System Identification, Lecture 2

Kristiaan Pelckmans (IT/UU, 2338)

Course code: 1RT880, Report code: 61800 - Spring 2013
F, FRI Uppsala University, Information Technology

23 January 2013
Lecture 2

• An Example.

• A Model Linear in the Parameter.

• Least Squares Estimation.

• Numerical Techniques.

• Matrix Decompositions.

• Principal Component Analysis.

• Indirect Techniques.
Recipe

• Given a set \( \{x_1, \ldots, x_n\} = \{x_i\}_{i=1}^n \) with \( x_i \in \mathbb{D} \).

• Apply those to a static function \( f_0 : \mathbb{D} \to \mathbb{R} \), and add some disturbances.

• Observe outcomes \( \{y_1, \ldots, y_n\} = \{y_i\}_{i=1}^n \subset \mathbb{R} \) so that

\[
y_i = f_0(x_i) + v_i, \ \forall i = 1, \ldots, n,
\]

with \( \{v_i\} 'small' \).

• We want to recover an as yet unknown parameter \( \theta \) such that \( f_0 \approx f_\theta \).

• ... or that \( f_\theta(x_i) \approx y_i \)

• Theory: converse
• Model class \( \{ f_\theta : \mathcal{D} \to \mathbb{R} \}_\theta \).

• Least Squares (LS) estimator:

\[
\theta_n = \arg\min_\theta \sum_{i=1}^n (y_i - f_\theta(x_i))^2
\]

• Tchebychev Approximation:

\[
\theta_n = \arg\min_\theta \max_{i=1,\ldots,n} |y_i - f_\theta(x_i)|
\]

• L1 Approximation:

\[
\theta_n = \arg\min_\theta \sum_{i=1}^n |y_i - f_\theta(x_i)|
\]

• L0 Approximation (where \(|z|_0 = 1\) iff \(z \neq 0\), and \(|z|_0 = 0\) iff \(z = 0\)):

\[
\theta_n = \arg\min_\theta \sum_{i=1}^n |y_i - f_\theta(x_i)|_0
\]
An Example

• Let \( \{y_1, \ldots, y_n\} = \{y_i\}_{i=1}^n \subset \mathbb{R} \) be a set of observed values. We want to find an as yet unknown parameter \( \theta_0 \in \mathbb{R} \) such that
\[
y_i = \theta_0 + v_i \approx \theta_0, \quad \forall i = 1, \ldots, n.
\]

• Best estimate?
\[
\theta_n = \arg\min_{\theta} V_n(\theta) = \frac{1}{2} \sum_{i=1}^{n} (y_i - \theta)^2
\]
Least Squares Estimate.

• Optimum? Equate the derivative to zero
\[
\frac{dV_n(\theta)}{d\theta} = -\sum_{i=1}^{n} (y_i - \theta) = 0
\]
Hence

\[ \theta_n = \frac{1}{n} \sum_{i=1}^{n} y_i \]

- Theory: is \( \theta_n \approx \theta_0 \)?

- Given observations \( \{(x_i, y_i)\}_{i=1}^{n} \subset \mathbb{R} \times \mathbb{R} \), find the best parameter \( \theta \in \mathbb{R} \) such that

\[ y_i = x_i \theta + v_i \approx x_i \theta, \quad \forall i = 1, \ldots, n. \]

then LS

\[ \theta_n = \arg\min_{\theta} V_n(\theta) = \frac{1}{2} \sum_{i=1}^{n} (y_i - x_i \theta)^2 \]

and equating the derivative to zero gives

\[ \frac{dV_n(\theta)}{d\theta} = \sum_{i=1}^{n} -x_i (y_i - x_i \theta) = 0 \]
and hence

\[ \theta_n = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \]

But...
A Model Linear in the Parameters

This method applicable for many models of such class. Other examples of models which are Linear In the Parameters (LIP)

• Linear model

\[ y_i = \sum_{j=1}^{d} x_{ij} \theta_j + v_i = x_i^T \theta + v_i, \ \forall i = 1, \ldots, n, \]

where \( x_i = (x_{i1}, \ldots, x_{id})^T \in \mathbb{R}^d \) and \( \theta = (\theta_1, \ldots, \theta_d)^T \in \mathbb{R}^d \).

Example ANOVA models.

