5. Extracting Estimates from PDFs

We've seen examples of how to manipulate probability distributions, but how do we extract an estimate from them? In this lecture we will focus exclusively on continuous random variables. Define:

x : unknown quantity of interest

z : observation related to x.

5.1 Maximum likelihood (ML)

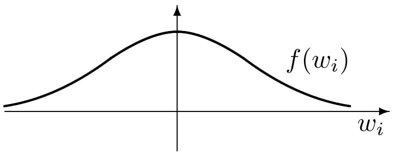
- Assumes that we can calculate $f(z|x)$.
- $\hat{x}^{ML} := \arg\max_x f(z|x)$, with $f(z|x)$, the conditional PDF, often called the *likelihood function.*
- What choice of parameters makes the observed measurements the most likely ones?

Example

$$z_1 = x + w_1$$
$$z_2 = x + w_2$$

w_1, w_2 are normally distributed, with zero mean, unit variance, independent:

$$f(w_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{w_i^2}{2}\right)$$
$$w_i \sim \mathcal{N}(0,1).$$



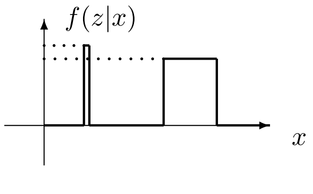
Then

$$f(z_1, z_2|x) = f(z_1|x)\, f(z_2|x)\,, \text{ since conditionally independent.}$$
$$f(z_i|x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z_i - x)^2}{2}\right)$$
$$\therefore f(z_1, z_2|x) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\left((z_1 - x)^2 + (z_2 - x)^2\right)\right)$$

Differentiate with respect to x and set to 0: $(z_1 - \hat{x}) + (z_2 - \hat{x}) = 0$, $\hat{x} = \dfrac{z_1 + z_2}{2}$.

- (**PSET 3: P1**) $w_i \sim \mathcal{N}(0, \sigma_i^2)$, independent.
- (**PSET 3: P2**) w_1, w_2 uniformly distributed on $[-1, 1]$, independent.

Why this is not always a good thing to do:



Date compiled: March 22, 2012

Example

$$z = Hx + w \qquad \text{with } z, w \in \mathbb{R}^m,\ x \in \mathbb{R}^n,\ \ m > n,\ \ w_i \sim \mathcal{N}(0,1),\ \text{independent,}$$

$$H = \begin{bmatrix} H_1 \\ H_2 \\ \vdots \\ H_m \end{bmatrix},\ \ H_i = \begin{bmatrix} h_{i1} & \dots & h_{in} \end{bmatrix},\ h_{ij} \in \mathbb{R}\ .$$

As before,

$$f(z|x) \propto \exp\left(-\frac{1}{2}\left((z_1 - H_1 x)^2 + \ldots + (z_m - H_m x)^2\right)\right)\ .$$

Differentiating with respect to x_j and setting to 0 gives:

$$(z_1 - H_1\hat{x})h_{1j} + (z_2 - H_2\hat{x})h_{2j} + \ldots + (z_m - H_m\hat{x})h_{mj} = 0, \qquad j = 1, \ldots, n$$

$$[h_{1j}\ h_{2j}\ \ldots\ h_{mj}]\,(z - H\hat{x}) = 0, \qquad j = 1, \ldots, n$$

$$H^T(z - H\hat{x}) = 0,\ \ H^T H\hat{x} = H^T z,\ \ \hat{x} = (H^T H)^{-1}H^T z$$

This is the least squares solution!

- We can thus give least squares a statistical interpretation: the maximum likelihood estimate when the errors are independent, zero mean, *same* variance and normally distributed.
- Recall the "standard" interpretation:

  $$w(x) = z - Hx, \quad \hat{x} = \arg\min_x\ w^T(x)w(x)\ .$$

- $w_i \sim \mathcal{N}(0, \sigma_i^2)$ results in *weighted* least squares (**PSET 3: P3**).

