CH5350: Applied Time-Series Analysis

Arun K. Tangirala

Department of Chemical Engineering, IIT Madras

Introduction to Estimation Theory

Motivation

Development of empirical models primarily involves selecting a model structure and forcing it to explain the data. This exercise of fitting is carried out by choosing parameters of the model in an optimal manner.

- The problem of determining optimal parameters essentially belongs to a larger class of problems that "estimate" unknowns (parameters) from knowns (observations), known as the parameter estimation problems.
- Additionally, it is important to know the signal properties prior and during model development. This is the classic problem of estimating signal properties.
- Finally, it is also common determining the underlying signals or states from measurements. Then, we have the state or signal estimation problem.

TSA $\equiv$ Estimation

At the heart of any time-series analysis exercise, is an **estimator**. The role of the estimator is to produce an estimate given information and other user inputs.

Understanding the fundamentals of estimation theory is critical to the development of a good, useful estimate of the model and **importantly** to also be able to state "how good the estimate is."

What does estimation theory offer?

The theory of estimation provides us with

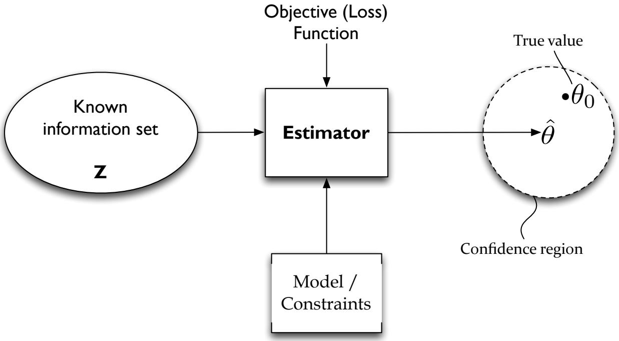
i. **Methods for estimating** the unknowns (model parameters, signals, etc.)

ii. **Means for assessing** the “goodness” of the resulting estimates.

iii. **Making confidence statements** about the true values

Elements of estimation theory

Estimation is the exercise of systematically inferring the unobserved or hidden variable from a **given information set** using a **mathematical map** between the unknowns and knowns, and a **criterion for estimation**.



The device that performs the estimation is said to be the **estimator**

Elements of estimation theory

1. **Information set $\mathbf{Z}$:** This is a key ingredient. An information set is typically the data (time-series, input-output, etc.). It may also contain other a priori information. To obtain good estimates, data should be *informative* w.r.t. the variables / parameters being estimated.
2. **Model (Constraints) $\mathcal{M}$:** The role of the model is to establish a connection between the information set and the space of the unknowns.
   Models can be specified in various forms: (i) differential and/or algebraic equations, (ii) approximation or a predictor functions and (iii) probability density (distribution) functions. Further, it may also include any known constraints (such as bounds) on the variables / parameters to be estimated.

Elements of estimation theory . . . contd.

3. **Obj. function** J **:** Specifies the goals that have to be achieved by the estimator.
    - Typically a minimization of some loss or a risk function or a distance measure. The actual form depends on the estimation problem.
    - Commonly used: Squared approximation errors, negative log-likelihood function, etc.
    - Multiple objectives and/or additional terms that reflect the cost of estimation, computational effort, penalty for violating constraints, *etc.* can also be included.
    - Critical to the form and quality of the final solution, and its implementation!
    - Complicated forms of J may provide better solutions but typically at the cost of increased computational burden.

Elements of estimation theory . . . contd.

4. **Estimator:** Essentially the mathematical device or expression that computes the estimate using the information $\mathbf{Z}$, the model $\mathcal{M}$ and the objective function J.
    - An estimator is also a *filter*. It “filters” the true solution from the given information. **Example:** Wiener and Kalman filters (estimators of signals / states)
    - As in the case of filters, estimators can be of various types - causal / non-causal, linear / non-linear, time-invariant / adaptive and so on.
    - Linear estimators are preferred to non-linear ones because of ease of implementation. However, the price paid may be the inefficiency and / or lack of *robustness*.
    - **Form of the estimator is dictated by the optimization problem.** An *a priori* form along with a criterion of estimation can be employed as well.

Elements of estimation theory . . . contd.

5. **Estimate:** This is the final quantity of interest, that is produced by the estimator. Conventionally, the parameter or the unknown to be estimated is denoted by $\boldsymbol{\theta}$ (could be a scalar or a vector), while the estimate itself is denoted by a hat, $\hat{\boldsymbol{\theta}}$.

$$\hat{\boldsymbol{\theta}} = g(\mathbf{Z})$$

   where the function $g(\mathbf{Z})$ is known as the *estimator function*.

