227-0689-00L: System Identification HS2023

Michele Zaffalon
Bruker BioSpin AG

December 4, 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 a vector containing the measurements $\{y_0, \dots, y_{N-1}\}$, $\Phi \in \mathbb{R}^{N \times p}$ is called the regressor and $\theta \in \mathbb{R}^p$ is the model parameter, the vector of p unknown parameters to be estimated; p is the model order. The noise vector $\epsilon = \begin{bmatrix} v_0 & \dots & v_{N-1} \end{bmatrix}^\top$ has zero mean $\mathbb{E}\left\{\epsilon\right\} = 0$ and covariance $\mathbb{E}\left\{\epsilon\epsilon^\top\right\} = R$, a symmetric positive definite matrix.

The maximum likelihood (ML) 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 $C^\top C = R^{-1}$.

[1] What is the expectation value of eq. (1.2)? Using eq. (1.5), I *think* the term $Y - \Phi\theta$ should be evaluated as

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

The average

$$\mathbb{E}\left\{(Y - \Phi\hat{\theta}_{\mathrm{ML}})^\top R^{-1}(Y - \Phi\hat{\theta}_{\mathrm{ML}})\right\} = \mathbb{E}\left\{\epsilon^\top \left(I - \Phi K \Phi^\top R\right)^\top R^{-1} \left(I - \Phi K \Phi^\top R\right)\epsilon\right\}$$

but I do not know what the term $\mathbb{E}\left\{\epsilon^\top R^{-1}\epsilon\right\}$ evaluates to. I expect this to be equal to N, at least this is what happens with an uncorrelated noise with constant variance.

Note also that while the matrix $vv^\top$ has rank 1, the variance matrix R for Gaussian distributed white noise is full rank because it is the sum of random vectors that span the whole space.

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

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

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]: $\operatorname{rank}(C\Phi) = p$. When this is 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.

[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 $\hat{\theta}_{\mathrm{ML}}$.

1.1 Bias, Covariance and MSE of the Least Squares Estimation

The linear estimator eq. (1.3) is unbiased[3]:

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

The covariance[4]

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

Note that *only* in the case of diagonal (*e.g.* uncorrelate) with all elements equal $R = \sigma^2 I_N$ the covariance reduces to $\sigma^2\left(\Phi^\top\Phi\right)^{-1}$. Lastly, we consider the mean squared error

$$\mathrm{MSE}\left(\hat{\theta}_{\mathrm{ML}}\right) = \underbrace{\left|\left|\mathrm{Bias}\left(\hat{\theta}_{\mathrm{ML}}\right)\right|\right|_2^2}_{=0} + \mathrm{tr}\left(\operatorname{cov}\left(\hat{\theta}_{\mathrm{ML}}\right)\right)$$

which reduces to $\sigma^2 N$ for the case of uncorrelated noise.

1.1.1 Geometric Interpretation of Least-Squares

The least squares problem

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

[3] Recalling that R and $\Phi^\top R^{-1}\Phi$ are symmetric matrices, letting $K \doteq (\Phi^\top R^{-1}\Phi)^{-1}$ and using eq. (1.1) into eq. (1.3), we obtain

$$\hat{\theta}_{\mathrm{ML}} = K\Phi^\top R^{-1}(\Phi\theta_0 + \epsilon) = \theta_0 + K\Phi^\top R^{-1}\epsilon \tag{1.5}$$

from which $\mathbb{E}\left\{\hat{\theta}_{\mathrm{ML}}\right\} = \theta_0$ since $\mathbb{E}\{\epsilon\} = 0$. Moreover

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

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

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:

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

which is the normal equation.

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

These are my considerations that are not part of the lecture.

- The *off-diagonal* elements of the covariance matrix $\operatorname{cov}(\theta)$ 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 $\operatorname{cov}(\theta)$'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.

Chapter 2

Regularized FIR Models

We have seen that the solution $\hat{\theta}_{\mathrm{ML}}$ of the least squares problem eq. (1.2)

$$\arg\min_{\theta} ||y - \Phi\theta||_2^2$$

is unbiased, see Sect. 1.1.

We can however choose to have a biased estimate to reduce the mean square error

$$\mathrm{MSE}\left(\hat{\theta}\right) = E||\hat{\theta} - \theta_0||_2^2.$$

This can be achieved if we modify the minimization problem by adding a regularization term

$$||Y - \Phi\theta||_2^2 + \gamma\theta^\top P^{-1}\theta \tag{2.1}$$

where γP^{-1} is a positive definite matrix. P is called the *kernel* or the regularization matrix[1].

Regularization prevents overfitting, reduces the sensitivity to noise and can improve the estimate by a proper choice of the kernel (*e.g.* for instance if the system is known to be stable, this information can be used to improve the estimation.)

The closed form mathematical solution to eq. (2.1) is given by

$$\hat{\theta}^{\mathrm{R}} = \left(\Phi^\top\Phi + \gamma P^{-1}\right)^{-1} \Phi^\top Y. \tag{2.2}$$

[1] When $P = \mathbb{I}_N$ the technique is called ridge regression, otherwise it goes under the name of Tikhonov regularization.

