227-0689-00L: System Identification HS2023

Michele Zaffalon (Bruker BioSpin AG)

Yannic Hofmann (ETHZ)

December 12, 2023

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

Contents

Chapter 1

Least-Squares Estimation

A word about notation: in class we used Y and Φ for the measurements and regressor matrix and ϵ for the (possibly correlated) noise. However because the noise is Gaussian, the equations below show that the terms scaled by the Cholesky decomposition C of the noise correlation matrix R are the relevant ones. With the scaling, the scaled noise e becomes uncorrelated and has variance $\sigma^2 = 1$. I personally like this approach.

Consider the model

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

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

The maximum likelihood (ML) probability[2] requires us to minimize

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

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

[2] What is the expectation value of eq. (1.2)? Using eq. (1.5)

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

The average

$$\mathbb{E}\left\{\left|\left|Y - \Phi\hat{\theta}_{\mathrm{ML}}\right|\right|_2^2\right\} = \mathbb{E}\left\{e^\top \left(I - \Phi K\Phi^\top\right)^\top \left(I - \Phi K\Phi^\top\right) e\right\} = \mathbb{E}\left\{e^\top \left(I - \Phi K\Phi^\top\right) e\right\}.$$

where C is the Cholesky decomposition of $R = CC^\top$ (and $R^{-1} = \left(C^{-1}\right)^\top C^{-1}$), $Y = C^{-1}\tilde{Y}$, $\Phi = C^{-1}\tilde{\Phi}$ and $e = C^{-1}\tilde{\epsilon}$. Note that the scaled noise e is uncorrelated and has unit variance:

$$\mathbb{E}\left\{ee^\top\right\} = \mathbb{E}\left\{C^{-1}\tilde{\epsilon}\tilde{\epsilon}^\top\left(C^{-1}\right)^\top\right\} = C^{-1}R\left(C^{-1}\right)^\top = I.$$

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

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

The solution exists provided Φ has full column rank[3]: $\mathrm{rank}\,(\Phi) = p$. When this is the case, the system is said to be *persistently excited*, see Sect.. and when given the freedom, one chooses Φ so that it is well conditioned.

Had we not scaled $\tilde{Y}$ and $\tilde{\Phi}$ by the noise covariance, the solution would have been

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

Numerically one should not form the normal equation directly because it squares the condition number of Φ 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 $\backslash$ operator

$$\hat{\theta}_{\mathrm{ML}} = \Phi\backslash Y. \tag{1.4}$$

The term $\mathbb{E}\left\{e^\top e\right\}$ evaluates to N. The other term evaluates to p: using $\Phi = U\Sigma V^\top$,

$$(\Phi^\top\Phi)^{-1} = \left(V\Sigma^\top U^\top U\Sigma V^\top\right)^{-1} = V\left(\Sigma^\top\Sigma\right)^{-1}V^\top$$

and

$$\Phi V\left(\Sigma^\top\Sigma\right)^{-1}V^\top\Phi^\top = U\Sigma\left(\Sigma^\top\Sigma\right)^{-1}\Sigma^\top U$$

where the term between the two U is a tall matrix of dimension $N \times p$ with ones on the top block's main diagonal (of dimension $p \times p$) and zeros in the bottom block.

[3]The usual warning holds for the rank. Instead we want to have the matrix Φ with the smallest condition number for the estimate to be numerically stable, which is a stronger condition than full rank.

1.1 Bias, Covariance and MSE of the Least Squares Estimation

The linear estimator eq. (1.3) is unbiased[4] (but see also Sect. 2.1.2):

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

The covariance[5]

$$\mathrm{cov}\left(\hat{\theta}_{\mathrm{ML}}\right) = \left(\Phi^\top \Phi\right)^{-1}. \tag{1.6}$$

Lastly, we consider the mean squared error

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

which reduces to N for the case of uncorrelated noise (I am not sure about this anymore).

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.

[4] Recalling that $\Phi^\top \Phi$ is a symmetric matrix, letting $K \doteq (\Phi^\top \Phi)^{-1}$ and using eq. (1.1) into eq. (1.3), we obtain

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

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

$$\begin{aligned}
\mathrm{cov}\left(\hat{\theta}_{\mathrm{ML}}\right) &\doteq \mathbb{E}\left\{\left(\hat{\theta}_{\mathrm{ML}} - \mathbb{E}\left\{\hat{\theta}_{\mathrm{ML}}\right\}\right)\left(\hat{\theta}_{\mathrm{ML}} - \mathbb{E}\left\{\hat{\theta}_{\mathrm{ML}}\right\}\right)^\top\right\} \\
&= \mathbb{E}\left\{\left(K\Phi^\top e\right)\left(K\Phi^\top e\right)^\top\right\} \\
&= \mathbb{E}\left\{K\Phi^\top e e^\top \Phi K\right\} = K\Phi^\top \mathbb{E}\left\{ee^\top\right\}\Phi K = K.
\end{aligned}$$

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

In other words, we require[6] 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 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 $\operatorname{cov}\left(\hat{\theta}_{\mathrm{ML}}\right)$ eq. (1.6) 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? Is this done in practice?

