CA684	Jia Yuan Yu
5: Regression, state estimation	
DCU	March 14, 2014

These notes are based on the notes "Statistique appliquée" by Biau and Tsybakov and the notes "Linear dynamical model, Kalman filtering and statistic" by Bolviken, Christophersen, and Storvik.

Supervised learning is not all about classification: another (harder) problem is regression. In classification, we have $X_j \in \mathbb{R}^d$ and $Y_j \in \{0,1\}$, and the variable Y_j is a label. In regression, we have $Y_j \in \mathbb{R}$: the variable Y_j is a quantity. We first present the regression problem and then a generalization of it: the estimation problem in state-space representation, which is the bread and butter of control engineering.

1 Regression function

Let $(X,Y) \in \mathbb{R}^d \times \mathbb{R}$ be a pair of random variables such that $\mathbb{E}\|X\|_2 < \infty$ and $\mathbb{E}Y^2 < \infty$. The function $g : \mathbb{R}^d \to \mathbb{R}$ such that

$$g^*(z) = \mathbb{E}(Y \mid X = z)$$

is called the regression function of Y. This function has the property of minimizing the mean-squared error, i.e.,

$$\mathbb{E}(Y - g^*(X))^2 = \min_h \mathbb{E}(Y - h(X))^2,$$

for all distributions.

2 Linear regression

Suppose that for all $j = 1, \ldots, n$:

$$Y_j = g(X_j) + \nu_j,$$

where $\{\nu_i\}$ are noise random variables that are independent and have zero mean. The function g is unknown. The regression problem is to find a function g_n based on $X_1, Y_1, \ldots, X_n, Y_n$ that estimates g.

The linear regression problem is a special case, where we assume additionally that there exists an (unknown) $\theta \in \mathbb{R}^d$ such that $g(z) = \theta^T z$[1]. In this case, the assumption on Y_j becomes:

$$Y_j = \theta^T X_j + \nu_j, \quad j = 1, \ldots, n, \tag{1}$$

so that the problem becomes estimating θ by $\hat{\theta}_n$.

[1] This denotes an inner product.

Example 2.1 (Polynomial regression)**.** Let $Z \in \mathbb{R}$, and define $X = (1, Z, \ldots, Z^{d-1}) \in \mathbb{R}^d$. Then, we have

$$\theta^T X = \theta^1 + \theta^2 Z + \ldots + \theta^d Z^{d-1},$$

so that polynomial regression on $\mathbb{R}^1$ can be formulated as multivariate linear regression.

3 Least squares

The least squares estimate of the parameter θ is

$$\hat{\theta}_n = \arg\min_{\omega \in \mathbb{R}^d} \sum_{j=1}^{n} (Y_j - X_j^T \omega)^2,$$

where

$$h(\omega) = \sum_{j=1}^{n} (Y_j - X_j^T \omega)^2$$

is convex and non-negative.

For $\hat{\theta}_n$ to be a minimum, we must have

$$\frac{\mathrm{d}h}{\mathrm{d}\theta^i}(\hat{\theta}_n) = 0, \quad \text{for all } i = 1, \ldots, d,$$

or, by taking the derivative:

$$2 \sum_{j=1}^{n} X_j (Y_j - X_j^T \omega) = 0. \tag{2}$$

Letting

$$B \triangleq \sum_{j=1}^{n} X_j X_j^T,$$

the above (2) becomes

$$\sum_{j=1}^{n} X_j Y_j = B \hat{\theta}_n.$$

Hence, if B is an invertible matrix, then we have

$$\hat{\theta}_n = B^{-1} \sum_{j=1}^{n} X_j Y_j. \tag{3}$$

3.1 Matrix notation

You will often see linear regression equations of (1) presented by a single matrix equation:

$$\vec{Y} = \mathbf{X}\theta + \vec{\nu},$$

where

$$\begin{aligned}\vec{Y} &= (Y_1, \ldots, Y_n)^T, \\ \mathbf{X} &= (X_1, \ldots, X_n)^T,\end{aligned}$$

