5. Extracting Estimates from Probability Distributions

We have discussed examples of how to manipulate probability distributions. For example, we computed probability density functions of an unknown quantity of interest (denoted by x) conditioned on given observations (denoted by z) that are related to this quantity. While the conditional PDF $f(x|z)$ captures the full information that one has about x given z in the Bayesian sense, you may be interested in just an estimate $\hat{x}$ of the random variable x (that is, a real number for a scalar CRV, for example, or a vector in $\mathbb{R}^n$ for a vector-valued CRV). In this lecture, we discuss different ways to extract estimates from PDFs. We mostly focus on CRVs, but the concepts carry over to DRVs.

5.1 Maximum Likelihood (ML)

This method is often used when $x \in \mathcal{X}$ is an unknown (constant) parameter without a (known) probabilistic description. For given observation z and observation model $f(z|x)$, the method seeks the value for the parameter x that makes the observation z most likely; that is,

$$\hat{x}^{\mathrm{ML}} := \arg\max_{x \in \mathcal{X}} f(z|x)\,.$$

In this context, $f(z|x)$ as a function of x is often called the *likelihood function*.

Example

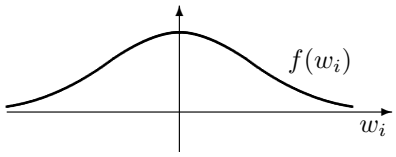
Consider two measurements of a scalar quantity $x \in \mathbb{R}$:

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

where w_1 and w_2 are two normally distributed, independent CRVs with zero mean and unit variance; that is,

$$f(w_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{w_i^2}{2}\right)$$

(Shorthand: $w_i \sim \mathcal{N}(0,1)$).



Since z_1 and z_2 are conditionally independent given x (given x, z_1 only depends on w_1, z_2 only depends on w_2, and w_1 and w_2 are independent),

$$f(z_1, z_2|x) = f(z_1|x)\, f(z_2|x)\ ,$$
$$f(z_i|x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z_i - x)^2}{2}\right)\ ,$$

and, therefore,

$$f(z_1, z_2|x) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\left((z_1 - x)^2 + (z_2 - x)^2\right)\right).$$

Date compiled: March 26, 2013

Differentiating with respect to x and setting to 0 yields: $(z_1 - \hat{x}) + (z_2 - \hat{x}) = 0$ and $\hat{x} = \frac{z_1+z_2}{2}$. That is, the ML estimate is the average (in this example).

Variations:

- $w_i \sim \mathcal{N}(0, \sigma_i^2)$, independent. (**PSET 3: P3**)
- w_1, w_2 uniformly distributed, independent. (**PSET 3: P4**)

Example (Generalization)

We generalize the previous example to m measurements:

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

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

and H is assumed to have full column rank.

- To compute $f(z|x)$ from $z = Hx+w$ and the PDF of w, we use a multivariable version of the change of variables formula for CRVs. While we could proceed just as in the previous example by arguing that the z_i's are conditionally independent given x and then computing $f(z_i|x)$ using the scalar change of variables formula, we present a different way here, which is also useful in other settings (for example, when the noise is not independent (**PSET 3: P5**)).

  Multivariable change of variables for CRVs
  Let g be a function mapping $y \in \mathbb{R}^n$ to $x \in \mathbb{R}^n$, $x = g(y)$, and assume that the determinant of the Jacobian matrix $\frac{\partial g}{\partial y}$ is nonzero for all y; that is,

  $$\det\left(\frac{\partial g}{\partial y}(y)\right) = \det\left(\begin{bmatrix} \frac{\partial g_1}{\partial y_1}(y) & \dots & \frac{\partial g_1}{\partial y_n}(y) \\ \vdots & & \vdots \\ \frac{\partial g_n}{\partial y_1}(y) & \dots & \frac{\partial g_n}{\partial y_n}(y) \end{bmatrix}\right) \neq 0 \quad \text{for all } y,$$

  where $\det(\cdot)$ is the determinant. Furthermore, assume that $x = g(y)$ has a unique solution for y in terms of x, say $y = h(x)$. Then:

  $$f_x(x) = f_y(h(x)) \left|\det\left(\frac{\partial g}{\partial y}(h(x))\right)\right|^{-1}.$$

  Note that the change of variables formula for a scalar CRV (see lecture #2) can be recovered from this.