  In the context of experiment design, this is done.

  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 best measurement points by minimizing one (or more elements of the covariance) is in general a non-convex problem.

[6] 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

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

[1]When $P = \mathbb{I}_N$ the technique is called ridge regression, otherwise it goes under the name of Tikhonov regularization. Compared to the ridge regression which only tries to decrease $||\theta||^2$, the Tikhonov regularization can use information about the system, see Sect. 2.1.1.

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 with constant variance 1: $\mathbb{E}\left\{ee^\top\right\}=I_N$.

The estimate eq. (2.2) has bias[2], covariance[3] and MSE

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

2.1.1 Choice of the Regularization Matrices

The MSE is minimized by this choice of parameters

$$\gamma^\star=1, \quad P^\star=\theta_0\,\theta_0^\top. \tag{2.3}$$

The optimal regularization matrix $P^\star$ is unknown[4] because it depends on the unknown θ_0 but the approximate knowledge of the solution helps to construct a "good" P.

---

[2] Let $K=\left(\Phi^\top\Phi+\gamma P^{-1}\right)^{-1}$ so that $\hat{\theta}^{\text{R}}=K\Phi^\top Y$. Plugging $Y=\Phi\theta_0+e$, we have $\mathbb{E}\left\{\hat{\theta}^{\text{R}}\right\}=K\Phi^\top\Phi\theta_0$. The bias is

$$\mathbb{E}\left\{\hat{\theta}^{\text{R}}\right\}-\theta_0=(K\Phi^\top\Phi-I)\theta_0=K(\Phi^\top\Phi-K^{-1})\theta_0=-\gamma KP^{-1}\theta_0.$$

[3] With the same notation as before, we have $\hat{\theta}^{\text{R}}-\mathbb{E}\left\{\hat{\theta}^{\text{R}}\right\}=K\Phi^\top e$. The covariance is

$$\text{cov}\left(\hat{\theta}^{\text{R}}\right)=\mathbb{E}\left\{\left(K\Phi^\top e\right)\left(K\Phi^\top e\right)^\top\right\}=K\Phi^\top\mathbb{E}\left\{ee^\top\right\}\Phi K=K\Phi^\top\Phi K.$$

[4] $P^\star=\theta_0\,\theta_0^\top$ has rank 1 and is therefore not invertible. To get a well-

Regularization functions are designed to encode prior knowledge on the unknown system; some choices will result in better estimates. The choice can be either subjective or based on cross-validation and different regularization functions/kernels give rise to different model classes. Technically there is no wrong regularization function, but there is an optimal one (in the sense of minimum MSE).

A quadratic regularization function based on the TC kernel is well-suited for exponentially decaying pulse responses of stable linear systems. In that case, a high order FIR model is used to estimate the first significant part of the pulse response of the unknown system, ignoring the tail. An FIR system is always stable, but its coefficients may not be exponentially decaying. An impulse response experiment may be used to reveal some properties of the system if the system and experimental configuration allow it. [Mohamed Abdalmoaty on Moodle]

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 $-1 \leq \rho \leq 1$: ρ describes correlations α and the decays.

A simpler version of it is the tuned-correlated (TC) kernel[5] where one lets $\rho = \sqrt{\alpha}$ in the DC kernel

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

defined problem, the inverse P^{-1} is replaced by the Moore-Penrose pseudoinverse P^{+}. It turns out that the solution of the problem defined with $P^{\dagger}$ is also equal to $(P\Phi^{\top}\Phi + \gamma I)^{-1}P\Phi^{\top}Y$. [Mohamed Abdalmoaty on Moodle]

[5]Despite having entries that are exponential, the TC kernel is *not* $\theta\theta^{\top}$

$$\theta\theta^{\top} = \begin{bmatrix} a & a^2 & a^3 & \dots \\ a^2 & a^3 & \dots & \\ a^3 & \dots & & \\ \dots & & & \end{bmatrix} \quad \text{with } \theta = \begin{bmatrix} a & a^2 & \dots \end{bmatrix}$$

coming from the truncation of the IIR filter $\frac{z^{-1}}{1-az^{-1}} \approx z^{-1}(1 + az^{-1} + a^2z^{-2} + \dots)$.