   Needless to state, $g(\mathbf{Z})$ implicitly depends on J and the model $\mathcal{M}$.

   Note: It is also conventional to have the same notation for the estimate and the estimator, *i.e.*, $\hat{\theta}$. The actual reference is understood based on the context.

Simple Example: Constant embedded in noise

Assume that we are interested in knowing the (constant) level $x[k]$ of fluid in a storage tank (no in and out flow).

- The level sensor that is being used for this purpose is known to provide an erroneous measurement $y[k]$. The true quantity of interest is therefore “hidden” or “unobserved” and has to be estimated from $y[k]$.
- Observation at a single instant in fact does offer an estimate of c. But, as we shall show later, this is too crude an estimate.
- Intuitively, using a set of observations $\{y[0], \cdots, y[N-1]\}$ we may obtain a better estimate (under certain conditions).

Simple example: Problem formulation

Given N observations $\{y[k]\}_{k=0}^{N-1}$ of a constant signal c, obtain the "best" estimate of c.

1. **Information set Z**: Observations $\{y[0], y[1], \cdots, y[N-1]\}$
2. **Model / Constraints:** $y[k] = c + e[k]$ where $e[k] \sim GWN(0, \sigma_e^2)$
3. **Criterion of estimation (fit):** Choose standard least squares criterion.

$$\text{minimize} \quad \sum_{k=0}^{N-1} (y[k] - \hat{y}[k])^2$$

where $\hat{y}[k] = c$ is the approximation or prediction of $y[k]$ from the model.

Simple example: Solution

Introduce

$$\varepsilon[k|\theta] = y[k] - \hat{y}[k|\theta] \qquad \text{(prediction error)}$$

$$\mathbf{y} = \begin{bmatrix} y[0] & \cdots & y[N-1] \end{bmatrix}^T; \qquad \hat{\mathbf{y}} = \begin{bmatrix} \hat{y}[0] & \cdots & \hat{y}[N-1] \end{bmatrix}^T$$

Then, the **least-squares estimate** of $\hat{c}$ is the solution to

$$\min_{\theta} \sum_{k=0}^{N-1} \varepsilon^2[k|\theta] = \min_{\theta} ||\mathbf{y} - \hat{\mathbf{y}}||_2^2 \qquad \text{subject to } \hat{y}[k|\theta] = c \qquad (1)$$

Simple example: Solution . . . contd.

Solution:

$$\boxed{\hat{c}^\star = \frac{1}{N}\sum_{k=0}^{N-1} y[k]} \qquad \text{(This is the sample mean!)} \tag{2}$$

The function $\hat{c}(\mathbf{y})$ is said to be the **estimator**, while $\hat{c}$ is the **estimate**.

Simple example: Error characteristics

One more unknown remains to be determined - variance of $e[k]$. The theory for estimating σ_e^2 appears later. For the present, we provide the expression for its estimate

$$\boxed{\hat{\sigma}_e^2 = \frac{1}{N-1}\sum_{k=0}^{N-1} \varepsilon^2[k|\hat{\theta}^\star] = \frac{\text{SSE (or SSR)}}{N-1}} \qquad (3)$$

where $\varepsilon[k|\hat{\theta}^\star]$ is the **residual** evaluated at the optimum. Notice that the RHS of (3) is the sample variance of the prediction errors.

Impact of objective function

Changing the cost criterion has a strong influence on the final solution.

Suppose we choose to minimize the 1-norm instead of the 2-norm, while keeping the predictor fixed. Then, the optimization problem and the solution are

$$\min_{\theta} \sum_{k=0}^{N-1} |\varepsilon[k|\theta]| = \min_{\theta} ||\mathbf{y} - \hat{\mathbf{y}}||_1 \qquad \text{subject to} \;\; \hat{y}[k|\theta] = c \tag{4}$$

Solution:

$$\hat{c}^{\star} = \text{Median}(\mathbf{y}) \tag{5}$$

Median estimator

- The median estimator in (5) is *non-linear*.
- Under the assumption that the errors in $y[k]$ are GWN, both estimators produce identical estimates. However, the "quality" of the estimates are quite different.
- Median offers a robust estimate of c while mean is very sensitive to the presence of outliers.

Simple example: Post-estimation analysis

Regardless of the estimator, **the value of the estimate changes when a different realization of data is used or a record of a different size is used.**

Any estimate is a random variable and a function of the sample size.