- We apply the multivariable change of variables formula with $z = g(x, w) = Hx+w$ and $w = h(z, x) = z - Hx$. Since $\det(\frac{\partial g}{\partial w}) = 1$, we obtain

  $$\begin{aligned} f(z|x) &= f_w(z - Hx) && \text{(by change of variables)} \\ &= f_{w_1}(z_1 - H_1x) \cdot \ldots \cdot f_{w_m}(z_m - H_mx) && \text{(by independence of } w_i\text{)} \\ &\propto \exp\left(-\frac{1}{2}\left((z_1 - H_1x)^2 + \ldots + (z_m - H_mx)^2\right)\right) \end{aligned}$$

  where $\propto$ denotes proportionality.

- Differentiating the above expression with respect to x_j and setting to 0 gives:

  $$\begin{aligned} (z_1 - H_1\hat{x})h_{1j} + (z_2 - H_2\hat{x})h_{2j} + \ldots + (z_m - H_m\hat{x})h_{mj} = 0, &\quad j = 1, \ldots, n \\ [h_{1j}\ h_{2j}\ \ldots\ h_{mj}]\,(z - H\hat{x}) = 0, &\quad j = 1, \ldots, n \end{aligned}$$

  and, combining the equations for all j, we get

  $$H^T(z - H\hat{x}) = 0,\ \ H^TH\hat{x} = H^Tz, \quad \text{and, finally,} \quad \hat{x} = (H^TH)^{-1}H^Tz \quad (H^TH \text{ is invertible}).$$

The obtained solution is the *least squares* (LS) 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" LS interpretation, minimizing a quadratic error:

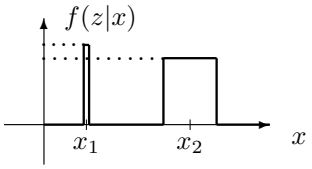
$$\epsilon(\hat{x}) = z - H\hat{x}, \qquad \hat{x}^{\mathrm{LS}} := \arg\min_{\hat{x}} \epsilon^T(\hat{x})\,\epsilon(\hat{x}) = (H^T H)^{-1} H^T z.$$

Variations:

- $w_i \sim \mathcal{N}(0, \sigma_i^2)$ results in *weighted* least squares. (**PSET 3: P3**)
- $w \sim \mathcal{N}(0, \Sigma)$, correlated noise. (**PSET 3: P5**)

Why is ML not always a good thing to do?

- The maximum may not be what you want. In the example below, the ML estimate is $\hat{x}^{\mathrm{ML}} = x_1$, but for robustness you may actually prefer x_2. (Small variations (shifts) in $f(z|x)$ cause x_1 to have likelihood zero, whereas the likelihood of x_2 remains the same.)



- You may have prior knowledge about x (its PDF $f(x)$). $\rightarrow$ MAP

5.2 Maximum a posteriori (MAP)

We can use the MAP estimate when x is a random variable with a known PDF. We already know what to do (Bayes' rule):

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

$$\hat{x}^{\mathrm{MAP}} := \arg\max_{x \in \mathcal{X}} f(z|x)\,f(x)$$

What choice of parameters are the most likely ones, given the observations and the prior belief about x?

Remarks:

- If $f(x)$ is constant, then $\hat{x}^{\mathrm{MAP}} = \hat{x}^{\mathrm{ML}}$; that is, if all values of x are a priori equally likely, then the estimates coincide.
- As for ML, we are maximizing a function over x, so the same criticism as mentioned above may apply.

