x}_{j+1}^{(i)}$   and   $\widehat{C}_{j+1} = \frac{1}{N} \sum_{i=1}^N (\widehat{x}_{j+1}^{(i)} - \widehat{m}_{j+1})(\widehat{x}_{j+1}^{(i)} - \widehat{m}_{j+1})^{\mathrm{T}}$.
    Correction step:
        draw $\varepsilon_{j+1}^{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Gamma)$,    $i = 1, \dots, N$,
        $y_{j+1}^{(i)} = y_{j+1} + s\varepsilon_{j+1}^{(i)}$,    $i = 1, \dots, N$,
        $K_{j+1} = \widehat{C}_{j+1} H^{\mathrm{T}} (H \widehat{C}_{j+1} H^{\mathrm{T}} + \Gamma)^{-1}$,
        $x_{j+1}^{(i)} = \widehat{x}_{j+1}^{(i)} + K_{j+1}(y_{j+1}^{(i)} - H\widehat{x}_{j+1}^{(i)})$,    $i = 1, \dots, N$.
end for
Output: Ensembles $\{x_j^{(i)}\}_{i=1}^N$, $j = 0, \dots, J$.
```

Remark:

- Setting the parameter $s = 1$ is suitable at approximating the Kalman filter in linear Gaussian settings: if each prediction particle $\widetilde{x}_{j+1}^{(i)}$ is distributed according to a non-degenerate Gaussian distribution, then in the linear Gaussian setting the "corrected" particle $x_{j+1}^{(i)}$ will be Gaussian with mean and covariance that agree with the usual Kalman filter formulae.
- Setting the parameter $s = 0$ is natural if viewing the algorithm as a sequential optimizer in problems where the filtering distributions are not well approximated by Gaussians.

Numerical example

We wish to track the state $x_k = \begin{bmatrix} p_k \\ v_k \end{bmatrix} \in \mathbb{R}^2$ of a moving particle at discrete times $t_k = k\Delta t$, $k = 0, 1, 2, \ldots$. The first component p_k corresponds to the position of the particle while the second component $v_k = \dot{p}_k$ is its velocity at time $k = 0, 1, 2, \ldots$. Let us also denote the *unknown* acceleration of the particle as $a_k = \dot{v}_k = \ddot{p}_k$ for $k = 0, 1, 2, \ldots$. We have the following dynamics:

$$\begin{cases} p_k = p_{k-1} + v_{k-1}\Delta t + \frac{1}{2}a_{k-1}(\Delta t)^2 \\ v_k = v_{k-1} + a_{k-1}\Delta t \end{cases} \Leftrightarrow x_k = \underbrace{\begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix}}_{=:M} x_{k-1} + \begin{bmatrix} \frac{1}{2}(\Delta t)^2 \\ \Delta t \end{bmatrix} a_{k-1}.$$

If we *assume* that $a_{k-1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, then the innovation process is $\xi_k := \begin{bmatrix} \frac{1}{2}(\Delta t)^2 \\ \Delta t \end{bmatrix} a_{k-1} \sim \mathcal{N}(0, \Sigma)$, where $\Sigma := \begin{bmatrix} \frac{1}{2}(\Delta t)^2 \\ \Delta t \end{bmatrix} \begin{bmatrix} \frac{1}{2}(\Delta t)^2 & \Delta t \end{bmatrix}$ $= \begin{bmatrix} \frac{1}{4}(\Delta t)^4 & \frac{1}{2}(\Delta t)^3 \\ \frac{1}{2}(\Delta t)^3 & (\Delta t)^2 \end{bmatrix}$. This yields the *evolution model*

$$x_k = Mx_{k-1} + \xi_k, \quad \xi_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma).$$

Meanwhile, we can only measure the location of the particle, so the *observation model* is given by

$$y_k = Hx_k + \varepsilon_k, \quad \varepsilon_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \gamma^2),$$

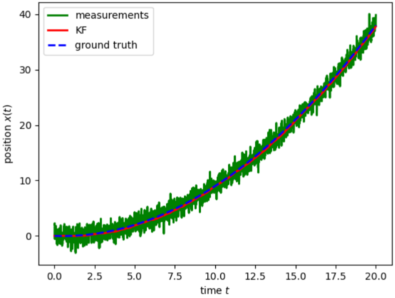
where $H := \begin{bmatrix} 1 & 0 \end{bmatrix}$ and y_k is a noisy measurement of the particle's location at time k, with the noise level assumed to be $\gamma := 1$.