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 DC kernel is a tridiagonal with the explicit form

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

2.1.2 Estimation Bias

> One cannot say that the linear least-squares method is unbiased in general. There are three elements at play: the data set, the model used, and the estimation method. Least-squares is just the estimation method. To check if it is unbiased, an assumption has to be made on the true system that generated the data, and the model used to get the closed form expression of the LS estimator. [Mohamed Abdalmoaty on Moodle]

In the following, the input u is known, the measurement noise e has zero-mean, the data is generated as $Y = \Phi\theta_0 + e$ with $\theta_0 \in R^p$; we estimate $\theta \in R^q$ in the linear regression model $Y = \Phi_q\theta + e$; and we use the least-squares estimation method: then $\hat{\theta} = \left(\Phi_q^\top \Phi_q\right)^{-1} \Phi_q^\top Y$.

Here some examples of source of biased estimate:

- The model does not match the generating data. One underestimates the length of the response by choosing the model order too small, $q < p = \tau_{\max}$: this is the truncation error of Sect.. (see slide 10–10). Indeed we are trying to fit the data with a model of order q but the model has order p. The estimate is biased because

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

  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.

  Note that here the regressor involves only the known (noise-free) input u;

- The data comes from an ARX system with $p = n + m$ parameters and $\theta_0 \in R^{n+m}$ is the true parameter. As opposed to the case above, here

Φ involves also the random outputs: for this reason the resulting LS estimator is *biased* even though we have a correct model order/structure (*i.e.*, the model that matches the data generating system);

- Finally, even with the correct order parameter, one can *on purpose* bias $\hat{\theta}$ by adding the regularization term to $||Y - \Phi\theta||_2$.

2.1.3 Why the LS May Perform Badly

Least-squares performs badly when the regression matrix Φ is ill-conditioned[6], 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 $\hat{\theta}_{\text{ML}}$. 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

[6]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

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

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.

Numerical example in `10_lect/regularization.m`, for the moment limited to ridge regression.

2.3 Numerical Solution of the Tikhonov Regularization Problem

For the numerical solution I am aware of three methods:

1. Solve directly the normal equation

$$\left(\Phi^\top \Phi + \gamma P^{-1}\right) \hat{\theta} = \Phi^\top y \tag{2.4}$$

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. (2.4) 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.

Chapter 3

Error Prediction Methods and Transfer Function Models

We assume that the complete system model is given by

$$y(k) = G(z)u(k) + H(z)e(k). \tag{3.1}$$

Prediction error-based identification methods estimate the transfer functions $G(z)$ and $H(z)$ by minimizing the objective function $J(\epsilon)$, a function of the prediction error $\epsilon(k)$

$$\epsilon(k) \doteq y(k) - \hat{y}(k|k-1).$$

The prediction $\hat{y}(k|k-1)$ for the time k is a function of the previous measurements and inputs at $k-1,\ k-2, \ldots$ only.

3.1 Prediction

We assume that $H(z)$ is monic[1] and stable[2]. In this case, given the filtered noise $v(k) = H(z)e(k)$, $e(k)$ can be reconstructed from $v(k)$ as

$$e(k) = H^{-1}(z)v(k). \tag{3.2}$$

[1] Monic means that $h(0) = 1$:

$$H(q) = 1 + \sum_{k=1}^{\infty} h(k)q^{-k}.$$

[2] $H(z)$ has only poles strictly inside the unit circle.

We now seek to predict $v(k)$ given the past values up to time $k-1$: this is called the *one-step ahead* estimate. Since $H(z)$ is monic, we split the filtered noise contribution into the term $e(k)$ and other terms up to time $k-1$

$$v(k) = H(z)e(k) = e(k) + (H(z) - 1)e(k) \tag{3.3}$$

The *predicted* filtered noise $\hat{v}(k|k-1)$ is

$$\hat{v}(k|k-1) = (H(z) - 1)e(k).$$

This can be intuitively understood because the error probability function distribution for $\{e\}$ has zero mean[3]: if we were left to guess $e(k)$, we would guess $e(k) = 0$. Making use of eq. (3.2), the one-step ahead estimate

$$\hat{v}(k|k-1) = \left(1 - H^{-1}(z)\right) v(k) \tag{3.4}$$

is determined only from the knowledge of the past values of v up to $k-1$.