Example

Consider the scalar observation

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

Then

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

By differentiating $f(x|z)$ with respect to x, setting to 0, and solving for $\hat{x}$, we get:

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

Notice the following special cases:

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

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

5.3 Minimum Mean Squared Error (MMSE)

The MMSE is the a posteriori estimate that minimizes the mean squared error:

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

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

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

The MMSE is simply the expected value conditioned on z.

Remarks:

- Compare this with the MAP estimate: the MAP is the maximum of the posterior PDF $f(x|z)$, while the MMSE is the mean of $f(x|z)$.
- Notice that the minimization above is defined without constraints on the admissible set for $\hat{x}$ (that is, $\hat{x} \in \mathbb{R}^n$). Therefore we can simply differentiate and set to zero to find the minimum, without having to worry about $\hat{x}$ being admissible.

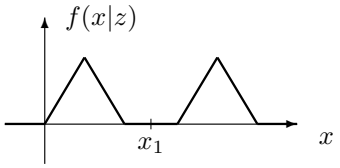
  For certain applications, one may want to restrict the admissible values for the estimate $\hat{x}$. For example, for a DRV with sample space $\mathcal{X}$, one may define

  $$\hat{x}^{\text{MMSE2}} := \arg\min_{\hat{x}\in\mathcal{X}} \mathrm{E}[(\hat{x}-x)^T(\hat{x}-x)\,|\,z]$$

  in order to ensure that $\hat{x}^{\text{MMSE2}} \in \mathcal{X}$, while $\hat{x}^{\text{MMSE}}$ as defined above is not necessarily in $\mathcal{X}$.

Why is MMSE not always a good thing to do?

Consider the bimodal distribution $f(x|z)$ below. The MMSE estimate is $\hat{x}^{\text{MMSE}} = x_1$. But the probability of x taking any value in a small neighborhood $[x_1-\Delta x, x_1+\Delta x]$ around x_1 is actually zero: $\Pr(x \in [x_1-\Delta x, x_1+\Delta x]\,|\,z) \approx 2\,f_{x|z}(x_1|z)\,\Delta x = 0$ (for small Δx). In the sense of "likelihood," this might be considered unsatisfactory. Nonetheless, x_1 is the estimate that minimizes the mean squared error.



Generally speaking, which type estimate (ML, MAP, MMSE, etc.) you should use, depends on the application (what is it that you want to minimize/maximize/achieve?).

5.4 Recursive Least Squares (RLS)

This section is a prelude to the standard way that the Kalman Filter is often derived. This approach bypasses PDFs and works directly with mean and variance.

Problem

Consider the observation model

$$z(k) = H(k)x + w(k) \qquad \text{with } z(k), w(k) \in \mathbb{R}^m,\ x \in \mathbb{R}^n.$$

- Prior knowledge: mean and variance of x, $\overline{x} := \mathrm{E}[x]$ and $P_x := \mathrm{E}[(x-\overline{x})(x-\overline{x})^T] = \mathrm{Var}[x]$, are given. Notice that we have no process model: x does not change (our knowledge of it changes, however).
- Measurement noise: zero-mean with known variance, $\mathrm{E}[w(k)] = 0$, $R(k) := \mathrm{Var}[w(k)]$.

- $\{x, w(1), w(2), \dots\}$ are independent.
- Typically, $n > m$; that is, fewer equations than unknowns at any particular time.

The objective is to compute an estimate $\hat{x}(k)$ of x from the observations $\{z(1), z(2), \dots, z(k)\}$ in the least squares sense (minimizing a quadratic error).

Standard LS

After collecting a whole bunch of data until time k, we can convert this to a standard, weighted least squares problem:

$$z = Hx + w \quad \text{with } z := \begin{bmatrix} z(1) \\ \vdots \\ z(k) \end{bmatrix}, \; H := \begin{bmatrix} H(1) \\ \vdots \\ H(k) \end{bmatrix}, \; w := \begin{bmatrix} w(1) \\ \vdots \\ w(k) \end{bmatrix}, \; R := \begin{bmatrix} R(1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & R(k) \end{bmatrix},$$

$$\hat{x}^{\text{LS}} = \arg\min_{\hat{x}} \epsilon^T(\hat{x})\,\epsilon(\hat{x}) = \arg\min_{\hat{x}} (z - H\hat{x})^T (z - H\hat{x}) = (H^T R^{-1} H)^{-1} H^T R^{-1} z.$$

The implementation of standard LS would mean that we have to solve a full LS problem (with growing dimensions) at every time step k.

