Chapter 6

Linear regression and least squares estimation

Up until now, we have dealt with deterministic systems. In this part we consider estimation of variables if the measurements are affected by noise/uncertainties, etc. First, we consider a static problem and then the dynamic one.

6.1 Linear regression

The regression problem can be formulated as *identify the unknown parameters θ, given a collection of known samples and a linear model that relates the samples to the parameters*. The known samples, generally denoted by $y(k)$, $k = 1, 2, \ldots, N$ are the *regressed variables*, a known vector $\varphi(k) = \begin{pmatrix}\varphi_1(k) & \varphi_2(k) & \ldots & \varphi_n(k)\end{pmatrix}$ for $k = 1, 2, \ldots, N$ contains the *regressors*, and the model relating them to the *unknown parameter vector θ* is given by

$$y(k) = \varphi^T(k)\theta$$

Such model are often used in function fitting, function approximation, time-series prediction, supervised learning, system identification, etc.

Writing the model for each data point, the system of equations becomes

$$\begin{aligned}
y(1) &= \varphi_1(1)\theta_1 + \varphi_2(1)\theta_2 + \ldots \varphi_n(1)\theta_n \\
y(2) &= \varphi_1(2)\theta_1 + \varphi_2(2)\theta_2 + \ldots \varphi_n(2)\theta_n \\
&\vdots \\
y(N) &= \varphi_1(N)\theta_1 + \varphi_2(N)\theta_2 + \ldots \varphi_n(N)\theta_n
\end{aligned}$$

or, in a vector notation,

$$\boldsymbol{y} = \Phi\theta \tag{6.1}$$

If the number of data point N is equal to the number of unknown parameters n and the equations are linearly independent, then the system (6.1) has a unique solution. However, this is rarely the case, as 1) usually there are more measurements than parameters; 2) the model is un general not exact, but only an approximation; and 3) the samples (measurements) are affected by noise/uncertainties, etc. Therefore, the problem of determining the unknown parameters θ becomes an *optimization problem*: *find θ that minimizes* $\sum_{k=1}^{N} \epsilon^2(k)$, where $\epsilon(k) = y(k) - \varphi^T(k)\theta$.

The solution of this problem can be computed as follows. Note that $F(N) = \sum_{k=1}^{N} \epsilon^2(k)$ can be written as $F(N) = (\boldsymbol{y} - \Phi\theta)^T(\boldsymbol{y} - \Phi\theta)$. Computing the gradient with respect to θ we obtain

$$\frac{\partial F(N)}{\partial \theta} = -2\Phi^T(\boldsymbol{y} - \Phi\theta)$$

i.e., the minimum of the function $F(N)$ is given by θ for which $\Phi^T\boldsymbol{y} = \Phi^T\Phi\theta$. If $\Phi^T\Phi$ is invertible, then θ is given by $\theta = (\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{y}$.

Example 6.1 *Consider the case of obtaining several (noisy) measurements of e.g., the distance from a given object,* $\boldsymbol{y} = \begin{pmatrix} 1 & 1.3 & 0.8 & 1.1 & 0.9 \end{pmatrix}$. *Our objective is to estimate the distance based on these noisy measurements, i.e., find θ, that minimizes* $\sum_{k=1}^{5}(y(k) - theta)^2$. *Note that for this particular case, the regressor* $\varphi = 1$. *Thus, we have*

$$\begin{aligned} y(1) &\approx \varphi(1)\theta = \theta = 1 \\ y(2) &\approx \varphi(1)\theta = \theta = 1.3 \\ y(3) &\approx \varphi(1)\theta = \theta = 0.8 \\ y(4) &\approx \varphi(1)\theta = \theta = 1.1 \\ y(5) &\approx \varphi(1)\theta = \theta = 0.9 \end{aligned}$$

We have $\Phi = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \end{pmatrix}^T$, $\Phi^T\Phi = 5$, *and θ estimated as*

$$\theta = \frac{1}{5}\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1.3 \\ 0.8 \\ 1.1 \\ 0.9 \end{pmatrix} = 1.02$$

In general, when a constant has to be estimated from direct measurements, the problem is reduced to computing the mean of the measurements.