Post-estimation: Burning questions

- **Accuracy:** How accurate is the estimate on the average?
- **Variance:** How does the estimate vary across different records? (precision)
- Does the given estimator produce an estimate with the least variability?
- What can we confidently say about the true value of c (call it c_0) from the optimal estimate?
- Will the estimate converge (to c_0) as we increase the sample size?

In order to answer these questions, we need measures for all the above qualifiers.

Notion of “Truth”

All measures for the goodness of estimators **rest on the notion of a “truth”** that essentially serves as a reference or target.

- In all subsequent discussions, we shall assume a “true” value for the unknown and represent it by $\boldsymbol{\theta}_0$.
- This concept is mostly fictitious because the process generating the knowns is usually much more complex than the model. However, the primary reason for assuming a true value is for the sake of evaluating and comparing the performance of estimators.

Properties of Estimators

The metrics used for assessing estimators are of two types:

1. **Statistical properties:** These represent the **ensemble behaviour** of the estimator, i.e., the performance averaged across all possible realizations.
2. **Asymptotic properties:** These characterize the **large sample ($N \to \infty$) behaviour** of the estimator.

Statistical properties of estimators

1. **Bias:** A measure of how accurate the estimator is on the average. Zero bias is useful, but sometimes accuracy is sacrificed for precision.
2. **Variance:** Measures the variability of the estimate across records with reference to $E(\hat{\boldsymbol{\theta}})$. A good estimator should have as low variance as possible.
3. **Efficiency:** Metric to compare the variance of a given estimator with that of a **minimum variance estimator**. An efficient estimator is highly desirable.
4. **Mean square error:** A measure of variability, but w.r.t. the true value $\boldsymbol{\theta}_0$. Ideally a minimum MSE estimator is desirable.
5. **Distribution of estimates:** Provides the uncertainty description of $\hat{\boldsymbol{\theta}}$. Crucial to the determination of the **confidence regions** for the true parameter $\boldsymbol{\theta}_0$.

Asymptotic properties of estimators

The term **asymptotic** refers to large sample behaviour in the limit, i.e., $N \to \infty$.

1. **Asymptotic bias:** Quantifies the statistical bias in the estimate as $N \to \infty$. Biased estimators with zero asymptotic bias are acceptable.
2. **Consistency:** A mandatory property for any estimator, it examines the asymptotic **convergence** of the estimate to the true value. Different forms of consistency arise depending on the notion of convergence (of sequence of RVs) that we work with.
3. **Asymptotic distribution:** Distribution or density of $\hat{\boldsymbol{\theta}}$ as $N \to \infty$. Theoretical results for finite sample size are usually very difficult to compute.

In addition, two indispensable tasks in estimation are **hypothesis testing** and construction of **confidence intervals**

Types of estimation problems

As discussed early in this lecture, estimation problems are of different types:

1. **Signal estimation:** Concerned with estimation of signals from measurements. It is a classical problem with wide applications.
2. **Parameter estimation:** In these problems, the unknowns are model parameters, while the models themselves may be of regression type or probability distribution functions.
3. **State estimation:** These are relatively recent. It is concerned with the inferring the states (of a state-space model) given input-output measurements.

All the three problems can be shown to be equivalent. But observing the distinction is helpful in practice.

Signal estimation

Goal: To estimate the signal(s) from the measurements.

Given measurements $\{\mathbf{Z}[0], \mathbf{Z}[1], \cdots, \mathbf{Z}[N-1]\}$, $\mathbf{Z}[.] \in \mathbb{R}^m$ and a *dynamical* model,

$$\mathbf{Z}[k+1] = \Phi(\mathbf{x}[k], \mathbf{Z}[k], \mathbf{v}[k])$$

estimate the signal $\mathbf{x}[k] \in \mathbb{R}^p$ and the properties of the stochastic signal $\mathbf{v}[k] \in \mathbb{R}^m$.
In addition, the random component $\mathbf{v}[k]$ is assumed to have a pdf:

$$\mathbf{v}[k] \sim f(\mathbf{v}; \xi)$$

where ξ is the vector of parameters characterizing the probability density $f(.)$.

The information set could also include possible input actions $\mathbf{u}[k]$.

Types of signal estimation problems

It is useful to classify the large class of signal estimation problems into three categories based on the times of available information and when we wish to estimate:

1. **Prediction:** Information is available up to k and future values of the signal $x[k+1], x[k+2], \cdots$ are of interest.

$$\text{Estimate } x[k+1], x[k+2], \cdots \qquad \text{given } \{\mathbf{Z}[0], \mathbf{Z}[1], \cdots, \mathbf{Z}[k]\}$$

Types of signal estimation problems ... contd.