5.2 Maximum a posteriori (MAP)

We use MAP when x is a random variable with a known PDF. We already know what to do:

$$f(x|z) = \frac{f(z|x)\,f(x)}{f(z)}$$

$$\hat{x}^{MAP} := \arg\max_x f(z|x)\,f(x)$$

What choice of parameters are the most likely ones, given the observations and the prior belief about x? Note: if $f(x)$ is constant, $\hat{x}^{MAP} = \hat{x}^{ML}$, i.e. if all values of x are a priori equally likely, the estimates coincide.

Example

Define

$$z = x + w,\ \text{scalar},\ w \sim \mathcal{N}(0,1),\ \ x \sim \mathcal{N}(\overline{x}, \sigma_x^2);\ x \text{ and } w \text{ are independent.}$$

$$f(z|x) \propto \exp\left(-\frac{1}{2}(z - x)^2\right),\ \ f(x) \propto \exp\left(-\frac{1}{2}\frac{(x - \overline{x})^2}{\sigma_x^2}\right)$$

Differentiate $f(x|z)$ with respect to x and set to 0:

$$\hat{x} = \frac{\overline{x}}{1 + \sigma_x^2} + \frac{\sigma_x^2}{1 + \sigma_x^2}z, \quad \text{a weighted sum.}$$

$$\sigma_x^2 = 0: \qquad \hat{x} = \overline{x}\ \text{(ML of prior)}$$

$$\sigma_x^2 \to \infty: \quad \hat{x} = z\ \text{(ML)}$$

PSET 3: P4: Show steps.

5.3 Minimum mean squared error (MMSE)

A posteriori estimate that minimizes the mean squared error.

$$\begin{aligned}\hat{x}^{MMSE} &:= \arg\min_{\hat{x}} \underset{x}{\mathrm{E}}[(\hat{x}-x)^T(\hat{x}-x)|z] \\ &= \arg\min_{\hat{x}} \left(\hat{x}^T\hat{x} - 2\hat{x}^T\mathrm{E}[x|z] + \mathrm{E}[x^Tx|z]\right)\end{aligned}$$

Differentiate with respect to $\hat{x}$ and set to 0:

$$\hat{x}^{MMSE} = \mathrm{E}[x|z]$$

This is simply the expected value conditioned on z. Compare this with the MAP estimate.

5.4 Recursive least squares

- A prelude to the standard way that the Kalman Filter is derived.
- Bypasses PDFs and works directly with mean and variance.

Model: $z(k) = H(k)x + w(k)$

- x: $\overline{x} := \mathrm{E}[x]$, $P_x := \mathrm{E}[(x-\overline{x})(x-\overline{x})^T] = \mathrm{Var}[x]$ are given. This is the prior knowledge.
- $w(k)$: $\mathrm{E}[w(k)] = 0$, $R(k) := \mathrm{Var}[w(k)]$ are given. This is the measurement noise model.
- $\{x, w(1), ...\}$ are independent.
- No process model: x does not change (our knowledge of it changes, however!)
- Typically, $\dim(x) > \dim(z(k))$, i.e. less equations than unknowns at any particular time.
- After collecting a whole bunch of data, we can convert this to a standard, weighted least squares problem:

$$z = \begin{bmatrix} z(1) \\ \vdots \\ z(k) \end{bmatrix}, \quad H = \begin{bmatrix} H(1) \\ \vdots \\ H(k) \end{bmatrix}, \quad w = \begin{bmatrix} w(1) \\ \vdots \\ w(k) \end{bmatrix}, \quad z = Hx + w$$

- Can we build an estimate in real-time? **Assume** the following form for the estimate at time k:

$$\hat{x}(k) = \hat{x}(k\text{-}1) + K(k)\Big(z(k) - H(k)\hat{x}(k\text{-}1)\Big)$$

  where

  $\hat{x}(k)$ is the estimate of x at time k, and
  $K(k)$ is the gain matrix at time k, the only design variable.