are the data, and $\vec{\nu} = (\nu_1, \ldots, \nu_d)$ is the vector of noise terms. In this case, we can write (3) as

$$\hat{\theta}_n = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{Y}.$$

3.2 Properties of least squares estimate $\hat{\theta}_n$

Theorem 3.1. *Suppose that $X_1, \ldots, X_n$ are deterministic, and that the matrix B is invertible. Furthermore, $\vec{\nu}$ is a random vector such that $\mathbb{E}\vec{\nu} = 0$ and there exists an unknown $\sigma > 0$ such that $V(\vec{\nu}) = \mathbb{E}\vec{\nu}\vec{\nu}^T = \sigma^2 I$. Then, we have*

$$\begin{aligned}\mathbb{E}\hat{\theta}_n &= \theta, \\ V(\hat{\theta}_n) &= \mathbb{E}[(\hat{\theta}_n - \mathbb{E}\hat{\theta}_n)(\hat{\theta}_n - \mathbb{E}\hat{\theta}_n)^T] = \sigma^2 B^{-1}.\end{aligned}$$

The first equation says that the estimator $\hat{\theta}_n$ is unbiased.

Proof. Observe that

$$\begin{aligned}\hat{\theta}_n &= B^{-1}\mathbf{X}^T\vec{Y} \\ \text{(assumption)} \quad &= B^{-1}\mathbf{X}^T(\mathbf{X}\theta + \vec{\nu}) \\ &= \theta + B^{-1}\mathbf{X}^T\vec{\nu}.\end{aligned}$$

The first claim follows by taking the expectation. Next, observe that

$$\begin{aligned}V(\hat{\theta}_n) &= \mathbb{E}[(\hat{\theta}_n - \theta)(\hat{\theta}_n - \theta)^T] \\ &= \mathbb{E}(B^{-1}\mathbf{X}^T\vec{\nu})(\vec{\nu}^T\mathbf{X}B^{-1}) \\ &= B^{-1}\mathbf{X}^T[\mathbb{E}\vec{\nu}\vec{\nu}^T]\mathbf{X}B^{-1} \\ \text{(assumption)} \quad &= \sigma^2 B^{-1}\mathbf{X}^T\mathbf{X}B^{-1} = \sigma^2 B^{-1}.\end{aligned}$$

□

If $\vec{\nu}$ is normally distributed, then we also have that $\hat{\theta}_n$ is distributed as $\mathcal{N}(\theta, \sigma^2 B^{-1})$.

4 State-space estimation problems

In this section, we generalize what was done in the previous section. The state-space estimation problem in control theory is similar to the regression problem. The (time-invariant) state-space model is

$$X_k = \Phi_{k-1} X_{k-1} + w_{k-1}, \quad k = 1, 2, \ldots,$$
$$Y_k = H_k X_k + v_k, \quad k = 1, 2, \ldots,$$

where the vector $X_k \in \mathbb{R}^n$ is the state vector, and $Y_k \in \mathbb{R}^m$ is the observation vector, and the initial state X_0 is deterministic and known. The matrices $\{\Phi_k\}$ and $\{H_k\}$ are deterministic matrices that are known. The sequences $\{w_k\}$ and $\{v_k\}$ are noise processes that are independent and have zero mean and covariance matrices $\mathbb{E} w_k w_k^T = Q_k$ and $\mathbb{E} v_k v_k^T = R_k$.

Example 4.1 (Linear regression in state-space form)**.** The state-space representation of the linear regression assumption (1) is

$$\theta_k = I \theta_{k-1} + \vec{0},$$
$$Y_k = X_k^T \theta_k + \vec{\nu}_k,$$

where θ_k is the state process and X_k is known.

