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 θ such that $f_0 \approx f_\theta$.
- ... or that $f_\theta(x_i) \approx y_i$
- Theory: converse

- Model class $\{f_\theta : \mathbb{D} \rightarrow \mathbb{R}\}_\theta$.
- Least Squares (LS) estimator:

$$\theta_n = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - f_\theta(x_i))^2$$

- Tchebychev Approximation:

$$\theta_n = \underset{\theta}{\operatorname{argmin}} \max_{i=1,\ldots,n} |y_i - f_\theta(x_i)|$$

- L1 Approximation:

$$\theta_n = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} |y_i - f_\theta(x_i)|$$

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

$$\theta_n = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} |y_i - f_\theta(x_i)|_0$$

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
$$y_i = \theta_0 + v_i \approx \theta_0,\ \forall i = 1, \ldots, n.$$

- Best estimate?
$$\theta_n = \operatorname*{argmin}_{\theta} V_n(\theta) = \frac{1}{2} \sum_{i=1}^{n} (y_i - \theta)^2$$

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

$$y_i = x_i\theta + v_i \approx x_i\theta, \ \forall i = 1, \ldots, n.$$

then LS

$$\theta_n = \underset{\theta}{\operatorname{argmin}}\, V_n(\theta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i\theta)^2$$

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

$$y_i = \sum_{j=1}^{d} x_{ij}\theta_j + v_i = \mathbf{x}_i^T\theta + v_i,\ \forall i = 1, \ldots, n,$$

  where $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})^T \in \mathbb{R}^d$ 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

$$y_i = \sum_{j=1}^{d} \phi_j(\mathbf{x}_i)\theta_j + v_i$$

  Example Splines, Wavelets, ...

- Nonlinear model

$$y_i = f(\mathbf{x}_i) + v_i, \ \forall i = 1, \dots, n,$$

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

$$y_i = \sum_{j=1}^{m} f_j(\mathbf{x}_i)\theta_j + v_i, \ \forall i = 1, \dots, n.$$

In matrix notation (linear model):

$$y_i = \mathbf{x}_i^T \theta + v_i, \ \forall i = 1, \dots, n,$$

equals

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1d} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{nd} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_d \end{bmatrix} + \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}$$

denoted as

$$\mathbf{y} = \Phi\theta + \mathbf{v}$$

Least Squares Estimation

- Least Squares Objective:

$$\theta_n = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} V_n(\theta) = \frac{1}{2} \left(\Phi\theta - \mathbf{y}\right)^T \left(\Phi\theta - \mathbf{y}\right)$$

- Or

$$V_n(\theta) = \frac{1}{2} \left(\mathbf{y}^T\mathbf{y} - 2(\mathbf{y}^T\Phi\theta) + \theta^T(\Phi^T\Phi)\theta\right)$$

- Solution by equating derivative to zero:

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

- or solve for θ_n (Normal Equations)