2.0.1 The James-Stein Estimator

The James-Stein estimator was the first biased estimator that had a smaller MSE compared to least squares on all the [1, page 3]. It can be cast into a ridge regression problem

$$||Y-\theta||^2+\gamma||\theta||^2, \qquad \gamma=\frac{(N-2)\sigma^2}{||Y||^2-(N-2)\sigma^2}$$

2.1 Bias, Covariance and MSE of the Regularized Least Squares Estimation

We assume uncorrelated noise[2] with constant variance σ^2: $\mathbb{E}\left\{\epsilon\epsilon^\top\right\}=\sigma^2 I_N$.

The estimate eq. (2.2) has bias, covariance[3] and MSE

$$\begin{aligned}
\text{Bias}\left(\hat{\theta}^{\mathrm{R}}\right) &= -\left(\Phi^\top\Phi+\gamma P^{-1}\right)^{-1}\gamma P^{-1}\theta_0\\
\text{cov}\left(\hat{\theta}^{\mathrm{R}}\right) &= \sigma^2\left(\Phi^\top\Phi+\gamma P^{-1}\right)^{-1}\Phi^\top\Phi\left(\Phi^\top\Phi+\gamma P^{-1}\right)^{-1}\\
\text{MSE}\left(\hat{\theta}^{\mathrm{R}}\right) &= \left|\left|\text{Bias}\left(\hat{\theta}^{\mathrm{R}}\right)\right|\right|_2^2+\text{tr}\left(\text{cov}\left(\hat{\theta}^{\mathrm{R}}\right)\right)
\end{aligned}$$

The MSE is minimized by this choice of parameters[4]

$$\gamma^\star=\sigma^2, \quad P^\star=\theta_0\,\theta_0^\top.$$

The optimal regularization matrix $P^\star$ is unknown because it depends on the unknown θ_0 but the approximate knowledge of the solution helps to construct a “good” P. Compared to the ridge regression which only tries to decrease $||\theta||^2$, the Tikhonov regularization can use information about the system.

[2] In Sect. 1 we considered the more general case of correlated noise $\mathbb{E}\left\{\epsilon\epsilon^\top\right\}=R$.

[3] Let $K=\left(\Phi^\top\Phi+\gamma P^{-1}\right)^{-1}$ so that $\hat{\theta}^{\mathrm{R}}=K\Phi^\top Y$. Plugging $Y=\Phi\theta_0+\epsilon$, we have

$$\hat{\theta}^{\mathrm{R}}-\mathbb{E}\left\{\hat{\theta}^{\mathrm{R}}\right\}=K\Phi^\top\epsilon$$

and

$$\text{cov}\left(\hat{\theta}^{\mathrm{R}}\right)=\mathbb{E}\left\{\left(K\Phi^\top\epsilon\right)\left(K\Phi^\top\epsilon\right)^\top\right\}=K\Phi^\top\mathbb{E}\left\{\epsilon\epsilon^\top\right\}\Phi K=\sigma^2K\Phi^\top\Phi K.$$

[4] Even for the $Y=\theta+\epsilon$ and an impulse response, so that Φ has only ones on the main diagonal (but it is not the identity matrix because in general Φ is a tall matrix), the forms of the bias and covariance for the optimal parameters do not have a special form.

2.1.1 Estimation Bias

As far as I understand, there are two sources of bias:

- for an FIR system, I can underestimate the length of the response by choosing the model order too small, $p < \tau_{\max}$: this is the truncation error.

  For least squares we said the estimation is unbiased: this is the case only when the order is correctly selected, otherwise it is too biased (see slide 10–10). Indeed we are trying to fit the data as

  $$\hat{\theta} = \arg\min_{\theta} ||Y - \Phi_p \theta|| = \left(\Phi_p^\top \Phi_p\right)^{-1} \Phi_p^\top Y$$

  but the model is $Y = \Phi\theta_0 + \epsilon$. The bias then becomes

  $$\mathbb{E}\left\{\hat{\theta}\right\} = \left(\Phi_p^\top \Phi_p\right)^{-1} \Phi_p^\top \Phi \theta_0 \neq \theta_0.$$

  (This means the trade-off bias-variance can also be done with LS by varying the order parameter, but to a lower extent than using regularization where one has extra parameters available.)

- With the order parameter correct, I can bias the estimator on purpose by setting it to a value different from θ_0.