For the model of eq. (3.1), the one-step ahead predictor

$$\hat{y}(k|k-1) = G(z)u(k) + \hat{v}(k|k-1)$$

can be rewritten with the help of the expression for $\hat{v}(k|k-1)$ as

$$\hat{y}(k|k-1) = H^{-1}(z)G(z)u(k) + \left(1 - H^{-1}(z)\right) y(k). \tag{3.5}$$

The prediction error[4]

$$\epsilon(k) = y(k) - \hat{y}(k|k-1) = H^{-1}(z)v(k) = e(k) \tag{3.6}$$

is the noise $e(k)$ that cannot be predicted: the *innovation* is the part of the output prediction that cannot be estimated from past measurements.

[3]Had the probability distribution $f_e(x)$ not had a zero mean, we would have to modify the predition according to

$$\hat{v}(k|k-1) = \arg\max_x f_e(x - m(k-1)), \qquad m(k-1) = (H(q) - 1)e(k).$$

[4]Using eq. (3.5)

$$\begin{aligned}
\epsilon(k) &= y(k) - \hat{y}(k|k-1) \\
&= y(k) - H^{-1}(z)G(z)u(k) - \left(1 - H^{-1}(z)\right) y(k) \\
&= H^{-1}(z)(y(k) - G(z)u(k)) \\
&= H^{-1}(z)v(k) = e(k)
\end{aligned}$$

3.1.1 Example: Moving Average

The model

$$v(k) = e(k) + ce(k-1) \rightarrow H(z) = 1 + cz^{-1}$$

is invertible when $|c| < 1$. The one-step ahead predictor eq. (3.4) can be expressed in terms of the error $e(k-1)$ using eq. (3.6)

$$\hat{v}(k|k-1) = \left(1 - H^{-1}(z)\right) H(z)e(k) = cz^{-1}e(k) = ce(k-1)$$

3.2 Family of Transfer-Function Models

The advantage of the transfer-function models is that they can be described by fewer parameters and that the inputs required to identify the system do not have to be persistently exciting as it is the case when one wants to identify frequency or time-response: we have seen in Sect. **??** that for frequency domain methods, the order of excitation must be double the number of complex estimates of the transfer function $G(e^{j\omega_n})$ since gain and phase must be determined for each frequency; for a time response one requires the same persistency order as the number of impulse response terms.

If we control the input, this requirement is easy to satisfy. If on the other hand the data is given, this may not be the case and in these situations, one is better off looking for transfer functions/state space representations because of the reduced numbers of parameters to identify.

Prediction error-based identification methods construct the prediction error

$$\epsilon(k, \theta) = y(k) - \hat{y}(k, \theta)$$

from the (one-step ahead) predictor $\hat{y}(k, \theta)$ which is based on the guesses $\hat{G}(z) = G(z, \theta)$ and $\hat{H}^{-1}(z) = H^{-1}(z, \theta)$, the guesses being parametrized by θ. The optimal $\theta^\star$ is the argument that minimizes the cost function $J = J(\epsilon)$

$$\theta^\star = \arg\min_\theta J(\epsilon(k, \theta)).$$

Typical choices for the objective functions $J(\epsilon)$ are the 2-norm $||\epsilon||_2^2$ or the maximum deviation, the ∞-norm $||\epsilon||_\infty$. The kind of minimization depends on how the models $G(q)$ and $H^{-1}(q)$ are parametrised: in general, the parametrization will not be linear and the optimization may not be convex. Note moreover that the minimization of $||\epsilon||_2^2$ is not equivalent to the least squres method, unless $H(z) = 1$:

$$y(k) = G(z)u(k) + e(k).$$

Since this approach does not require an a-priori knowlege of the system, it is also called *black-box* approach.

3.2.1 Equation Error Model Structure (ARX)

The ARX model[5]

$$y(k) = B(z)u(k) + (1 - A(z))\, y(k) + e(k) \tag{3.7}$$

is a simple input-output relationship where the error enters as a direct term. We take

$$A(z) = 1 + a_1 z^{-1} + \ldots + a_n z^{-n}, \qquad B(z) = b_1 z^{-1} + \ldots + b_m z^{-m}$$

note that $A(z)$ is monic and $B(z)$ does not have a constant term, *i.e.* the model has no feed-through. It corresponds to the model of eq. (3.1) with

$$G(z) = \frac{B(z)}{A(z)}, \qquad H(z) = \frac{1}{A(z)}.$$

and generates the one-step ahead predictor

$$\hat{y}(k|k-1) = (1 - A(z))\, y(k) + B(z)u(k) \tag{3.8}$$

either by plugging $G(z)$ and $H^{-1}(z)$ in eq. (3.5) or by using the result of eq. (3.6) that $\hat{y}(k|k-1) = y(k) - e(k)$ in eq. (3.7).

