KTH ROYAL INSTITUTE
OF TECHNOLOGY

Estimation theory – Lecture 4

- Ch. 8 Least Squares Estimation
- Ch. 9 Method of Moments

Magnus Jansson

Least Squares (LS)

We assume $x(n) \approx s(n;\theta)$ and form the LS criterion:

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} [x(n) - s(n;\boldsymbol{\theta})]^2$$

The LS estimate is obtained by minimizing $J(\boldsymbol{\theta})$, generally by using nonlinear optimization tools.

LS requires no statistical assumptions. Good or bad estimator? For analysis, we need assumptions! It is reasonable to believe LS is good if the residuals are small compared to the signal, at least on the average.

Linear LS

For linear signal models:

$$\mathbf{s} = \begin{bmatrix} s(0;\boldsymbol{\theta}) \\ \vdots \\ s(N-1;\boldsymbol{\theta}) \end{bmatrix} = \mathbf{H}\boldsymbol{\theta}$$

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} [x(n) - s(n;\boldsymbol{\theta})]^2 = [\mathbf{x} - \mathbf{H}\boldsymbol{\theta}]^T[\mathbf{x} - \mathbf{H}\boldsymbol{\theta}]$$

The minimizing argument is

$$\hat{\boldsymbol{\theta}}_{\text{LS}} = [\mathbf{H}^T\mathbf{H}]^{-1}\mathbf{H}^T\mathbf{x}$$

cf. LMVU and BLUE if residual is white (Gaussian).

Geometrical interpretations

Note that the estimated signal is

$$\hat{\mathbf{s}} = \mathbf{H}[\mathbf{H}^T\mathbf{H}]^{-1}\mathbf{H}^T\mathbf{x} = \mathbf{\Pi}_{\mathbf{H}}\mathbf{x}$$

where $\mathbf{\Pi}_{\mathbf{H}}$ is the orthogonal projection matrix onto the span of $\mathbf{H}$.

Further, note that the residual at the optimal solution is

$$\varepsilon = \mathbf{x} - \hat{\mathbf{s}} = (\mathbf{I} - \mathbf{\Pi}_{\mathbf{H}})\mathbf{x}$$

where $\mathbf{\Pi}_{\mathbf{H}}^{\perp} = (\mathbf{I} - \mathbf{\Pi}_{\mathbf{H}})$ is the orthogonal projection matrix onto the orthogonal complement of span $\mathbf{H}$, or the nullspace of $\mathbf{H}^T$. The minimum of the LS criterion is

$$J(\hat{\boldsymbol{\theta}}_{\text{LS}}) = \mathbf{x}^T\mathbf{\Pi}_{\mathbf{H}}^{\perp}\mathbf{x}$$

Geometrical interpretations cont'd

Observation:

$$\varepsilon \perp \operatorname{span} \mathbf{H}$$

This is the *orthogonality condition*! The residual is orthogonal to the regressors (columns of **H**) at optimality.

Weighted LS

If the quality of the data samples (the different equations) differs, it makes sense to use a weighting:

$$J_{\mathbf{W}}(\boldsymbol{\theta}) = [\mathbf{x} - \mathbf{H}\boldsymbol{\theta}]^T \mathbf{W} [\mathbf{x} - \mathbf{H}\boldsymbol{\theta}]$$

where **W** is a positive definite weighting matrix.

Consider the special case of a diagonal weighting matrix, $\mathbf{W} = \operatorname{diag}(\{w_n\}_{n=0}^{N-1})$, $w_n \geq 0$:

$$J_{\mathbf{W}}(\theta) = \sum_{n=0}^{N-1} w_n [x(n) - s(n;\theta)]^2$$

If $x(k)$, say, is an outlier (very noisy), then use $w_k \approx 0$.
If, on the other hand, $x(k)$ is error free, then use a "large" w_k.

WLS

In general,

$$\hat{\boldsymbol{\theta}}_{\text{WLS}} = [\mathbf{H}^T \mathbf{W} \mathbf{H}]^{-1} \mathbf{H}^T \mathbf{W} \mathbf{x}$$

Cf. BLUE: If the residual $\mathbf{x} - \mathbf{H}\boldsymbol{\theta}$ is zero mean with covariance matrix **C**, it is optimal to use $\mathbf{W} = \mathbf{C}^{-1}$.

Order-recursive LS

Sometimes we want to estimate models of increasing complexity, like e.g., fitting polynomial models of increasing order to data.
In such cases, where models are nested, it is possible to compute models recursively in model order and possibly save some computations. See Chapter 8.6 for details.

Exponentially Weighted Recursive LS

For tracking applications/adaptive filtering it is useful to consider the criterion:

$$J_\lambda^n(\theta) = \sum_{k=0}^{n} \lambda^{n-k}[x(k) - s(k;\theta)]^2$$

where $0 < \lambda \leq 1$ is the *forgetting factor.*

Most recent data at time n are given higher weights relative to those of old data.

Minimizer can be computed as before but this is not efficient.

