Chapter 7

Model Selection and Model Validation

"How can we be sure that an estimated model serves its future purpose well?"

Suppose I were a teacher, and you as a student had to work out a case-study of an identification experiment. The question now is how to verify wether you attempt is fruitful? Or are your efforts more fruitful than the ones of your colleagues? When is an approach not acceptable? The question of how to come up with a preference amongst different models for a given system is in practice more important than the actual method used for estimating one model: that is, even if your toolbox contains a method which gives a useless result on the task at hand, proper model selection will reveal this weak approach and prefer other tools. The aims of model selection are as follows

- Given a model class, parameter estimation techniques give you the best model in this class. Model selection on the other hand describes how to arrive at this model class at the first place. This amounts to the decision of (i) what sort of model structure suffices for our needs (e.g. ARX, BJ, State Space), and (ii) what model orders we would need (e.g. an ARX(1,1) or a ARX(100,10)).

- Which (stochastic) assumptions are reasonable to drive the analysis? Are the conditions valid under which the employed parameter estimation techniques 'work' in the studied case?

- Is the model we have identified in some sense close to the real system? Or perhaps more realistically, is the model we have identified sufficient for our needs? We will refer to this aim as *model validation*.

Of course those objectives will be entangled in practice, and be closely related to the parameter estimation task at hand. It is as such no surprise that the same themes as explored in earlier chapters will pop up in a slightly different form. In fact, more recent investigations argue for a closer integration of parameter estimation and model selection problems at once, a theme which we will explore in the Chapter on nonlinear modeling.

The central theme of model selection and model validation will be to avoid the effect of 'over-fitting'. This effect is understood as follows

Definition 25 (Overfitting) *If we have a large set of models in the model class with respect to the number of data, it might well be possible that the estimated model performs well on the data used to tune the parameters to, but that this model performs arbitrary bad in new cases.*

The following is a prototypical example

Example 47 (Fitting white noise) *Let $\{e_t\}_t$ be zero mean white noise with variance σ^2. Consider the system*

$$y_t = e_t, \ \forall t = -\infty, \dots, \infty, \tag{7.1}$$

and suppose we observe corresponding to y_t an input φ_t which is unrelated. Consider the case where the estimated model contains ('remembers') all mappings from observed inputs to corresponding outputs $\{\varphi_t \to y_t\}$. Then the estimated error on the set used for building up the model will be zero (i.e. it can be reconstructed exactly). The error on new data will be σ^2 in case $\ell(e) = e^2$.

The tasks of model selection and model validation is characterized by different trade-offs one has to make. A trade-off will in general arise as one has to pay a price for obtaining more accurate or complex models. Such trade-offs come into the form of variance of the estimates, complexity of the algorithms to be used, or even approaches which lead necessarily to unsuccessful estimates.

> "Essentially, all models are wrong, but some are useful", G.Box, 1987

This is a mantra that every person who deals with (finite numbers of) observed data implements in one way or another.

- *Bias-Variance Trade-off:* In general one is faced with a problem of recovering knowledge from a finite set of observations, referred to as an 'inverse problem'. If the model class is 'large' and contains the 'true' system, the bias of a technique might be zero, but the actual deviation of the estimated parameters from the true one might be large ('large variance'). On the other hand, if the model class is small, it might be easy to find an accurate estimate of the best candidate in this model class ('low variance'), but this one might be far off the 'true' system ('large bias). This intuition follows the bias-variance decomposition Lemma given as

  Lemma 8 (Bias-Variance Decomposition) *Let θ_n be an estimate of $\theta_0 \in \mathbb{R}^d$ using a random sample of size n, then*

  $$\mathbb{E}\|\theta_0 - \theta_n\|_2^2 = \mathbb{E}\|\theta_0 - \mathbb{E}[\theta_n]\|_2^2 + \mathbb{E}\|\mathbb{E}[\theta_n] - \theta_n\|_2^2, \tag{7.2}$$

  where $\mathbb{E}\|\theta_0 - \mathbb{E}[\theta_n]\|_2$ is often referred to as the bias*, and $\mathbb{E}\|\mathbb{E}[\theta_n] - \theta_n\|_2^2$ as the variance associated to the estimator θ_n.*