4.1 State estimation problem

Assuming that the initial state $X(0)$ and the observations $Y_1, \ldots, Y_k$ are known at time k, the state-estimation problem at time k is to construct an estimator $\hat{X}_k$ for X_k. Of particular interest are linear estimator of the form

$$\hat{X}_k = a_k^k Y_k + a_k^{k-1} Y_{k-1} + \ldots + a_k^1 Y_1,$$

where the coefficients $\{a_k = (a_k^1, \ldots, a_k^k)\}$ are learned from the data $X_0, Y_1, \ldots, Y_k$.

4.2 Kalman filter

For integers k and ℓ, let $\vec{Y}_\ell = (Y_1, \ldots, Y_\ell)^T$, and we define

$$\hat{X}(k \mid \ell) = \mathbb{E}(X_k \mid \vec{Y}_\ell).$$

It is well-known (Gauss-Markov Theorem) that this $\hat{X}(k \mid \ell)$ is optimal[2] in the mean square error sense:

$$\mathbb{E}\Big[(\hat{X}(k \mid \ell) - X_k)^T (\hat{X}(k \mid \ell) - X_k)\Big] \le \inf_{h: \mathbb{R}^{\ell \times n} \to \mathbb{R}^n} \mathbb{E}\Big[(h(\vec{Y}_\ell) - X_k)^T (h(\vec{Y}_\ell) - X_k)\Big].$$

[2] Think of $\hat{X}(k \mid \ell)$ is the analogue of the regression function g^* and the Bayes classifier g^*: it is our baseline, but may be hard to compute.

Of particular interest is $\hat{X}(k \mid k) = \mathbb{E}(X_k \mid \vec{Y}_k)$, which is the estimator constructed with all the data seen at time k[3]. The question is: Can we compute $\hat{X}(k \mid k)$ efficiently? It turns out that yes, provided that the noise processes w_k and v_k are Gaussian, $\hat{X}(k \mid k)$ can be computed in a recursive and linear fashion. This is the so-called Kalman filter.

We need two more definitions first: the estimation error and error covariance matrix:

$$
\begin{aligned}
e(k \mid \ell) &= \hat{X}(k \mid \ell) - X_k, \\
P(k \mid \ell) &= \mathbb{E}e(k \mid \ell)e(k \mid \ell)^T.
\end{aligned}
$$

We are now ready to present the Kalman filter. The Kalman filter is the recursive sequence of linear estimators: for $k = 1, 2, \ldots,$

$$
\begin{aligned}
\hat{X}(k \mid k-1) &= \Phi_{k-1}\hat{X}(k-1 \mid k-1), \\
\hat{X}(k \mid k) &= \hat{X}(k \mid k-1) + M_k[Y_k - H_k\hat{X}(k \mid k-1)],
\end{aligned}
$$

where M_k is computed:

$$
\begin{aligned}
P(k \mid k-1) &= \Phi_{k-1}P(k-1 \mid k-1)\Phi_{k-1}^T + Q_{k-1}, \\
S_k &= H_kP(k \mid k-1)H_k^T + R_k, \\
M_k &= P(k \mid k-1)H_k^TS_k^{-1}, \\
P(k \mid k) &= (I - M_kH_k)P(k \mid k-1).
\end{aligned}
$$

The proof can be found in standard textbooks. The fact that the Kalman filter is linear and recursive makes it very efficient.

4.3 Kalman filter for linear regression

Recall the state-space representation of linear regression

$$
\begin{aligned}
\theta_k &= I\theta_{k-1} + \vec{0}, \\
Y_k &= X_k^T\theta_k + \vec{\nu}_k,
\end{aligned}
$$

where θ_k is the state process and X_k is known.

Using the Kalman filter, we obtain $\hat{\theta}(n \mid n)$, which coincides with the least-square estimate $\hat{\theta}_n$ of (3) The distinction of the Kalman filter is that is it recursive: we can recycle the computation until the previous time step for the current time step.

[3]Think of Y_k as the analogue of Y_n and $\hat{X}(k \mid k)$ as the analogue of $\hat{X}_n(X_0, Y_1, \ldots, Y_n)$.