- Intuitive: $z(k) = H(k)\hat{x}(k\text{-}1) \Rightarrow \hat{x}(k) = \hat{x}(k\text{-}1)$
- Let $e(k) := x - \hat{x}(k)$ be the estimate error. Then

$$\begin{aligned}e(k) &= x - \hat{x}(k\text{-}1) - K(k)\Big(H(k)x + w(k) - H(k)\hat{x}(k\text{-}1)\Big) \\ &= e(k\text{-}1) - K(k)H(k)e(k\text{-}1) - K(k)w(k) \\ &= \Big(I - K(k)H(k)\Big)e(k\text{-}1) - K(k)w(k)\end{aligned}$$

$$\mathrm{E}[e(k)] = \Big(I - K(k)H(k)\Big)\mathrm{E}[e(k\text{-}1)]$$

  Therefore, if we set $\hat{x}(0) = \overline{x}$, then $\mathrm{E}[e(0)] = 0$, $\mathrm{E}[e(k)] = 0 \; \forall \; k$. Independent of $K(k)$! **Unbiased estimator.**

- Choosing $K(k)$: Minimize $\mathrm{E}[e^T(k)e(k)]$ (MMSE):

$$J(k) := \mathrm{E}[e^T(k)e(k)] = \mathrm{E}[\mathrm{trace}\big(e(k)e^T(k)\big)] = \mathrm{trace}\big(P(k)\big)$$

  where $\mathrm{trace}(\cdot)$ is the sum of the diagonal terms, and $P(k) := \mathrm{Var}[e(k)]$.

- One can show (**PSET 3: P7**) that

$$\begin{aligned} P(k) &= \Big(I - K(k)H(k)\Big)P(k\text{–}1)\Big(I - K(k)H(k)\Big)^T + K(k)R(k)K^T(k) \\ &= P(k\text{–}1) - K(k)H(k)P(k\text{–}1) - P(k\text{–}1)H^T(k)K^T(k) \\ &\qquad\qquad +K(k)\Big(H(k)P(k\text{–}1)H^T(k) + R(k)\Big)K^T(k) \end{aligned}$$

  Note the $K(k)$ trade-off.

- We want to minimize $J(k) = \mathrm{trace}(P(k))$, as a function of $K(k)$. A necessary condition is that the derivatives with respect to each element of $K(k)$ are 0.

- We need the following results:

$$\frac{\partial\, \mathrm{trace}(ABA^T)}{\partial A} = 2AB \text{ if } B = B^T \quad (\textbf{PSET 3: P5}\text{: } 2\times 2 \text{ case})$$

$$\frac{\partial\, \mathrm{trace}(AB)}{\partial A} = B^T \quad (\textbf{PSET 3: P6}\text{: } 2\times 2 \text{ case})$$

$$\mathrm{trace}(C) = \mathrm{trace}(C^T) \quad \therefore \frac{\partial\, \mathrm{trace}(B^TA^T)}{\partial A} = B^T.$$

  Therefore the necessary condition $\dfrac{\partial J(k)}{\partial K(k)} = 0$ results in

$$K(k)\Big(H(k)P(k\text{–}1)H^T(k) + R(k)\Big) = P(k\text{–}1)H^T(k)$$

$$K(k) = P(k\text{–}1)H^T(k)\Big(H(k)P(k\text{–}1)H^T(k) + R(k)\Big)^{-1} \; (\textbf{PSET 3: P8})\ .$$

Summary

1. Initialize. $\hat{x}(0) = \overline{x}$, $P(0) = P_x = \mathrm{Var}[x]$

2. Observe. $z(k) = H(k)x + w(k)$

   Update. $K(k) = P(k\text{–}1)H^T(k)\Big(H(k)P(k\text{–}1)H^T(k) + R(k)\Big)^{-1}$ (can be pre-computed)

   $\hat{x}(k) = \hat{x}(k\text{–}1) + K(k)\Big(z(k) - H(k)\hat{x}(k\text{–}1)\Big)$ (real time)

   $P(k) = \Big(I - K(k)H(k)\Big)P(k\text{–}1)\Big(I - K(k)H(k)\Big)^T + K(k)R(k)K^T(k)$ (can be pre-computed)