  This result follows directly by working out the squares as

  $$\mathbb{E}\|\theta_0-\theta_n\|_2^2 = \mathbb{E}\|(\theta_0-\mathbb{E}[\theta_n])+(\mathbb{E}[\theta_n]-\theta_n)\|_2^2 = \mathbb{E}\|(\theta_0-\mathbb{E}[\theta_n])\|_2^2+\mathbb{E}\|(\mathbb{E}[\theta_n]-\theta_n)\|_2^2+2\mathbb{E}[(\theta_0-\mathbb{E}[\theta_n])^T(\mathbb{E}[\theta_n]-\theta_n)] \tag{7.3}$$

  and since $\mathbb{E}[(\theta_0-\mathbb{E}[\theta_n])^T(\mathbb{E}[\theta_n]-\theta_n)]$ equals $(\theta_0-\mathbb{E}[\theta_n])^t\mathbb{E}[\mathbb{E}[\theta_n]-\theta_n] = ((\theta_0-\mathbb{E}[\theta_n]))^T 0_d = 0$, since θ_0 and $\mathbb{E}[\theta_n]$ are deterministic quantities.

- *Algorithmic Issues:* In practice, when the data can only point to a specific model in a model class with large uncertainty, the parameter estimation problem will often experience algorithmic or numeric problems. Of course, suboptimal implementations could give problems even if the problem at hand is not too difficult. Specifically, in case one has to use heuristics, one might want to take precautions against getting stuck in 'local optima', or algorithmic 'instabilities'.

The theory of algorithmic, stochastic or learning complexity studies the theoretical link between either, and what a 'large' or 'small' model class versus a 'large' number of observations mean. In our case it is sufficient to focus on the concept of Persistency of Excitation (PE), that is, a model class is not too large w.r.t. the data if the data is PE of sufficient order. This notion is in turn closely related to the condition number of the sample covariance matrix, which will directly affect numeric and algorithmic properties of methods to be used for estimation.

7.1 Model Validation

Let us first study the 'absolute' question: 'is a certain model sufficient for our needs?' Specifically, we will score a given model with a measure how well it serves its purpose. This section will survey some common choices for such scoring functions.

7.1.1 Cross-validation

A most direct, but till today a mostly unrivaled choice is to assess a given model on how well it performs on 'fresh' data. Remember that the error on the data used for parameter estimation might be spoiled by overfitting effects, that is, the model might perform well on the specific data to which the estimated model is tuned but can perform very badly in new situations. It is common to refer to the performance of the model on the data used for parameter estimation as the *training performance*. Then, the training performance is often a biased estimate of the actual performance of the model. A more accurate estimate of the performance of the model can be based on data which is in a sense independent of the data used before.

The protocol goes as follows

1. Set up a first experiment and collect signals $\{u_t\}_{t=1}^{n}$ and $\{y_t\}_{t=1}^{n}$ for $n > 0$;
2. Estimate parameters $\hat{\theta}_n$ (model) based on those signals;
3. Set up a new experiment and collect signals $\{u_t^v\}_{t=1}^{n^v}$ and $\{y_t^v\}_{t=1}^{n^v}$ for $n^v > 0$;
4. Let $\ell : \mathbb{R} \to \mathbb{R}$ as before a loss function. Score the model as

$$V_{n^v}^v(\hat{\theta}_n) = \frac{1}{n^v} \sum_{t=1}^{n^v} \ell\left(y_t^v - f_{\hat{\theta}_n,t}\right), \tag{7.4}$$

where $f_{\hat{\theta}_n,t}$ is the shorthand notation for the predictor based on the past signals $\{y_s^v\}_{s<t}$ and $\{u_s^v\}_{s\leq t}$, and the parameters $\hat{\theta}_n$.

The crucial bit is that inbetween the two experiments, the studied system is left long enough so that the second experiment is relatively 'independent' from what happened during the first one. This issue becomes more important if we have only one set of signals to perform estimation and validation on. A first approach would be to divide the signals in two non-overlapping, consecutive blocks of length (usually) 2/3 and 1/3. The first one is then used for parameter estimation ('training'), the second black is used for model validation. If the blocks are small with respect to the model order (time constants), transient effects between the training block and validation block might affect model validation. It is then up to the user to find intelligent approaches to avoid such effects, e.g. by using an other split training-validation.