6.2 Mathematical intermezzo: notions of statistics

A *random variable* is a mapping from a set of experimental outcomes $\mathcal{X}$to a set of real numbers. Note that, a random variable X is a function defined on a sample space. A specific value x is called a *realization* of the random variable. The fundamental property of a random variable X is its *probability distribution function* (PDF), defined as

Consider a set $\mathcal{X}$ containing possible outcomes x for an event. A corresponding discrete random variable X is described by the *probability mass function* that lists the probabilities of all individual values $p(x_i)$, $x_i \in \mathcal{X}$. The probabilities sum up to 1, i.e., $P(\mathcal{X}) = \sum_{x \in \mathcal{X}} p(x) = 1$.

Consider now an interval $\mathcal{X}$. The corresponding continuous-valued random variable X is described the by *probability density functions* that defines the probability of obtaining a value in some sub-interval $[a,\, b] \subset \mathcal{X}$, $P(X \in [a,\, b]) = \int_a^b f(x)dx$. The probabilities of the random variable having a value in the whole interval $\mathcal{X}$ is 1, i.e., $P(\mathcal{X}) = \int_{x \in \mathcal{X}} f(x)dx = 1$.

Example 6.2 ***todo: uniform, Gauss***

The *mean* or *expected value* of a random variable is defined as

$$E(X) = \begin{cases} \sum_{x \in \mathcal{X}} p(x)x & \text{discrete} \\ \int_{x \in \mathcal{X}} f(x)dx & \text{continuous} \end{cases}$$

A function $g : \mathcal{X} \to \mathbb{R}$ that depends on a random variable X is itself a random variable, and its expected value is

$$E(g(X)) = \begin{cases} \sum_{x \in \mathcal{X}} p(x)g(x) & \text{discrete} \\ \int_{x \in \mathcal{X}} f(x)g(x)dx & \text{continuous} \end{cases}$$

The *variance* can be considered as the “spread” around the expected value and is given by

$$Var(X) = E\left((X - E(X))^2\right) = E(X^2) - (E(X))^2$$

The *standard deviation* σ is the square root of the variance, i.e., $\sigma = \sqrt{Var(X)}$.

Two events are independent if the occupance of one event has no effect on the probability of the occurrence of the other event. The *joint probability* density or mass function, as it may be the case, if $f_{XY}(x,y)$. Two random variables X and Y are *independent* if $P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$. This implies that $f_{XY}(x,y) = f_X(x)f_Y(y)$.

The *covariance* of two random variables X and Y is

$$Cov(X,Y) = E\left((X - E(X))(Y - E(Y))^T\right) = E(XY^T) - E(X)E(Y)^T$$

A covariance matrix is positive semidefinite.

The *correlation* of X and Y is $E(XY^T)$ and is defined as (assuming both X and Y column vectors)

$$E(XY^T) = \begin{pmatrix} E(X_1Y_1) & \dots & E(X_1Y_m) \\ \vdots & \vdots & \vdots \\ E(X_nY_1) & \dots & E(X_nY_m) \end{pmatrix}$$

Two random variables are *uncorrelated* if $E(XY^T) = E(X)E(Y)^T$.

Let us consider now two *stochastic processes* $X(t)$ and $Y(t)$. The *cross-correlation* of $X(t)$ and $Y(t)$ is defined as $E(X(t_1)Y(t_2)^T)$. The random processes $X(t)$ and $Y(t)$ are *uncorrelated* if $E(X(t_1)Y(t_2)^T) = E(X(t_1))E(Y(t_2))$, $\forall t_1, t_2$. The *cross-covariance* of $X(t)$ and $Y(t)$ is

$$Cov_{XY}(t1, t2) = E\left((X(t_1) - E(X(t_1)))(Y(t_2) - E(Y(t_2)))^T\right)$$

$X(t)$ is called *white noise* if the random variable $X(t_1)$ is independent from the random variable $X(t_2)$, $\forall t_1 \neq t_2$.