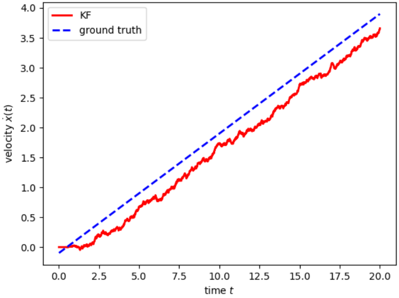
We can now implement the Kalman filter for this model problem. We can assume that the initial position of the particle is perfectly known:

$$x_0 = \mathbb{E}[x_0] = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad C_0 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{2\times 2}.$$

The Kalman filter can be used to obtain the mean and covariance of the (Gaussian) filtering distribution $(p_k, v_k)|y_1, \ldots, y_k$ for $k = 1, 2, 3, \ldots$. We can plot the means of the filtered positions $(t_k, \mathbb{E}[p_k|y_1, \ldots, y_k])$ and velocities $(t_k, \mathbb{E}[v_k|y_1, \ldots, y_k])$.

The implementation is left as an exercise.





The true trajectory of the particle was $x(t) = 0.1(t^2 - t)$ (left figure) with velocity $\dot{x}(t) = 0.2t - 0.1$ (right figure). The observations at time points $t_k = k\Delta t$, with $\Delta t = 0.01$ and $k = 1, 2, \ldots$, were contaminated with centered, unit-variance Gaussian noise (left figure). The red graphs correspond to the means of the filtered positions $(t_k, \mathbb{E}[p_k|y_1, \ldots, y_k])$ (left figure) and velocities $(t_k, \mathbb{E}[v_k|y_1, \ldots, y_k])$ (right figure).

Remarks:

- The Kalman filter is optimal in the sense that it gives the best estimator of the mean in an online setting.
- In the linear case ($G(\cdot) = M\cdot$), the ensemble Kalman filter converges to the Kalman filter. When applicable, the ensemble Kalman filter is much more efficient than particle filters. A primary advantage of ensemble methods is that they can provide good state estimation even when the number of particles is *not* large.

Appendix: General evolution-observation model and particle filters

General evolution-observation model and particle filters

Consider the more general model

$$\begin{aligned} x_{j+1} &= G(x_j, \xi_{j+1}), \quad j = 0, 1, \ldots, J-1, \\ y_{j+1} &= F(x_{j+1}, \varepsilon_{j+1}), \quad j = 0, 1, \ldots, J-1. \end{aligned}$$

The functions F and G are assumed to be known. We also assume that $\xi_{j+1} \perp x_j$ and $\varepsilon_{j+1} \perp x_{j+1}$.

Observation and evolution models may be cumbersome or impossible to linearize (e.g., non-differentiable or no closed form). One may try Monte Carlo methods to simulate the distributions by random samples.

The goal in *particle filter* methods is to produce sequentially an ensemble of random samples $\{x_j^{(1)}, \ldots, x_j^{(N)}\}$ distributed according to the conditional probability distributions $f(x_{j+1}|y_1, \ldots, y_j)$ (prediction) or $f(x_j|y_1, \ldots, y_j)$ (filtering). The vectors $x_j^{(i)}$ are called *particles* of the sample, hence the name particle filter.

One straightforward particle filter method is known as the *sampling importance resampling* filter (also known as *SIR* or *bootstrap filter*).

Sampling importance resampling

1. Set $j = 0$ and generate an initial sample $S_0 = \{x_0^{(i)}\}_{i=1}^N$ by drawing from the density $f(x_0)$. (This may require MCMC if the initial density is complicated, e.g., non-Gaussian.)
2. **Prediction:** Draw $\xi_{j+1}^{(i)}$ from the distribution of ξ_{j+1} and set $\widehat{x}_{j+1}^{(i)} = G(x_j^{(i)}, \xi_{j+1}^{(i)})$ for $1 \leq i \leq N$. Let $\widehat{S}_{j+1} = \{\widehat{x}_{j+1}^{(i)}\}_{i=1}^N$.
3. **Correction:** Assume that from the observational model $y_j = F(x_j, \varepsilon_j)$, we can calculate the likelihood density $Cf(y_j|x_j)$, $j = 1, 2, \ldots, J$, up to a multiplicative constant $C > 0$.[†] Calculate the importance of each propagated particle

   $$\widehat{w}_{j+1}^{(i)} = Cf(y_{j+1}|\widehat{x}_{j+1}^{(i)}), \quad 1 \leq i \leq N,$$

   and compute their *relative importance*

   $$w_{j+1}^{(i)} = \frac{\widehat{w}_{j+1}^{(i)}}{W}, \quad W = \sum_{i=1}^N \widehat{w}_{j+1}^{(i)}.$$

   Resampling: draw a new sample $S_{j+1} = \{x_{j+1}^{(1)}, \ldots, x_{j+1}^{(N)}\}$ from the sample $\widehat{S}_{j+1}$, with the probability of drawing $\widehat{x}_{j+1}^{(i)}$ set equal to $w_{j+1}^{(i)}$. Set $j \leftarrow j + 1$ and return to step 2.

[†] E.g., if $y_j = F(x_j) + \varepsilon_j$, $\varepsilon_j \sim \mathcal{N}(0, \gamma^2 I)$, then $f(y_j|x_j) \propto \exp(-\frac{1}{2\gamma^2}\|y_j - F(x_j)\|^2)$.