• Basis functions \( \{ \phi_j : \mathbb{R}^d \to \mathbb{R} \}_{j=1}^{m} \) and

\[ y_i = \sum_{j=1}^{d} \phi_j(x_i) \theta_j + v_i \]

Example Splines, Wavelets, . . . .
• Nonlinear model

\[ y_i = f(x_i) + v_i, \ \forall i = 1, \ldots, n, \]

with unknown \( f : \mathbb{R}^d \rightarrow \mathbb{R} \). Dictionaries of candidate solutions \( \mathcal{F} = \{ f_j : \mathbb{R}^m \rightarrow \mathbb{R} \} \) where \( f \in \mathcal{F} \). Then useful model

\[ y_i = \sum_{j=1}^{m} f_j(x_i)\theta_j + v_i, \ \forall i = 1, \ldots, n. \]

In matrix notation (linear model):

\[ y_i = x_i^T \theta + v_i, \ \forall i = 1, \ldots, n, \]

equalsls

\[
\begin{bmatrix}
  y_1 \\
  \vdots \\
  y_n
\end{bmatrix} =
\begin{bmatrix}
  x_{11} & \cdots & x_{1d} \\
  \vdots & \ddots & \vdots \\
  x_{n1} & \cdots & x_{nd}
\end{bmatrix}
\begin{bmatrix}
  \theta_1 \\
  \vdots \\
  \theta_d
\end{bmatrix} +
\begin{bmatrix}
  v_1 \\
  \vdots \\
  v_n
\end{bmatrix}
\]
denoted as

\[ y = \Phi \theta + v \]
Least Squares Estimation

• Least Squares Objective:

\[ \theta_n = \arg\min_{\theta \in \mathbb{R}^d} V_n(\theta) = \frac{1}{2} (\Phi \theta - y)^T (\Phi \theta - y) \]

• Or

\[ V_n(\theta) = \frac{1}{2} (y^T y - 2(y^T \Phi \theta) + \theta^T (\Phi^T \Phi) \theta) \]

• Solution by equating derivative to zero:

\[ \frac{dV_n(\theta)}{d\theta} = -(\Phi^T y) + (\Phi^T \Phi) \theta = 0 \]

• or solve for \( \theta_n \) (Normal Equations)

\[ (\Phi^T \Phi) \theta_n = \Phi^T y \]
or in vector notation

\[ \sum_{i=1}^{n} x_i (y_i - x_i^T \theta) = 0_d. \]

- If the inverse \((\Phi^T \Phi)^{-1}\) exists.

\[ \theta_n = (\Phi^T \Phi)^{-1} \Phi^T y \]

Figure 1: Orthogonal Projection
Least Squares Estimation, Ct’d

• Suppose 2 inputs exactly the same.

• Suppose an input can be written as a linear combination of the other inputs.

• Suppose inputs ’almost’ equal.

• $m \rightarrow n$.

$\rightarrow \Phi$ contains $d(m)$ linear independent vectors.
Numerical Techniques

Given an invertible matrix $A = A^T \in \mathbb{R}^{m \times m}$ and $b \in \mathbb{R}^m$ in the column space of $A$, find a solution $x \in \mathbb{R}^m$ such that

$$Ax = b$$

- Gauss and Gauss-Jordan elimination.
- Conjugate Gradient Methods.
- Triangular Structure. Try to rephrase as $A'x = b'$ with $A'$ diagonal. Therefore we use the matrix result that any positive definite matrix $A = A^T$ can be written as

$$A = \begin{bmatrix} u_{11} & \cdots & u_{1m} \\ \vdots & & \vdots \\ u_{m1} & \cdots & u_{mm} \end{bmatrix}^T \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1m} \\ 0 & q_{22} & \cdots & q_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & q_{mm} \end{bmatrix}$$
or $A = U^T Q$ with $U^T U = I_n$. Then

$$Ax = b \iff UAx = Ub \iff Qx = Ub$$

and solve by backwards elimination.
Matrix Decompositions

Let $A \in \mathbb{C}^{m \times n}$ be a matrix.

EVD:

- Define an eigenpair $(x, \lambda) \in \mathbb{R}^n \times \mathbb{R}$ as

  $$Ax = \lambda x$$

  and $\|x\|_2 = 1$.

- $n$ different eigenpairs $\{(x_i, \lambda_i)\}_{i=1}^n$

  $$AX = X\Lambda$$

where $X = (x_1, \ldots, x_n) \in \mathbb{C}^{n \times n}$ and

$$\Lambda = \text{diag} \begin{bmatrix} \lambda_1 & 0 \\ \vdots & \ddots \\ 0 & \lambda_n \end{bmatrix} \in \mathbb{R}^{n \times n}$$
• If $A = A^*$, then

(i) All eigenvalues real.
(ii) $\{x_i\}$ orthogonal, or $X^T X = XX^T = I_n$.