6.3 Least-squares estimation

6.3.1 Batch least-squares estimation

Let us now return to the problem of estimating a constant $\boldsymbol{x}$, under the assumption that each element of the measurement vector $\boldsymbol{y}$ is a linear combination of the elements of $\boldsymbol{x}$, with the addition of some measurement noise v, i.e.,

$$\begin{aligned} y(1) &= C_{11}x_1 + C_{12}x_2 + \cdots + C_{1n}x_n + v_1 \\ y(2) &= C_{21}x_1 + C_{22}x_2 + \cdots + C_{2n}x_n + v_2 \\ &\vdots \\ y(k) &= C_{k1}x_1 + C_{k2}x_2 + \cdots + C_{kn}x_n + v_k \end{aligned}$$

or, in matrix form

$$\boldsymbol{y} = C\boldsymbol{x} + \mathbf{v}$$

The measurement residual (measurement *error* in previous chapters) is defined as $\boldsymbol{e}_y = \boldsymbol{y} - C\widehat{\boldsymbol{x}}$. The most probable value of $\boldsymbol{x}$ is the vector $\widehat{\boldsymbol{x}}$ that minimizes $\boldsymbol{e}_y^T\boldsymbol{e}_y$, leading to the solution in Section 6.1.

Consider now the case when the variance of the measurement noise may be different for each measurement, i.e., $E(v_i^2) = \sigma_i^2$, $i = 1, 2, \dots, k$. The measurement covariance matrix is

$$R = E(\mathbf{v}\mathbf{v}^T) = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \dots & \sigma_k^2 \end{pmatrix}$$

In such a case, one may exploit this knowledge and minimize the *weighted* sum of squares,

$$\begin{aligned}F &= \frac{e_{y1}^2}{\sigma_1^2} + \cdots + \frac{e_{yk}^2}{\sigma_k^2}\\ &= \boldsymbol{e}_y^T R^{-1} \boldsymbol{e}_y\\ &= (\boldsymbol{y} - C\widehat{\boldsymbol{x}})^T R^{-1} (\boldsymbol{y} - C\widehat{\boldsymbol{x}})\\ &= \boldsymbol{y}^T R^{-1} \boldsymbol{y} - 2\widehat{\boldsymbol{x}}^T C^T R^{-1} \boldsymbol{y} + \widehat{\boldsymbol{x}}^T C^T R^{-1} X \widehat{\boldsymbol{x}}\end{aligned}$$

The partial derivative of F wrt. $\widehat{\boldsymbol{x}}$ is

$$\frac{\partial F}{\partial \widehat{\boldsymbol{x}}} = 2\widehat{\boldsymbol{x}}^T C^T R^{-1} C - 2\boldsymbol{y}^T R^{-1} C$$

resulting in

$$\begin{aligned}C^T R^{-1} \boldsymbol{y} &= C^T R^{-1} C \widehat{\boldsymbol{x}}\\ \widehat{\boldsymbol{x}} &= (C^T R^{-1} C) C^T R^{-1} \boldsymbol{y}\end{aligned}$$

Note that the above computations require that 1) there are *enough* measurements – at least as many as unknowns to be estimated and 2) the measurement noise matrix R is nonsingular, i.e., each measurement is corrupted by some noise.

6.3.2 Recursive least-squares estimation

A problem presented by the previous approach is that if the measurements are obtained sequentially, the matrix C has to be augmented for each measurement and the estimate needs to be recomputed. This may even be computationally unfeasible for a very large number of measurements.

In what follows, we present a *recursive* estimator that updates the estimate after obtaining a new measurement $\boldsymbol{y}$.

A *linear recursive estimator* has the form

$$\begin{aligned}\boldsymbol{y}(k) &= C_k \boldsymbol{x} + \mathbf{v}(k)\\ \widehat{\boldsymbol{x}}(k) &= \widehat{\boldsymbol{x}}(k-1) + K_k(\boldsymbol{y}(k) - C_k \widehat{\boldsymbol{x}}(k-1))\end{aligned}$$

i.e., the estimate $\widehat{\boldsymbol{x}}(k)$ is computed based on the previous estimate and the new measurement. K_k is the estimator gain matrix and $\boldsymbol{y}(k) - C\widehat{\boldsymbol{x}}(k-1)$ is the correction term.