2. **Filtering:** An estimation of the signal at the k^{th} (*present*) instant given information up to the k^{th} (*present*) instant.

$$\text{Given } \mathbf{Z}_k = \{\mathbf{Z}[0], \mathbf{Z}[1], \cdots, \mathbf{Z}[k]\} \text{ estimate } x[k]$$

The filtered estimate is denoted by $\hat{x}[k|\mathbf{Z}_k]$ or simply $\hat{x}[k|k]$. Indispensable in most applications (e.g., Wiener filter, Kalman filter).

Types of signal estimation problems . . . contd.

3. **Smoothing:** It relies on both *past* and *future* data to estimate signal at the *present*.

   $$\text{Estimate } x[k] \quad \text{given } \mathbf{Z}_N = \{\mathbf{Z}[0], \cdots, \mathbf{Z}[k-1], \mathbf{Z}[k], \mathbf{Z}[k+1], \cdots, \mathbf{Z}[N-1]\}$$

   The underlying operation is **non-causal** and the resulting estimate is denoted by $\hat{x}[k|\mathbf{Z}_N], \;\; 0 \le k \le N-2$.

Parameter estimation

Goal: Estimate parameters of a regression model or a probability distribution function.

1. **Model parameter estimation:** A standard identification problem. Given the model structure, estimate the parameters (optimally) of the model from the given data.
2. **Estimation of statistical parameters:** For a specified form of the p.d.f. (or the c.d.f.), estimate the parameters (optimally) of the density (distribution) function.

Historically, least squares methods were developed for the former, and the maximum likelihood estimation for the latter. Gradually, both methods have been employed for to address both types of problems.

State estimation

Goal: To estimate **states** from given observations.

Largely popularized by Kalman (1960) in his seminal paper on Kalman filter.

Given measurements $\mathbf{y}[k] \in \mathbb{R}^m$ and input actions $\mathbf{u}[k] \in \mathbb{R}^n$ and a *state-space* model

$$\mathbf{x}[\mathbf{k}+\mathbf{1}] = \Phi(\mathbf{x}[\mathbf{k}], \mathbf{u}[k], \mathbf{w}[k]) \tag{6a}$$

$$\mathbf{y}[k] = \Gamma(\mathbf{x}[k], \mathbf{u}[k], \mathbf{v}[k]) \tag{6b}$$

$$\mathbf{w}[k] \sim f_{\mathbf{w}}(\mathbf{w}; \xi_{\mathbf{w}}) \tag{6c}$$

$$\mathbf{v}[k] \sim f_{\mathbf{v}}(\mathbf{v}; \xi_{\mathbf{v}}) \tag{6d}$$

estimate the signal $\mathbf{x}[k] \in \mathbb{R}^p$ and the statistical properties of the state noise, $\mathbf{w}[k] \in \mathbb{R}^p$ and process noise, $\mathbf{v}[k] \in \mathbb{R}^m$.

Gaussian density functions for the state and process noise with a linear model is widely studied.

Other classifications

i. *Point estimators:* Those that produce single-valued estimates (more common).

ii. *Interval estimators:* Deliver estimates in a certain range or an interval.

iii. *Non-parametric*: The exact form model or density function is unknown, but the space to which it belongs is known. No prejudice towards a class of estimators. Avoids any errors due to misspecification, but larger computational and mathematical burden.

iv. *Semi-parametric*: The predictor form is known but the probability density function is known. These are popular in econometrics.

v. *Parametric*: Both the predictor and the density function forms are known.

Most of the identification literature are parametric estimation problems and we shall only deal with these class of problems in this course.

Estimation methods

1. (Generalized) **Method of moments:** The principle is to bring theoretical moments of the density function close or match with sample moments.
2. **Least-squares (LS) methods:** Principle is to minimize the distance between the observations and the approximations. A historic and natural approach.
3. **Maximum likelihood estimation:** Finds the parameters that maximizes the likelihood of obtaining the given observations
4. **Bayesian estimators:** These methods fuse known or *a priori* information with the given data to **estimate parameters on an interval.** Practical and powerful.

LS methods are extremely popular because of computational ease while MLE methods are widely used for their efficiency. All four classes of estimators are equivalent under certain conditions.

Summary

In this lecture, we obtained an overview of estimation theory, in particular:

- Concept of estimator and elements of an estimation problem - namely, information set, objective function or criterion of fit and model / constraints.
- Properties of an estimator such as bias (accuracy), variance (precision), efficiency, consistency and distribution of estimates
- Types of estimation problems signal, parameter and state estimation
- Classes of estimators - GMM, LS, MLE and Bayesian.