While the model covers a limited set of real-world problems (notably those where the perturbation act as a force), the predictor eq. (3.8) has the advantage of forming the linear regressor

$$\hat{y}(k, \theta) = \varphi^\top(k)\theta$$

the *parameter vector* θ being the unknown coefficients of the polynomials $A(z)$ and $B(z)$

$$\theta = \begin{bmatrix} a_1 & \ldots & a_n & b_1 & \ldots & b_m \end{bmatrix}^\top$$

and the *regressor vector* $\varphi(k)$ being the output and input terms

$$\varphi^\top(k) = \begin{bmatrix} -y(k-1) & \ldots & -y(k-n) & u(k-1) & \ldots & u(k-m) \end{bmatrix}.$$

[5] From $y(k) = \frac{B(z)}{A(z)}u(k)$, rearranging the terms

$$A(z)y(k) = B(z)u(k) \rightarrow y(k) = B(z)u(k) + (1 - A(z))y(k).$$

3.2.2 Estimate Bias

The ARX model structure is applied also to systems that do not have this noise model for the simple reason that it is linear. One must be aware that this induce a bias in the estimates $G(z, \theta^\star) = \frac{B(z,\theta^\star)}{A(z,\theta^\star)}$ and $H^{-1}(z, \theta^\star)$, see the numerical example in the slides 9.30.

3.2.3 ARMAX Model Structure

The ARX model structure is not very flexible with regards to the noise model: indeed it requires the noise to have the particular structure[6] $1/A(z)$. The ARMAX transfer function model is in the form

$$A(z)y(k) = B(z)u(k) + C(z)e(k) \tag{3.9}$$

where $A(z)$ and $B(z)$ are as in ARX and $C(z)$ is monic. It corresponds to the model of eq. (3.1) with

$$G(z) = \frac{B(z)}{A(z)}, \qquad H(z) = \frac{C(z)}{A(z)}$$

and with the one-step ahead predictor[7]

$$\hat{y}(k|\theta) = (1 - A(z))y(k) + B(z)u(k) + (C(z) - 1)\underbrace{(y(k) - \hat{y}(k|\theta))}_{\doteq\epsilon(k,\theta)}. \tag{3.10}$$

Introducing the regression vector

$$\varphi^\top(k, \theta) = \begin{bmatrix} -y(k-1) & \dots & u(k-1) & \dots & \epsilon(k-1, \theta) & \dots \end{bmatrix}$$

eq. (3.10) induces the pseudolinear regression

$$\hat{y}(k, \theta) = \varphi^\top(k, \theta)\theta.$$

[6] It has the advantage (together with FIR models) of forming a linear regressor.

[7] The one-step predictor is more easily obtained by plugging the expressions for $G(z)$ and $H^{-1}(z)$ into eq. (3.5).

Working out the expression directly is cumbersome: by using eq. (3.6) in the model eq. (3.9), we have

$$\hat{y}(k|k-1) = y(k) - e(k) = \frac{B(z)}{A(z)}u(k) + \left(\frac{C(z)}{A(z)} - 1\right)e(k).$$

By plugging in the expression for the noise

$$e(k) = C^{-1}(z)(A(z)y(k)) - B(z)u(k))$$

we obtain

$$C(z)\hat{y}(k|k-1) = B(z)u(k) + (C(z) - A(z))y(k).$$

Some manipulation is still required to bring it into the final form eq. (3.10).

3.2.4 Constrained Minimization

The optimization-based algorithm

$$\min_{\theta,\epsilon} ||\epsilon||_2$$
$$\text{subject to } Y = \Phi(\epsilon)\theta + \epsilon$$

has a non-linear dependence on θ in the affine constrained.

A reference implementation is in `11_lect/SysID_ARMAX.m`.

3.2.5 General Family of Model Structures

The most general family of model structure is [2, Sect. 4]

$$A(z)y(k) = \frac{B(z)}{F(z)}u(k) + \frac{C(z)}{D(z)}e(k) \tag{3.11}$$

Some of the common cases that we have seen so far are summarized in Table 3.1.

Polynomials	Name of Model Structure
B	FIR
A B	ARX
A B C	ARMAX
A B C D	ARARMAX
B F C D	Box-Jenkins

Table 3.1: Some common black-box SISO models using the polynomials of eq. (3.11).

Bibliography

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

[2] *System Indentification: Theory for the User*. Second edition. Prentice Hall PTR, 2009.