The estimation error mean can be computed as – note that we are estimating a *constant* $\boldsymbol{x}$ –

$$\begin{aligned}E(\boldsymbol{e}(k)) &= E(\boldsymbol{x} - \widehat{\boldsymbol{x}}(k))\\ &= E(\boldsymbol{x} - \widehat{\boldsymbol{x}}(k-1) - K_k(C_k \boldsymbol{x} + \mathbf{v}(k) - C_k \widehat{\boldsymbol{x}}(k-1)))\\ &= E(\boldsymbol{e}(k-1) - K_k C_k (\boldsymbol{x} - \widehat{\boldsymbol{x}}(k-1)) - K_k \mathbf{v}(k))\\ &= (I - K_k C_k) E(\boldsymbol{e}(k-1)) - K_k E(\mathbf{v}(k))\end{aligned}$$

If $E(\boldsymbol{e}(k-1)) = 0$ and $E(\mathbf{v}(k)) = 0$, then $E(\boldsymbol{e}(k)) = 0$. This means that if the noise is zero-mean and the initial estimate $\widehat{\boldsymbol{x}}(0)$ is equal to the expected value of $\boldsymbol{x}$, then the expected value of $\widehat{\boldsymbol{x}}(k)$ will be equal to $\boldsymbol{x}$, i.e., *on average* the estimate $\widehat{\boldsymbol{x}}$ will be equal to $\boldsymbol{x}$. Such an estimator is called *unbiased.* Note that this property holds independently of the value of K_k.

Next, we determine the optimal K_k that minimizes the sum of the variances of the estimation errors at step k, $E(\boldsymbol{e}(k)^T\boldsymbol{e}(k))$. The *estimation error covariance matrix* is defined as

$$P_k = E(\boldsymbol{e}(k)\boldsymbol{e}(k)^T)$$

Note that trace$(P_k) = E(\boldsymbol{e}(k)^T\boldsymbol{e}(k))$.

A recursive expression of P_k can be obtained as

$$\begin{aligned}
P_k &= E(\boldsymbol{e}(k)\boldsymbol{e}(k)^T)\\
&= E\left(((I - K_kC_k)\boldsymbol{e}(k-1) - K_k\mathbf{v}(k))\,((I - K_kC_k)\boldsymbol{e}(k-1) - K_k\mathbf{v}(k))^T\right)\\
&= (I - K_kC_k)E(\boldsymbol{e}(k-1)\boldsymbol{e}(k-1)^T)(I - K_kC_k)\\
&- K_kE(\mathbf{v}(k)\boldsymbol{e}(k-1)^T)(I - K_kC_k) - (I - K_kC_k)E(\boldsymbol{e}(k-1)\mathbf{v}(k)^T)K_k^T\\
&+ K_kE(\mathbf{v}(k)\mathbf{v}(k)^T)K_k^T
\end{aligned}$$

Since the estimation error at time $k-1$, $\boldsymbol{e}(k-1)$ is independent of the noise at time k, $\mathbf{v}(k)$, $E(\boldsymbol{e}(k-1)\mathbf{v}(k)^T) = E(\boldsymbol{e}(k-1))E(\mathbf{v}(k)) = 0$, and $E(\mathbf{v}(k)\boldsymbol{e}(k-1)^T) = 0$. Furthermore, $E(\boldsymbol{e}(k-1)\boldsymbol{e}(k-1)^T) = P_{k-1}$ and $E(\mathbf{v}(k)\mathbf{v}(k)^T) = R_k$. Thus, we obtain

$$P_k = (I - K_kC_k)P_{k-1}(I - K_kC_k) + K_kR_kK_k^T$$

Note that this form of P_k guarantees that – assuming it has been initialized at a positive definite matrix and R_k is positive – it will be positive definite. It is also consistent with the intuition that if the uncertainty in the measurement increases, i.e., R_k increases, then the uncertainty of the estimate also increases.