2.1.2 Choice of the Regularization Matrices

A commonly used regularization matrix is the diagonally-correlated (DC) kernel

$$[P]_{ij} = c\alpha^{\frac{i+j}{2}} \rho^{|i-j|}$$

with $0 \leq c$, $0 \leq \alpha \leq 1$ and $-1 \leq \rho \leq 1$. A simpler version of it is the tuned-correlated (TC) kernel where one lets $\rho = \sqrt{\alpha}$ in the DC kernel

$$[P]_{ij} = c\alpha^{\max\{i,j\}} = \begin{bmatrix} \alpha & \alpha^2 & \alpha^3 & \dots & \alpha^n \\ \alpha^2 & \alpha^2 & \alpha^3 & \dots & \alpha^n \\ \alpha^3 & \alpha^3 & \alpha^3 & \dots & \alpha^n \\ \dots & & & & \\ \alpha^n & \alpha^n & \alpha^n & \dots & \alpha^n \end{bmatrix}$$

The top left entries are large, the bottom right small. Since the inverse enters the minimization problem (cost function), the top left part is small whereas the bottom down are large and induce a larger cost.

The inverse of the TC kernel is a tridiagonal with the explicit form

$$[P{-}1]_{ij} = \frac{c_{ij}}{1-\rho^2}(-1)^{i+j}\alpha^{-\frac{i+j}{2}}\rho^{|i-j|}, \qquad c_{ij} = \begin{cases} 1+\rho^2 & \text{if } i=j=2,\dots,p-1 \\ 0 & \text{if } |i-j|>1 \\ 1 & \text{otherwise} \end{cases}$$

2.1.3 Why the LS May Perform Badly

Least-squares performs badly when the regression matrix Φ is ill-conditioned[5], as a result of not being persistently excited and a small perturbation of the measurement data Y can have a large impact on the estimate $\hat{\theta}_{\mathrm{ML}}$. Regularization helps in these cases. (Is there an easy way to show that $\Phi^\top\Phi + \gamma P^{-1}$ has a larger minimum singular value?)

2.2 Cross-Validation Methods

They are methods to select the discrete model orders (in least-squares or ARX) or the regularization parameters (*e.g.* γ and the kernel parameter α in regularized least squares).

A popular method is called hold-out cross-validation and consists of the following steps:

1. split the data into two (equal but not necessarily) non-overlapping parts, one for estimation and the other for validation. The way in which the data is split depends on the type of data: for data generated by dynamical systems, where the data is sequential in time, the order is important. Here the first N_{id} pairs of control inputs and output values $\{u(k), y(k)\}$ are used for identification and the remaining for validation. For static data, the order is not important: one could choose the odd-indexed pairs for identification and the even-indexed pairs for validation;

2. use the estimation data to estimate a model for each mode structure or each regularization parameter. In the case of the continuous parameter

[5] The condition number of a matrix is defined as the ratio between the largest σ_1 and the smallest σ_n singular value of the matrix

$$\mathrm{cond}(A) = \frac{\sigma_1}{\sigma_n}.$$

Note that from a numerical point of view, the matrix being full rank is not a guarantee that the LS problem can be stably solved.

γ, one estimates the model on a set of gridded values. If there are two continuous parameters, *e.g.* γ and α, the gridding is done for both parameters;

3. select the model structure / regularization parameters that give a model with least prediction error on validation data;

4. use the selected model structure / regularization parameters and the complete data set to estimate a final model.

2.3 Numerical Solution of the Tikhonov Regularization Problem

For the numerical solution I am aware of two methods:

1. Solve directly the normal equation

   $$\left(\Phi^\top \Phi + \gamma P^{-1}\right) \hat{\theta} = \Phi^\top y \tag{2.3}$$

   using MATLAB's backslash operator: although this squares Φ's condition number when constructing $\Phi^\top \Phi$, at least it does not require to explicitly construct the inverse;

2. Use the generalized SVD decomposition. We first manipulate eq. (2.1) to obtain

   $$||y - \Phi\theta||_2^2 + ||D\theta||_2^2$$

   by letting D be the Cholesky decomposition of γP^{-1}: that is $D^\top D = \gamma P^{-1}$.

   There exists different definitions of generalized SVD: MATLAB implements the call `[U,V,X,C,S] = gsvd(Φ, D)` where U and V are unitary matrices, and C and S are nonnegative diagonal matrices such that

   $$\Phi = UCX^\top \qquad D = VSX^\top \qquad C^2 + S^2 = I.$$

   Inserting the decompositions into eq. (2.3) gives

   $$X^\top \theta = (UC)^\top y$$

   which can be solved as $\theta = X^\top \backslash (UC)^\top y$;

3. Eq. (2.1) can also be rewritten as

   $$\left|\left|\begin{bmatrix}\Phi \\ D\end{bmatrix}\theta - \begin{bmatrix}y \\ 0\end{bmatrix}\right|\right|_2^2$$

   and solved by backslash.

   I tested the three approaches on prob. 8 and found the fastest to be the third which does not construct the normal equation and is therefore better conditioned at the small cost of increasing the problem size by the r rows of P^{-1}. The first method is twice slower and the second is 100 (!) times slower: the problem is most likely the call to `gsvd`. This is strange because this is the method normally suggested, but probably only for larger regularization matrices.

Bibliography

[1] *Regularized System Identification: Learning Dynamic Models from Data.* Springer, 2021.