Recursive solution

Define $\{\mathbf{h}^T(k)\}$ as the rows of $\mathbf{H}$.

1. Initialization: λ, $\hat{\boldsymbol{\theta}}(0) = 0$, $\mathbf{P}(0) = \alpha\mathbf{I}$ with a "large" α
2. For $n = 1, 2, 3 \ldots$, iterate

$$\mathbf{K}(n) = \frac{\mathbf{P}(n-1)\mathbf{h}(n-1)}{\lambda + \mathbf{h}^T(n-1)\mathbf{P}(n-1)\mathbf{h}(n-1)}$$

$$\hat{\boldsymbol{\theta}}(n) = \hat{\boldsymbol{\theta}}(n-1) + \mathbf{K}(n)[x(n-1) - \mathbf{h}^T(n-1)\hat{\boldsymbol{\theta}}(n-1)]$$

$$\mathbf{P}(n) = \frac{1}{\lambda}\left[\mathbf{P}(n-1) - \mathbf{K}(n)\mathbf{h}^T(n-1)\mathbf{P}(n-1)\right]$$

Requires matrix vector multiplications but no matrix inverses.

Separable LS

Sometimes: $\mathbf{s}(\boldsymbol{\theta}) = \mathbf{H}(\boldsymbol{\alpha})\boldsymbol{\beta}$ with $\boldsymbol{\theta} = [\boldsymbol{\alpha}^T \ \boldsymbol{\beta}^T]^T$.

That is, $\boldsymbol{\alpha}$ enters non-linearly, while $\boldsymbol{\beta}$ enters linearly.
LS criterion:

$$J(\boldsymbol{\alpha},\boldsymbol{\beta}) = [\mathbf{x} - \mathbf{H}(\boldsymbol{\alpha})\boldsymbol{\beta}]^T[\mathbf{x} - \mathbf{H}(\boldsymbol{\alpha})\boldsymbol{\beta}]$$

For a given $\boldsymbol{\alpha}$ we can explicitly minimize w.r.t. to $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}) = [\mathbf{H}^T(\boldsymbol{\alpha})\mathbf{H}(\boldsymbol{\alpha})]^{-1}\mathbf{H}^T(\boldsymbol{\alpha})\mathbf{x}$$

Concentrated LS criterion:

$$J(\boldsymbol{\alpha},\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha})) = \mathbf{x}^T[\mathbf{I} - \mathbf{H}(\boldsymbol{\alpha})[\mathbf{H}^T(\boldsymbol{\alpha})\mathbf{H}(\boldsymbol{\alpha})]^{-1}\mathbf{H}^T(\boldsymbol{\alpha})]\mathbf{x}$$

Find $\boldsymbol{\alpha}$ by non-linear optimization (cf. the frequency estimation problem in last lecture).

Connections between LS and ML

Data:

$$x(n) = s(n;\boldsymbol{\theta}) + w(n)$$

for $n = 0, 1, \ldots, N-1$; and $w(n)$ iid $\sim \mathcal{N}(0, \sigma^2)$.

In vector form:

$$\mathbf{x} = \mathbf{s}(\boldsymbol{\theta}) + \mathbf{w} \quad \sim \mathcal{N}(\mathbf{s}(\boldsymbol{\theta}), \sigma^2\mathbf{I}_N)$$

Probability density function:

$$p(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{x} - \mathbf{s}(\boldsymbol{\theta}))^T(\mathbf{x} - \mathbf{s}(\boldsymbol{\theta}))\right\}$$

Connections between LS and ML cont'd

MLE:

$$\hat{\boldsymbol{\theta}}_{\text{ML}}(\mathbf{x}) = \arg\max_{\boldsymbol{\theta}} p(\mathbf{x};\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \ln[p(\mathbf{x};\boldsymbol{\theta})]$$
$$= \arg\min_{\boldsymbol{\theta}} \{-\ln[p(\mathbf{x};\boldsymbol{\theta})]\}$$

or equivalently,

$$\hat{\boldsymbol{\theta}}_{\text{ML}}(\mathbf{x}) = \arg\min_{\boldsymbol{\theta}} \left\{ (\mathbf{x}-\mathbf{s}(\boldsymbol{\theta}))^T (\mathbf{x}-\mathbf{s}(\boldsymbol{\theta})) \right\}$$

which is exactly the LS problem.

Connections between LS and ML cont'd

Colored noise case: $\mathbf{x} = \mathbf{s}(\boldsymbol{\theta}) + \mathbf{w} \quad \sim \mathcal{N}(\mathbf{s}(\boldsymbol{\theta}), \mathbf{C})$

Probability density function

$$p(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{(2\pi)^{N/2}(\det\{\mathbf{C}\})^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}-\mathbf{s}(\boldsymbol{\theta}))^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{s}(\boldsymbol{\theta})) \right\}$$

MLE:

$$\hat{\boldsymbol{\theta}}_{\text{ML}}(\mathbf{x}) = \arg\min_{\boldsymbol{\theta}} \{-\ln[p(\mathbf{x};\boldsymbol{\theta})]\}$$
$$= \arg\min_{\boldsymbol{\theta}} \left\{ (\mathbf{x}-\mathbf{s}(\boldsymbol{\theta}))^T \mathbf{C}^{-1} (\mathbf{x}-\mathbf{s}(\boldsymbol{\theta})) \right\}$$

which is exactly the Weighted LS problem with $\mathbf{W} = \mathbf{C}^{-1}$, the optimal weighting.