Let us now compute K_k that minimizes the sum of the variances of the estimation errors at step k, i.e., the *trace* of P_k. Note that

$$\frac{\partial \text{trace}(ABA^T)}{\partial A} = 2AB$$

where A and $B = B^T$ are matrices of appropriate dimensions. Thus,

$$\frac{\partial P_k}{\partial K_k} = 2(I - K_kC_k)P_{k-1}(-C_k^T) + 2K_kR_k$$

Setting the above equal to zero results in

$$\begin{aligned}
K_k(R_k + C_kP_{k-1}C_k) &= P_{k-1}C_k^T\\
K_k &= P_{k-1}C_k^T(R_k + C_kP_{k-1}C_k)^{-1}
\end{aligned}$$

Thus, the recursive least square estimator can be summarized as:

1. Initialize the estimate $\widehat{\boldsymbol{x}}(0)$ and its covariance matrix P_0. If $\boldsymbol{x}$ is perfectly known, then $P_0 = 0$; if no knowledge is available, then $P_0 = \infty I$.

2. At each step

   - Obtain the measurement, assuming that it is given by $\boldsymbol{y}(k) = C\boldsymbol{x} + \mathbf{v}(k)$, where $\mathbf{v}(k)$ is a white noise, with covariance R_k.
   - Update the estimate and its covariance as:

$$\begin{aligned} K_k &= P_{k-1}C_k^T(R_k + C_k P_{k-1} C_k)^{-1} \\ \widehat{\boldsymbol{x}}(k) &= \widehat{\boldsymbol{x}}(k-1) + K_k(\boldsymbol{y}(k) - C_k\widehat{\boldsymbol{x}}(k-1)) \\ P_k &= (I - K_k C_k)P_{k-1}(I - K_k C_k) + K_k R_k K_k^T \end{aligned}$$

Note that there are alternative forms, see e.g., the developments in **?**, for the covariance and gain matrices. For instance, often used forms are

$$\begin{aligned} P_k &= (I - K_k C_k)P_{k-1} \\ P_k &= (P_{k-1}^{-1} + C_k^T R_k^{-1} C_k)^{-1} \\ K_k &= P_k C_k^T R_k^{-1} \end{aligned}$$

6.4 Propagation of states and covariances

Consider now the discrete-time system

$$\boldsymbol{x}(k) = A_{k-1}\boldsymbol{x}(k-1) + B_{k-1}\boldsymbol{u}(k-1) + \mathbf{w}(k-1)$$

where $\boldsymbol{u}$ is a known input and $\mathbf{w}(k-1)$ is a white noise with covariance Q_{k-1}.

It can easily be seen that the expected value of the state at time k is

$$\begin{aligned} E(\boldsymbol{x}(k)) &= E(A\boldsymbol{x}(k-1) + B\boldsymbol{u}(k-1) + \mathbf{w}(k-1)) \\ &= AE(\boldsymbol{x}(k-1)) + B\boldsymbol{u}(k-1) \end{aligned}$$

Regarding the covariance, **todo:write derivation**, we obtain

$$\begin{aligned} P_k &= E((\boldsymbol{x}(k) - E(\boldsymbol{x}(k)))(\boldsymbol{x}(k) - E(\boldsymbol{x}(k)))^T) \\ &= A_{k-1}P_{k-1}A_{k-1}^T + Q_{k-1} \end{aligned}$$

In many cases the noise does not enter directly the state (or measurement, as it may be the case) equation, but is multiplied with some matrix, i.e., we have

$$\boldsymbol{x}(k) = A_{k-1}\boldsymbol{x}(k-1) + B_{k-1}\boldsymbol{u}(k-1) + M_{k-1}\tilde{\mathbf{w}}(k-1)$$

Note that if $\tilde{\mathbf{w}}(k-1)$ is a white noise with covariance Q_{k-1} then the covariance of $M_{k-1}\tilde{\mathbf{w}}(k-1)$ is $M_{k-1}Q_{k-1}M_{k-1}^T$. Thus, the covariance matrix of $\boldsymbol{x}(k)$ will be $A_{k-1}P_{k-1}A_{k-1}^T + M_{k-1}Q_{k-1}M_{k-1}^T$.