7.1.2 Information Criteria

In case cross-validation procedures become too cumbersome, or lead to unsatisfactory results e.g. as not enough data is available, one may resort to the use of an Information Criterion (IC) instead. An information criterion in general tries to correct analytically for the overfitting effect in the training performance error. They come in various flavors and are often relying on statistical assumptions on the data. The general form of such an IC goes as follows

$$w_n = V_n(\theta_n)\left(1 + \beta(n,d)\right), \tag{7.5}$$

where $\beta : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ is a function of the number of samples n and the number of free parameters d. In general β should decrease with growing n, and increase with larger d. Moreover β should tend to zero if $n \to \infty$. An alternative general form is

$$\tilde{w}_n = n \log V_n(\theta_n) + \gamma(n,d), \tag{7.6}$$

where $\gamma : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ penalizes model structures with larger complexity. The choice of $\gamma(n,d) = 2d$ gives Akaike's IC (AIC):

$$\underset{n}{\mathrm{AIC}}(d) = n \ln V_n(\theta_n) + 2d, \tag{7.7}$$

where θ_n is the LS/ML/PEM parameter estimate. It is not too difficult to see that the form w_n and $\tilde{w}_n$ are approximatively equivalent for increasing n.

Several different choices for γ have appeared in the literature, examples of which include the following:

(FPE): The *Finite Prediction Error* (FPE) is given as

$$\mathrm{FPE}(d) = V_n(\theta_n)\frac{1 + d/n}{1 - d/n}, \tag{7.8}$$

which gives an approximation on the prediction error on future data. This criterium can be related to AIC's and the significance test for testing the adequacy of different model structures.

(BIC): The *Bayesian Information Criterion* (BIC) is given as

$$\mathrm{BIC}(d) = -2 \ln V_n(\theta_n) + d \ln(n), \tag{7.9}$$

where θ_n is the ML parameter estimate, and the correction term is motivated from a Bayesian perspective.

In general, an individual criterion can only be expected to give consistent model orders under strong assumptions. In practice, one typically choses a model which minimizes approximatively all criteria simultaneously.

7.1.3 Testing

In case a stochastic setup were adopted, a model comes with a number of stochastic assumptions. If those assumptions hold, the theory behind those methods ensures (often) that the estimates are good (i.e. efficient, optimal in some metric, or leading to good approximations). What is left for the practitioner is that we have to verify the assumptions for the task at hand. This can be approached using statistical significance testing. If such a test gave evidence that the assumptions one adopted do not hold in practice, it is only reasonable to go back to the drawing table or the library. The basic ideas underlying significance testing go as follows.

> 'Given a statistical model, the resulting observations will follow a derived statistical law. If the actual observations are not following this law, the assumed model cannot be valid.'

This inverse reasoning has become the basis of the scientific method. Statistics being stochastic is not about deterministic laws, but will describe how the observations would be like most probable. As such it is not possible to refute a statistical model completely using only a sample, but merely to accumulate evidence that it is not valid. Conversely, if no such evidence for the contrary is found in the data, it is valid practice to go ahead as if the assumptions were valid. Note that this does not mean that the assumptions are truly valid!

The translation of this reasoning which is often used goes as follows

$$Z_1, \ldots, Z_n \sim \mathcal{H}_0 \Rightarrow T(Z_1, \ldots, Z_n) \sim \mathcal{D}_n(\beta, \mathcal{H}_0), \tag{7.10}$$

where

n: The number of samples, often assumed to be large (or $n \to \infty$)

Z_i: An observation modeled as a random variable

$\sim$: means here '*is distributed as* '

$\Rightarrow$: Implies

$\mathcal{H}_0$: or the *null distribution*

$T(\ldots)$: or the *statistic*, which is in turn a random variable.

$\mathcal{D}_n(\beta, \mathcal{H}_0)$: or the *limit distribution* of the statistic under the null distribution $\mathcal{H}_0$. Here β denotes a parameter of this distribution.

Typically, we will have asymptotic null distributions (or 'limit distributions'), characterized by a PDF f. That is, assuming that n tends to infinity, $\mathcal{D}_n(\beta, \mathcal{H}_0)$ tends to a probability law with PDF $f_{\beta,\mathcal{H}_0}$. Shortly, we write that

$$T(Z_1, \ldots, Z_n) \to f_{\beta,\mathcal{H}_0}. \tag{7.11}$$

Now, given a realization $z_1, \ldots, z_n$ of a random variable $Z'_1, \ldots, Z'_n$, the statistic $t_n = T(z_1, \ldots, z_n)$ can be computed for this actual data. A statistical hypothesis test checks wether this value t_n is likely to occur in the theoretical null distribution $\mathcal{D}_n(\beta, \mathcal{H}_0)$. That is, if the value t_n were rather unlikely to occur under that model, one must conclude that the statistical model which were assumed to underly $Z_1, \ldots, Z_n$ are also not too likely. In such a way one can build up *evidence*

for the assumptions **not** to be valid. Each test comes with its associated $\mathcal{H}_0$ to test, and with a corresponding test statistic. The derivation of the corresponding limit distributions is often available in reference books and implemented in standard software packages.

