System Identification

Supplementary notes: lecture 2

Roy Smith

2 Data fitting and statistics

2.1 Bias, variance, and the MSE

The derivation of the result,

$$\text{MSE}(\hat{G}) \;=\; \text{var}\Big(\hat{G}\Big) \;+\; \text{Bias}^2(\hat{G}).$$

is relatively straightforward.

$$\begin{aligned}
MSE(\hat{G}) \;&=\; E\left\{|G \,-\, \hat{G}|^2\right\} \\
&=\; E\left\{\Big|(G - E\{\hat{G}\}) \,-\, (\hat{G} - E\{\hat{G}\})\Big|^2\right\} \\
&=\; E\left\{|G - E\{\hat{G}\}|^2\right\} \;+\; E\left\{|\hat{G} - E\{\hat{G}\}|^2\right\} \qquad \text{(note } |G-E\{G\}| \text{ is a constant)} \\
&\qquad -\; E\left\{2\,\text{real}\Big((G - E\{\hat{G}\})(\hat{G} - E\{\hat{G}\})^*\Big)\right\} \\
&=\; \text{Bias}^2(\hat{G}) \;+\; \text{var}\Big(\hat{G}\Big) \\
&\qquad -\; E\left\{2\,\text{real}\Big((G - E\{\hat{G}\})(\hat{G} - E\{\hat{G}\})^*\Big)\right\}
\end{aligned}$$

To show that the last term is actually zero:

$$\begin{aligned}
E\left\{(G - E\{\hat{G}\})(\hat{G} - E\{\hat{G}\})^*\right\} \;&=\; \\
&\qquad E\left\{G\hat{G}^* - GE\{\hat{G}^*\} - E\{\hat{G}\}\hat{G}^* + E\{\hat{G}\}E\{\hat{G}^*\}\right\} \\
&=\; GE\{\hat{G}^*\} - GE\{\hat{G}^*\} - E\{\hat{G}\}E\{\hat{G}^*\} \;+\; E\{\hat{G}\}E\{\hat{G}^*\} \\
&=\; 0.
\end{aligned}$$

2.2 Maximum-Likelihood Estimation

If our model errors are assumed to be Gaussian, then it is straightforward to use Maximum Likelihood to estimate linear model parametrisations. Consider a model of the form,

$$y \;=\; X\,\theta \;+\; v, \tag{1}$$

where $y \in \mathcal{R}^N$ is our vector of measurements, and v is a noise vector. The parameter vector, $\theta \in \mathcal{R}^d$, is linearly related to the output measurements, y. The matrix X may be a function of the measurements and the input, but in the linear model case it cannot be a function of θ or v. We also assume that the noise, v, has known covariance,

$$\Sigma_v \;:=\; E\{\,(v \,-\, \mu_v \mathbf{1})\,(v \,-\, \mu_v \mathbf{1})^T\}.$$

We are assuming that $v(k)$ is stationary and so it has a constant mean-value, μ_v. Note that $\Sigma_v \in \mathcal{R}^{N\times N}$ and $\Sigma_v = \Sigma_v^T > 0$. If the the noise, $v(k)$, is identically distributed then,

$$\Sigma_v \;=\; \sigma_v^2 I_N,$$

where σ_v^2 is the variance of each $v(k)$. Using the more general covariance matrix, Σ_v, allows us to consider noise modeled as filtered white noise. We can consider frequency dependent noise in this framework.

If $v(k)$ is modeled as filtered white noise, then we can write down the probability density function of the probability distribution for the vector, v,

$$f(v) \;=\; \frac{1}{2\pi^{N/2}\,|\det(\Sigma_v)|^{1/2}}\;\mathrm{e}^{\;-\frac{1}{2}(v-\mu_v\mathbf{1})^T\Sigma_v^{-1}(v-\mu_v\mathbf{1})}.$$

This is a multi-variate Gaussian distribution.

Using the model equation in (1), we can write down a probability density function for y as a function of the parameter θ. This is simply,

$$f(y;\theta) \;=\; \frac{1}{2\pi^{N/2}\,|\det(\Sigma_v)|^{1/2}}\;\mathrm{e}^{\;-\frac{1}{2}(y-X\theta-\mu_v\mathbf{1})^T\Sigma_v^{-1}(y-X\theta-\mu_v\mathbf{1})}.$$

Substituting a given measurement vector, y, into $f(y;\theta)$ gives the Likelihood function,

$$\mathcal{L}(\theta) \;=\; f(y;\theta)\;|_{y=\text{measured data vector}}\,.$$

The Maximum Likelihood estimate, θ_{ML}, is the value of θ that maximises the Likelihood function, or equivalently, the value of θ that maximises the log-likelihood function.

$$\begin{aligned}\theta_{\mathrm{ML}} \;=\; \underset{\theta}{\mathrm{argmax}}\,\mathcal{L}(\theta) \;&=\; \underset{\theta}{\mathrm{argmax}}\;\ln\left(\mathcal{L}(\theta)\right)\\ &=\; C \;-\; \frac{1}{2}(y-X\theta-\mu_v\mathbf{1})^T\Sigma_v^{-1}(y-X\theta-\mu_v\mathbf{1}).\end{aligned}$$

The constant term,

$$C \;=\; \ln\left(\frac{1}{2\pi^{N/2}\,|\det(\Sigma_v)|^{1/2}}\right)$$

is independent of θ and so doesn't affect the value of θ achieving the maximum.

To determine $\theta_{\rm ML}$ we differentiate $\ln(\mathcal{L}(\theta))$ and find the values of θ making the derivative equal to zero. The derivative with respect to θ is,

$$\frac{\partial \ln(\mathcal{L}(\theta))}{\partial \theta} \;=\; (y \,-\, X\theta - \mu_v \mathbf{1})^T \Sigma_v^{-1} X.$$

Setting this equal to zero implies that the Maximum Likelihood estimate must satisfy,

$$X^T \Sigma_v^{-1} X\, \theta_{\rm ML} \;=\; X^T \Sigma_v^{-1} (y \,-\, \mu_v \mathbf{1}).$$

In theory, this can be written as,

$$\theta_{\rm ML} \;=\; \left(X^T \Sigma_v^{-1} X\right)^{-1} X^T \Sigma_v^{-1} (y \,-\, \mu_v \mathbf{1}),$$

although from a numerical point of view, this is a rather poor way of performing the calculation. Note that for $\theta_{\rm ML}$ to be unique we require that

$$\text{rank}(X) \;\geqslant\; d.$$

A numerically better approach to calculating the solution is to use numerically stable methods to find the least-squares solution to,

$$\Sigma_v^{-1/2} X\, \theta_{\rm ML} \;=\; \Sigma_v^{-1/2} (y \,-\, \mu_v \mathbf{1}).$$

The square-root of Σ_v can be calculated via a Cholesky decomposition.

For models where either $f(y;\theta)$, or any monotone increasing function of $f(y;\theta)$, is linear in θ, the Maximum Likelihood estimate is easily derived via linear algebra. Unfortunately such models are not all that common. The approach of maximising the Likelihood (or log-likelihood) function still applies to models that are non-linear in θ, it's just much harder to do—both theoretically and numerically.