Similarly, given the measurement equation

$$\boldsymbol{y}(k) = C_k\boldsymbol{x}(k) + N_k\tilde{\mathbf{v}}(k)$$

where $\mathbf{v}(k)$ is a zero mean white noise with covariance R_k, then the covariance matrix of the measurement is $C_kP_kC_k^T + N_kR_kN_k^T$.

Consider now the continuous-time system

$$\dot{\boldsymbol{x}} = A\boldsymbol{x} + B\boldsymbol{u}$$

It is quite straightforward that, by taking the expectation of both sides, we have

$$E(\dot{\boldsymbol{x}}) = AE(\boldsymbol{x}) + B\boldsymbol{u}$$

For the covariance matrices, using a limiting argument based on the discrete-time case, see **?**, one can get

$$\dot{P} = AP + PA^T + Q$$

6.5 The discrete-time Kalman filter

With all the notions introduced until nw, we are ready to formulate the discrete-time Kalman filter. Consider the discrete-time system

$$\begin{aligned}\boldsymbol{x}(k) &= A_k\boldsymbol{x}(k-1) + B_k\boldsymbol{u}(k-1) + \mathbf{w}(k-1)\\ \boldsymbol{y}(k) &= C_k\boldsymbol{x}(k) + \mathbf{v}(k)\end{aligned}$$

where $\mathbf{w}$ and $\mathbf{v}$ are zero-mean white noises with covariance matrices Q_k and R_k, respectively.

After the estimate of the state $\widehat{\boldsymbol{x}}(0)$ and the covariance matrix P_0 are initialized, based on the state equation in (6.5) a *prediction* of the next state and the corresponding covariance can be made, thus:

$$\begin{aligned}\boldsymbol{x}_{pred} &= A_k\boldsymbol{x}(k-1) + B_k\boldsymbol{u}(k-1)\\ P_{pred} &= A_kP_{k-1}A_k^T + Q_{k-1}\end{aligned}$$

The predicted state $\boldsymbol{x}_{pred}$ is usually called the *a priori* (before measurement) estimate and is sometimes denoted by $\boldsymbol{x}^-(k)$, where the $^-$ denotes that a measurement is not yet available. Similarly, P_{pred} can be denoted by P_k^-.

Once a measurement becomes available, the prior estimate can be *updated* according to the derivations from Section 6.3.2, i.e.,

$$K_k = P_{pred} C_k^T (C_k P_{pred} C_k^T + R_k)^{-1}$$
$$\widehat{\boldsymbol{x}}(k) = \boldsymbol{x}_{pred} + K_k(\boldsymbol{y} - C_k \boldsymbol{x}_{pred})$$
$$P_k = (I - K_k C_k) P_{pred} (I - K_k C_k)^T + K_k R_k K_k^T$$

The updated estimate $\widehat{\boldsymbol{x}}(k)$ is the *a posteriori* estimate.

Note that since the system dynamics is determined by the stochastic processes **w** and **v**, both the state $\boldsymbol{x}$ and its estimate $\widehat{\boldsymbol{x}}$ will be random variables. If vw and **w** are zero-mean, uncorrelated and white, the Kalman filter is the *optimal linear* solution to the problem of minimizing the weighted norm of the estimation error $\boldsymbol{e} = \boldsymbol{x} - \widehat{\boldsymbol{x}}$. In such a case, the innovation $\boldsymbol{y} - C_k \boldsymbol{x}_{pred}$ is also zero-mean and white with covariance $C_k P_{pred} C_k^T + R_k$. Since for a given application the innovation can be measured and the mean and covariance approximated, this property can be used to verify the model and the noise statistics. Modelling errors and numerical errors are two of the primary causes for which the filter may diverge on a real system, even though the theory is correct. Modelling errors may sometimes be compensated by adding fictitious process noise.