MOM – Method of Moments

We observe data $\{x(n)\}_{n=0}^{N-1}$ from some assumed PDF $p(\mathbf{x};\boldsymbol{\theta})$.
Let $\mu_k = \mathrm{E}\{x^k(n)\} = h_k(\boldsymbol{\theta})$.
If $\boldsymbol{\theta} \in \mathbb{R}^p$, consider e.g. the equations:

$$\boldsymbol{\mu} = \mathbf{h}(\boldsymbol{\theta}); \quad \text{where } \boldsymbol{\mu} = [\mu_1 \ \ldots \ \mu_p]^T; \mathbf{h} = [h_1 \ \ldots \ h_p]^T$$

Assume we can solve for $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} = \mathbf{h}^{-1}(\boldsymbol{\mu})$$

Now replace $\boldsymbol{\mu}$ by (consistent) sample estimates:

$$\hat{\mu}_k = \frac{1}{N}\sum_{n=0}^{N-1} x^k(n); \quad k = 1, 2, \ldots, p$$

MOM - cont'd

This results in a natural method of moments estimator:

$$\hat{\boldsymbol{\theta}}_{\text{MOM}} = \mathbf{h}^{-1}(\hat{\boldsymbol{\mu}})$$

- May need more than p equations.
- May use/need other functions of the moments.
- Cf. Yule-Walker methods for estimating parameters of autoregressive models, which are based on equations formed from sample covariance estimates
- Cf. Subspace based estimation methods that are typically based on functions (eigen- or singular value decompositions) of sample covariance matrices

MOM cont'd

- The basic idea is to form equations based on true ensemble averages in terms of unknown parameters and then replace ensemble averages by consistent sample averages.
- Consistency will then typically be inherited by the solution of the equations.
- Variance? Could be very large.

Check Chapter 9.5 for a technique to analyze estimators. (Cf. also the analysis of ML in Appendix 7.B.)

Asymptotically Best Consistent (ABC) Estimation

(Söderström, Stoica “System Identification” Complement C4.4)

Special nonlinear regression model:

$$Y_N = g(\theta_0) + e_N \qquad Y_N, e_N \in \mathbb{R}^M, \theta_0 \in \mathbb{R}^p$$

The noise/residual e_N has the asymptotic covariance matrix

$$\lim_{N\to\infty} N\,\mathrm{E}\{e_N e_N^T\} = R(\theta_0).$$

(This means that Y_N is a root-N consistent estimator of $g(\theta_0)$.)

Objective

Find a consistent estimator

$$\hat{\theta} = f(Y_N)$$

of θ_0 that is ABC in the sense that it has the smallest asymptotic covariance matrix

$$P_M(\hat{\theta}) = \lim_{N\to\infty} N\,\mathrm{E}\{(\hat{\theta}-\theta_0)(\hat{\theta}-\theta_0)^T\}.$$

Derivations

Note that $\lim_{N\to\infty}\hat{\theta} = f(g(\theta_0))$. For consistency we must have $\lim_{N\to\infty}\hat{\theta} = f(g(\theta_0)) = \theta_0$. Since this should hold for any θ_0,

$$\frac{\partial f(g)}{\partial g}\frac{\partial g(\theta)}{\partial \theta} = I$$

or $FG = I$ with obvious definitions of the matrices. A Taylor expansion shows that $P_M(\hat{\theta}) = FRF^T$. One can show that, under the constraint $FG = I$, this is minimized by $F = [G^T R^{-1} G]^{-1} G^T R^{-1}$, and the lower bound on $P_M(\hat{\theta})$ is $[G^T R^{-1} G]^{-1}$.

ABC estimator

The lower bound can be achieved by the estimator

$$\hat{\theta} = \arg\min_{\theta} V(\theta)$$

$$V(\theta) = \frac{1}{2}\left[Y_N - g(\theta)\right]^T R^{-1}(\theta)\left[Y_N - g(\theta)\right]$$

It can be shown that $\hat{\theta}$ is a consistent estimator of θ_0 and

$$P_M(\hat{\theta}) = \left[G^T R^{-1}(\theta_0) G\right]^{-1}$$

where

$$G = \left.\frac{\partial g(\theta)}{\partial \theta}\right|_{\theta=\theta_0} \in \mathbb{R}^{M\times p}$$

Remarks

- One can replace $R(\theta)$ in $V(\theta)$ by any consistent estimate of $R(\theta_0)$ without changing the asymptotic properties of the estimator.
- Similar theory/estimators as ABC exist under different names in literature; notably, the “generalized least squares” (GLS) estimator in statistics/econometrics.
- It can also be viewed as a generalization of optimally weighted MOM.