• If $A = A^*$, then (Rayleigh coefficient)

$$\lambda_i = \frac{x_i^T A x_i}{x_i^T x_i}$$

Moreover if $\lambda_1 \geq \cdots \geq \lambda_n$

$$\lambda_1 = \max_x \frac{x^T A x}{x^T x}$$

and

$$\lambda_n = \min_x \frac{x^T A x}{x^T x}$$
• Eigen Value Decomposition (EVD) for matrix $A = A^*$ is unique when all eigenvalues are distinct:

$$AU = U\Lambda$$

• Matrix operations, what is $A^{-1}$ when $A = A^T$? Formally,

$$A^{-1} = \sum_{k=1}^{\infty} (I_n - A)^k$$

Let $A = U^T \Lambda U$ then

$$A^{-1} = \sum_{k=1}^{\infty} U^T (I_n - \Lambda)^k U = U^T \text{diag}(\lambda^1, \ldots, \lambda^n) U$$

using the geometric expansion $\sum_{k=0}^{\infty} a^k = \frac{1}{1-a}$ if $|a| < 1$ (Geometric Series).
SVD:

- For any $A \in \mathbb{C}^{m \times n}$, there exist orthonormal matrices $U \in \mathbb{C}^{m \times m}$ and $V \in \mathbb{C}^{n \times n}$ and a 'diagonal' matrix $\Sigma = \mathbb{R}^{m \times n}$ such that

$$A = U \Sigma V^*$$

where $U^* U = UU^* = I_m$ and $VV^* = V^*V = I_n$. The columns of $U$ are the left singular vectors, the columns of $V$ the right singular vectors. The diagonal elements of $\Sigma$ denoted as $\{\sigma_1, \ldots, \sigma_n\}$ the singular values.

- If the matrix $A \in \mathbb{R}^{m \times n}$ is rank $r$, then

$$A = [U_1 \quad U_2] \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \vdots \\ 0 & \cdots & \sigma_r \end{bmatrix} \begin{bmatrix} V_1^* \\ V_2^* \end{bmatrix}$$

- Optimal rank $s \leq r$ approximation:

$$\hat{B} = \arg\min_{B \in \mathbb{R}^{m \times n}} \|A - B\|_F \text{ s.t. rank}(B) = s$$
with \( \|A\|_F = \text{tr} A^T A \) the Frobenius norm, is given by

\[
\hat{B} = \sum_{j=1}^{s} \sigma_j u_i v_j^T = U\Sigma(s)V^T
\]
Principal Component Analysis

Try to find 'hidden structure' in the data.

• Given \( \{x_1, \ldots, x_n\} \subset \mathbb{R}^d \).

• Try to find \( \{v_1, \ldots, v_n\} \subset \mathbb{R}^m \) such that \( v_i \) contains the same 'information' as \( x_i \).

• Optimization problem

\[
\hat{w} = \arg\max_{w \in \mathbb{R}^n} \|w^T \Phi\|_2 \quad \text{s.t.} \quad w^T w = 1
\]

or

\[
\hat{V} = \arg\min_{\{V_j, w_j\}} \left\| X - \sum_{j=1}^{m} V_j w_j \right\|_F.
\]
Figure 2: Examples of Principal Component Analysis.
Indirect Techniques

- Solve normal equations.
- Via SVD.
  \[ \theta_n = (\Phi^T \Phi)^{-1}(\Phi^T y) \]

  or
  \[ (V \Sigma^T U^T U \Sigma V^T)^{-1}(V \Sigma^T U^T y) \]
  \[ = V \Sigma_*^{-2} V^T V \Sigma^T U^T y \]
  \[ = V \Sigma_*^{-T} U^T y \]

- Via Pseudo-inverse.
- Via QR Decomposition
- In MATLAB
  1. \[ >> \theta = \text{inv}(X'X) \times (X'Y) \]
  2. \[ >> \theta = \text{pinv}(X) \times Y \]
  3. \[ >> \theta = X \backslash Y \]
Conclusions

- LS $\rightarrow$ Normal equations!
- Example (LS=average).
- Regression (linear in the parameters) models describe a large class of dynamical models.
- The LS estimator is fundamental in SI and can be derived from various perspectives.
- We have assumed that $\Phi$ is deterministic. We run into problems when this matrix is a function of stochastic variables (ARX).