A number of classical tests are enumerated:

F-test: The one-sample z-test checks wether a univariate sample $\{y_i\}_{i=1}^n$ originating from a normal distribution with given variance has mean value zero or not. The null-hypothesis is that the sample is sampled i.i.d. from zero mean Gaussian distribution with variance σ^2. The test statistic is computed as

$$T_n\left(\{y_i\}_{i=1}^n\right) = \frac{\sum_{i=1}^n y_i}{\sigma\sqrt{n}}, \tag{7.12}$$

and when the null-distribution were valid it would be distributed as a standard normal distribution, i.e. $T_n \to \mathcal{N}(0,1)$. In other words, if $Y_1, \ldots, Y_n$ are i.i.d. samples from f_0, then $f_T \to \mathcal{N}(0,1)$ when n tends to be large (in practice $n > 30$ is already sufficient for the asymptotic results to kick in!). Based on this limit distribution one can reject the null-hypothesis with a large probability if the test-statistic computed on the observed sample would have a large absolute value.

χ^2-test: Given is a set of n i.i.d. samples $\{y_i\}_{i=1}^n$ following a normal distribution. The standard χ^2-test checks wether this normal distribution has a pre-specified standard deviation σ_0. The test statistic is given as

$$T_n\left(\{y_i\}_{t=1}^n\right) = \frac{(n-1)s_n^2}{\sigma_0^2}, \tag{7.13}$$

where the sample variance is computed as $\frac{1}{n}\sum_{i=1}^n (y_i - m_n)^2$, and the sample mean is given as $\frac{1}{n}\sum_{i=1}^n y_i$. Then the limit distribution of this statistic under the null-distribution is known to follow a χ^2-distribution with $n-1$ degrees of freedom, the PDF and CDF of this distribution is computed in any standard numerical software package.

Example 48 (Lady Tasting Tea) *(Wikipedia) - The following example is summarized from Fisher, and is known as the Lady tasting tea example. Fisher thoroughly explained his method in a proposed experiment to test a Lady's claimed ability to determine the means of tea preparation by taste. The article is less than 10 pages in length and is notable for its simplicity and completeness regarding terminology, calculations and design of the experiment. The example is loosely based on an event in Fisher's life. The Lady proved him wrong.*

1. *The null hypothesis was that the Lady had no such ability.*
2. *The test statistic was a simple count of the number of successes in 8 trials.*
3. *The distribution associated with the null hypothesis was the binomial distribution familiar from coin flipping experiments.*
4. *The critical region was the single case of 8 successes in 8 trials based on a conventional probability criterion (< 5%).*
5. *Fisher asserted that no alternative hypothesis was (ever) required.*

If and only if the 8 trials produced 8 successes was Fisher willing to reject the null hypothesis effectively acknowledging the Lady's ability with > 98% confidence (but without quantifying her ability). Fisher later discussed the benefits of more trials and repeated tests.

7.1.4 Testing LTI Models

In the context of testing the results of a model of a linear system, the following tests are often used. The following setup is typically considered. Given timeseries $\{u_t\}_t$ and $\{Y_t\}_t$ as well as (estimated) parameters $\hat{\theta}$ of a model structure $\mathcal{M}$. Then, one can compute the corresponding optimal predictions $\{\hat{y}\}_t$ and the prediction errors (or residuals) $\{\hat{\epsilon}_t\}_t$ corresponding to this estimate $\hat{\theta}$. Now, a common statistical model assumes that $\epsilon_t = Y_t - \hat{y}_t(\theta_0)$ (the innovations) is zero mean, white stochastic noise. Then, if $\theta_0 \approx \hat{\theta}$, also the estimated innovations $\{\hat{\epsilon}_t\}_t$ would be similar to zero mean Gaussian noise. A statistical test could then be used to collect evidence for the $\{\hat{\epsilon}_t\}_t$ not too be a white noise sequence, hence implying that the parameter estimate is not adequate. Two typical tests which check the whiteness of a sequence $\{\epsilon_t\}_{t=1}^n$ go as follows.

