227-0689-00L: System Identification HS2023

Michele Zaffalon
Bruker BioSpin AG

November 30, 2023

Unreviewed notes for Prof. Smith's class. Use at own risk!

Contents

Chapter 1

Least-Squares Estimation

Consider a model

$$Y = \Phi\theta_0 + \epsilon \tag{1.1}$$

where $Y = \begin{bmatrix} y_0 & \dots & y_{N-1} \end{bmatrix}^\top$ is typically a vector containing the measurements $\{y_0, \dots, y_{N-1}\}$, $\theta \in \mathbb{R}^p$ is the unknown parameter to be estimated, $\Phi \in \mathbb{R}^{N\times p}$ is called the regressor. The noise vector $\epsilon = \begin{bmatrix} v_0 & \dots & v_{N-1} \end{bmatrix}^\top$ has zero mean $E\{\epsilon\} = 0$ and covariance $E\{\epsilon\epsilon^\top\} = R$, a symmetric positive definite matrix.

The maximum likelihood probability[1] requires us to minimize

$$(Y - \Phi\theta)^\top R^{-1}(Y - \Phi\theta) = ||C(Y - \Phi\theta)||_2^2 \tag{1.2}$$

where C is the Cholesky decomposition of the symmetric positive definite matrix $C^\top C = R^{-1}$.

The minimum of eq. (1.2) is found by setting the gradient of the expression with respect to θ to zero, which gives the equation

$$\left(\Phi^\top R^{-1}\Phi\right)^{-1} \hat{\theta}_{\mathrm{ML}} = \Phi^\top R^{-1} Y.$$

This is called the normal equation. The *mathematical* solution is given by

$$\hat{\theta}_{\mathrm{ML}} = \left(\Phi^\top R^{-1}\Phi\right)^{-1} \Phi^\top R^{-1} Y \tag{1.3}$$

and it exists provided that $C\Phi$ has full rank[2]: $\mathrm{rank}(C\Phi) = p$. When this is

[1] What is the average value of eq. (1.2)? I expect this to be equal to N. At least this is what happens with a diagonal noise covariance matrix.

[2] The usual warning holds for the rank. Instead we want to have the matrix $C\Phi$ with the smallest condition number for the estimate to be numerically stable, which is a stronger condition than full rank, to uniquely determine the best estimate $\hat{\theta}_{\mathrm{ML}}$.

the case, the system is said to be *persistently excited*, see Sect.. Given the freedom to choose Φ, one must select it such that $C\Phi$ is persistently excited.

Numerically one should not form the normal equation directly because it squares the condition number of $C\Phi$ and rely either on the QR decomposition or on the SVD to solve eq. (1.2). This is taken care automatically by MATLAB when using the backslash \ operator

$$\hat{\theta}_{\mathrm{ML}} = (C\Phi)\backslash(CY). \tag{1.4}$$

From this, it is clear that the quantities that matter are $C\Phi$ and CY.

1.1 Bias and Covariance of the Least-Squares Estimation

The linear estimator eq. (1.3) is unbiased. To see why,

$$\hat{\theta}_{\mathrm{ML}} = \left(\Phi^\top R^{-1}\Phi\right)^{-1}\Phi^\top R^{-1}(\Phi\theta_0+\epsilon) = \theta_0 + \left(\Phi^\top R^{-1}\Phi\right)^{-1}\Phi^\top R^{-1}\epsilon$$

using eq. (1.1) into eq. (1.3), and

$$E\{\hat{\theta}_{\mathrm{ML}}\} = \theta_0.$$

The covariance[3] is[4]

$$\mathrm{cov}\{\hat{\theta}_{\mathrm{ML}}\} = \left(\Phi^\top R^{-1}\Phi\right)^{-1}.$$

Note that *only* in the case $R = \sigma^2 I$ (diagonal matrix with all elements equal) the covariance reduces to σ^{-2}.

[3] Recalling that $R = R^\top$, $\left(\Phi^\top R^{-1}\Phi\right)^\top = \Phi^\top R^{-1}\Phi$, we have

$$\begin{aligned}
\mathrm{cov}\{\hat{\theta}_{\mathrm{ML}}\} &= E\left\{(\hat{\theta}_{\mathrm{ML}}-\theta_0)(\hat{\theta}_{\mathrm{ML}}-\theta_0)^\top\right\} \\
&= E\left\{\left(\left(\Phi^\top R^{-1}\Phi\right)^{-1}\Phi^\top R^{-1}\epsilon\right)\left(\left(\Phi^\top R^{-1}\Phi\right)^{-1}\Phi^\top R^{-1}\epsilon\right)^\top\right\} \\
&= E\left\{\left(\Phi^\top R^{-1}\Phi\right)^{-1}\Phi^\top R^{-1}\epsilon\epsilon^\top R^{-1}\Phi\left(\Phi^\top R^{-1}\Phi\right)^{-1}\right\} \\
&= \left(\Phi^\top R^{-1}\Phi\right)^{-1}\Phi^\top R^{-1}E\left\{\epsilon\epsilon^\top\right\}R^{-1}\Phi\left(\Phi^\top R^{-1}\Phi\right)^{-1} \\
&= \left(\Phi^\top R^{-1}\Phi\right)^{-1}.
\end{aligned}$$

[4] Is there a way to understand the form of covariance matrix without going through the calculation?

1.1.1 Geometric Interpretation of Least-Squares

The least squares problem

$$||b - Ax||_2$$

has the following geometric interpretation: the solution is that for which the residuals $v \doteq b - Ax$ are outside (i.e. orthogonal) of the space spanned by A. In other words, we require[5] the scalar product $\langle Az, v\rangle$ to be zero for all z:

$$\begin{aligned} 0 = \langle Az, v\rangle = (Az)^\top (b - Ax) = z^\top \left(A^\top b - A^\top Ax\right) \qquad \forall z \\ \rightarrow A^\top Ax = A^\top b \end{aligned}$$

which is the normal equation.

1.2 Random: The Covariance Matrix and the Choice of the Measurement Points

- The *off-diagonal* elements of the covariance matrix $\mathrm{cov}(\theta)$ represent the correlations between the errors of the variables θ. It is therefore not justified to discard them and take θ's standard deviations as the square root of $\mathrm{cov}(\theta)$'s diagonal elements because one discards the correlations: the ball of probability is in general an ellipse with the axes not parallel to the variable directions.

- Given the freedom to choose the measurement points, is there a "best" way of placing them?

  There are two factors that determine the covariance matrix: the choice of basis and the choice of points. The choice of basis is determined by the variables that one wants to extract: a linear transformation between one basis and the other will also transform the covariance matrix and the only concern may be the numerical stability (although one should expect that the measurement errors dominate).

  Determining the position of the measurement points by minimizing one (or more elements of the covariance) is in general a non-convex problem. However this is done.

[5] I believe the proof given in class is not correct: to span the full column space of A, one has to multiply by a generic vector z; in class $z = x$ was taken.