Recursive LS

Can we build an estimate in real-time; that is, update the current estimate of x recursively when a new measurement $z(k)$ comes in? We *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 given all measurements up to time k, $\{z(1), z(2), \dots, z(k)\}$, and

$K(k)$ is the gain matrix at time k, the only design variable.

- The estimator structure is intuitive: if $z(k) = H(k)\hat{x}(k\text{–}1)$ (measurement is consistent with the estimate), then $\hat{x}(k) = \hat{x}(k\text{–}1)$ (estimate remains unchanged).
- We analyze the estimation error $e(k) := x - \hat{x}(k)$. We have

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

  and, for the mean,

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

  If we initialize the estimator with $\hat{x}(0) = \overline{x}$, then $\mathrm{E}[e(0)] = 0$, and from the above recursion, we get $\mathrm{E}[e(k)] = 0$ for all k. That is, the estimator is unbiased (independent of the choice of $K(k)$).

- We now choose $K(k)$ to minimize the mean squared error

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

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

  – One can show that **(PSET 3: P9)**

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

– We want to minimize $J(k) = \text{trace}(P(k))$, 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: (proof for 2×2 case in **PSET 3: P7+8**)

$$\frac{\partial \,\text{trace}(ABA^T)}{\partial A} = 2AB \;\; (\text{for } B = B^T) \quad \text{and} \quad \frac{\partial \,\text{trace}(AB)}{\partial A} = B^T,$$

where, for a scalar function $g(A)$ and a matrix $A \in \mathbb{R}^{m \times n}$, $\frac{\partial g(A)}{\partial A}$ denotes the partial derivative of $g(A)$ with respect to A defined as follows:

$$\frac{\partial g(A)}{\partial A} = \begin{bmatrix} \frac{\partial g(A)}{\partial a_{11}} & \cdots & \frac{\partial g(A)}{\partial a_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial g(A)}{\partial a_{m1}} & \cdots & \frac{\partial g(A)}{\partial a_{mn}} \end{bmatrix}.$$

Furthermore, since $\text{trace}(C) = \text{trace}(C^T)$, we have

$$\frac{\partial \,\text{trace}(B^T A^T)}{\partial A} = \frac{\partial \,\text{trace}(AB)}{\partial A} = B^T.$$

– Using these results, the necessary condition $\frac{\partial J(k)}{\partial K(k)} = 0$ results in (**PSET 3: P10**)

$$K(k)\big(H(k)P(k\text{–}1)H^T(k) + R(k)\big) = P(k\text{–}1)H^T(k)$$
$$K(k) = P(k\text{–}1)H^T(k)\big(H(k)P(k\text{–}1)H^T(k) + R(k)\big)^{-1} \qquad (\text{inverse exists if } R(k) > 0 \;\forall\; k).$$

Summary (RLS algorithm)

Initialization: $\hat{x}(0) = \overline{x}$, $P(0) = P_x = \text{Var}[x]$

Recursion:

Observe: $z(k)$

Update:
$$K(k) = P(k\text{–}1)H^T(k)\big(H(k)P(k\text{–}1)H^T(k) + R(k)\big)^{-1}$$
$$\hat{x}(k) = \hat{x}(k\text{–}1) + K(k)\big(z(k) - H(k)\hat{x}(k\text{–}1)\big)$$
$$P(k) = \big(I - K(k)H(k)\big)P(k\text{–}1)\big(I - K(k)H(k)\big)^T + K(k)R(k)K^T(k)$$

The matrices $K(k)$ and $P(k)$ can be pre-computed from the problem data P_x, $\{H(\cdot)\}$, and $\{R(\cdot)\}$.