- (Portmanteau):
$$T_n(\{\epsilon_i\}) = \frac{n}{\hat{r}_0^T} \sum_{\tau=1^m} \hat{r}_\tau^2, \tag{7.14}$$
where the sample auto-covariances are computed as $\hat{r}_\tau = \frac{1}{n}\sum_{i=1}^{n-\tau} \epsilon_i \epsilon_{i+\tau}$. This test statistic follows a χ^2 distribution with m degrees of freedom, if $\{\epsilon_t\}$ where indeed samples from a zero mean, white noise sequence. So, if the test-statistic computed using the estimated innovations $T_n(\{\hat{\epsilon}_t\})$ were really large, evidence is collected for rejecting the null-hypothesis - the estimate $\hat{\theta}$ were close to the true value θ_0. This reasoning can be quantified exactly using the above expressions.

- (Normal): A simple test for checking where a auto-covariance at a lag $\tau > 0$ is zero based on a sample of size n is given by the statistic
$$T_n(\{\epsilon_i\}) = \sqrt{n}\frac{\hat{r}_\tau^2}{\hat{r}_0^2}, \tag{7.15}$$
with a distribution under the null-hypothesis (i.e. $r_\tau = 0$) which tends to a normal distribution with unit variance and zero mean when $n \to \infty$.

- (Cross-Correlation Test:) Now let us shift gears. Assume that the input timeseries $\{U_t\}_t$ is stochastic as well. If the model were estimated adequate, no dynamics are left in the residuals. Hence, it makes sense to test wether there are cross-correlations left between input signals and residuals. This can be done using the statistic
$$T_n(\{\epsilon_i\}) = \sqrt{n}\hat{\mathbf{r}}^T(\hat{r}_0^2\hat{R}_u)^{-1}\hat{\mathbf{r}}, \tag{7.16}$$
where for given m and τ' one has the sample quantities
$$\begin{cases} \hat{r} = \left(\hat{r}_{\tau'+1}^{u,\epsilon}, \ldots, \hat{r}_{\tau'+m}^{u,\epsilon}\right)^T \\ \hat{r}_\tau^{u,\epsilon} = \frac{1}{n}\sum_{t=1-\min(0,\tau)}^{n-\max(\tau,0)} U_t\epsilon_{t+\tau} \\ \hat{R}_u = \frac{1}{n}\sum_{t=m+1}^{n} U_t' U_t'^T \\ U_t' = \left(U_{t-1}, \ldots, U_{t-m}\right)^T. \end{cases} \tag{7.17}$$
This test statistic has a distribution under the null-hypthesis (i.e. $r^{\epsilon u} = 0$) which tends to a χ^2 distribution with m degrees of freedom when $n \to \infty$.

- (Sign Test) Rather than looking at second moments, it was argued to look at different properties of the residuals. For example one could calculate the number of flips of signs in consequent values. This lead to the statistic

$$T_n(\{\epsilon_i\}) = \frac{1}{\sqrt{n/2}} \left(\sum_{t=1}^{n-1} I(\epsilon_t \epsilon_{t+1} < 0) - \sqrt{n/2} \right), \tag{7.18}$$

  with $I(z)$ equal to one if z is true, and zero otherwise. This statistic has a (sample) distribution under the null-hypthesis (i.e. $\{\epsilon_i\}$ were zero mean white) which tends to a normal distribution with unit variance and zero mean when $n \to \infty$.

Evidently, not all test are equally muscular. When applying a test, there i always a chance the the null-hypothesis were rejected even when it were actually true, or vice versa. The former risk - the so-called type 1 risk) of false positives is captured by the threshold α used in the test. In general, when decreasing this risk factor, one necessarily increases the risk of a false negative. However, this risk is much more difficult to characterize, and requires a proper characterization of the alternative hypothesis.

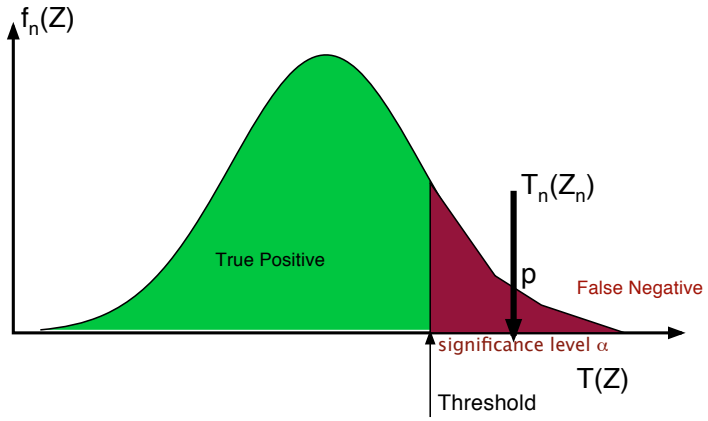

Figure 7.1: Schematic illustration of a statistical hypothesis test. A test is associated with a statistic T_n computed on a sample of size n. One is then after accepting or rejecting a null-hypothesis $\mathbf{H}_0$ underlying possibly the sample. If the sample follows indeed $\mathbf{H}_0$, one can derive theoretically the corresponding null-distribution of T_n. If the statistic computed on the sample is rather atypical under this null-distribution, evidence is found that $\mathbf{H}_0$ is not valid. If the statistic computed on the sample is likely under this null-distribution, no evidence is found to reject $\mathbf{H}_0$. The exact threshold where to draw a distinction between either to conclusions is regulated by a significance level $0 < \alpha < 1$.

7.2 Model Class Selection

Let us study the 'relative' question: 'which of two models is most useful for our needs?' The approach will be to score both models with a given model validation criterion, and to prefer the one with the best score.

The classical way to compare two candidate models are the so-called 'goodness-of-fit hypothesis tests'. Perhaps the most common one is the likelihood ratio test. Here we place ourselves again in a proper stochastic framework. Let $\hat{\theta}_1$ be the maximum likelihood estimator of the parameter θ_0 in the model structure M_1, and let $\hat{\theta}_2$ be the maximum likelihood estimator of the parameter θ_0' in the model structure M_2. We can evaluate the likelihood function $L_{Z_n}(\hat{\theta}_1)$ in the sample Z_n under model M_1 with parameter $\hat{\theta}_1$, as well as the likelihood function $L'_{Z_n}(\hat{\theta}_2)$ in the sample Z_n under model M_2 with parameter $\hat{\theta}_2$. Then we can compute the test statistic

$$T_n(Z_n) = \frac{L_{Z_n}(\hat{\theta}_1)}{L'_{Z_n}(\hat{\theta}_1)}. \tag{7.19}$$

Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two different model structures, such that $\mathcal{M}_1 \subset \mathcal{M}_2$. That is, they are hierarchically structured. For example, both are ARX model structures but the orders of $\mathcal{M}_1$ are lower than the orders of $\mathcal{M}_2$. A related approach based on the loss functions is given makes use of the following test statistic

$$T_n(Z_n) = \frac{V_n^1 - V_n^2}{V_n^2}, \tag{7.20}$$

where V_n^1 is the minimal squared loss of the corresponding LSE in $\mathcal{M}_1$, and V_n^2 is the minimal squared loss of the corresponding LSE in $\mathcal{M}_2$. Lets consider the null-hypothesis $\mathbf{H}_0$ that the model structure $\mathcal{M}_1$ were describing the observations adequately enough. Then it is not too difficult to derive that the sample distribution under the null-hypothesis tends to a χ^2 distribution with degrees of freedom $|\theta_2|_0 - |\theta_1|_0$ (i.e. the difference in number of parameters of either model). This test is closely related to the F-test introduced earlier.

A more pragmatic approach is to use the validation criteria as follows. Consider a collection of estimated models in different model structures. For example, fix the model sort, and constructs estimates of the parameters for various model orders. Then you can score each candidate using an appropriate validation criterion. The model structure (order) leading to the best score is then obviously preferred. This approach is typically taken for model design parameters with no direct physical interpretation.

7.2.1 A Priori Considerations

Model selection is but a formalization of the experience of model building. One could easily imagine that a person which is experienced in the field does not need to implement one of the above methods to guide his design decisions. But in order to assist such an expert, a fe tricks of the trade are indispensable, a few of which are enumerated here.

- Plot the data in different ways.
- Look out for trends or seasonal effects in the data.
- Is an LTI model sufficient for our needs?
- Look at the spectra of the data.
- Try to capture and to explain the noise process. Where does the noise come from in your case study? Consequently, what is

- Try a naively simple model, and try to figure out how you see that this is not sufficient for you. Again, if you manage to formalize the exact goal you're working to, you're halfway the modeling process.