$$(\Phi^T\Phi)\theta_n = \Phi^T\mathbf{y}$$

or in vector notation

$$\sum_{i=1}^{n} \mathbf{x}_i (y_i - \mathbf{x}_i^T \theta) = 0_d.$$

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

$$\theta_n = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y}$$

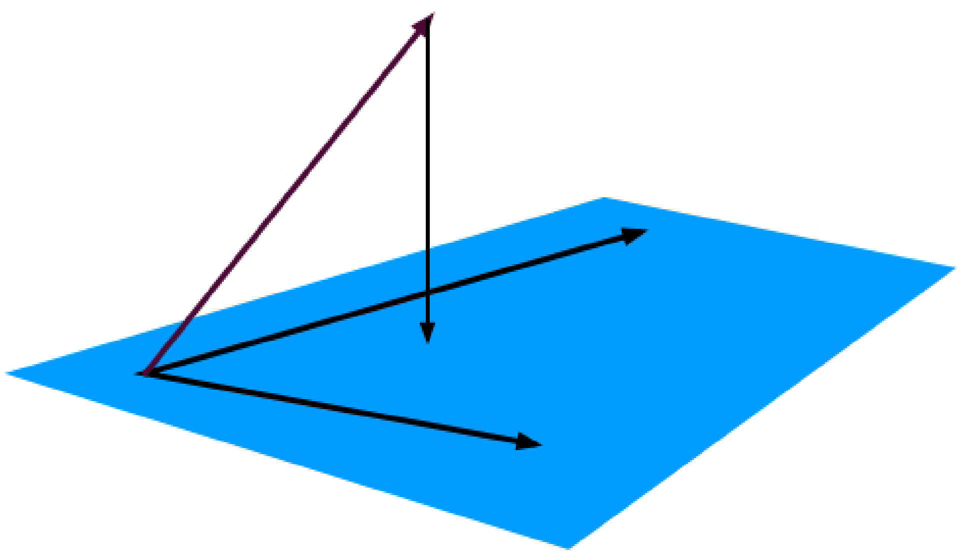

Figure 1: Orthogonal Projection

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.
- $m \to n$.

$\to \Phi$ contains d (m) linear independent vectors.

Numerical Techniques

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

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

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

$$\mathbf{A} = \begin{bmatrix} u_{11} & \dots & u_{1m} \\ \vdots & & \vdots \\ u_{m1} & & u_{mm} \end{bmatrix}^T \begin{bmatrix} q_{11} & q_{12} & \dots & q_{1m} \\ 0 & q_{22} & & q_{2m} \\ \vdots & & \ddots & \\ 0 & \dots & & q_{mm} \end{bmatrix}$$

or $\mathbf{A} = \mathbf{U}^T\mathbf{Q}$ with $\mathbf{U}^T\mathbf{U} = I_n$. Then

$$\mathbf{A}\mathbf{x} = \mathbf{b} \Leftrightarrow \mathbf{U}\mathbf{A}\mathbf{x} = \mathbf{U}\mathbf{b} \Leftrightarrow \mathbf{Q}\mathbf{x} = \mathbf{U}\mathbf{b}$$

and solve by backwards elimination.

Matrix Decompositions

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

EVD:

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

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{x}$$

and $\|\mathbf{x}\|_2 = 1$.

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

$$\mathbf{A}\mathbf{X} = \mathbf{X}\Lambda$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{C}^{n \times n}$ and

$$\Lambda = \operatorname{diag} \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} \in \mathbb{R}^{n \times n}$$

- If $\mathbf{A} = \mathbf{A}^*$, then

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

- If $\mathbf{A} = \mathbf{A}^*$, then (Rayleigh coefficient)

$$\lambda_i = \frac{\mathbf{x}_i^T \mathbf{A} \mathbf{x}_i}{\mathbf{x}_i^T \mathbf{x}_i}$$

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

$$\lambda_1 = \max_{\mathbf{x}} \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}$$

and

$$\lambda_n = \min_{\mathbf{x}} \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}$$

- Eigen Value Decomposition (EVD) for matrix $\mathbf{A} = \mathbf{A}^*$ is unique when all eigenvalues are distinct:

$$\mathbf{A}\mathbf{U} = \mathbf{U}\Lambda$$

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

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

Let $\mathbf{A} = \mathbf{U}^T \Lambda \mathbf{U}$ then

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

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

SVD:

- For any $\mathbf{A} \in \mathbb{C}^{m\times n}$, there exist orthonormal matrices $\mathbf{U} \in \mathbb{C}^{m\times m}$ and $\mathbf{V} \in \mathbb{C}^{n\times n}$ and a 'diagonal' matrix $\Sigma = \mathbb{R}^{m\times n}$ such that
$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^*$$
where $\mathbf{U}^*\mathbf{U} = \mathbf{U}\mathbf{U}^* = I_m$ and $\mathbf{V}\mathbf{V}^* = \mathbf{V}^*\mathbf{V} = I_n$. The columns of $\mathbf{U}$ are the left singular vectors, the columns of $\mathbf{V}$ the right singular vectors. The diagonal elements of Σ denoted as $\{\sigma_1, \ldots, \sigma_n\}$ the singular values.

- If the matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is rank r, then
$$\mathbf{A} = \begin{bmatrix}\mathbf{U}_1 & \mathbf{U}_2\end{bmatrix} \begin{bmatrix}\sigma_1 & & & 0 \\ & \ddots & & \vdots \\ & & \sigma_r & 0 \\ 0 & \ldots & 0 & 0\end{bmatrix} \begin{bmatrix}\mathbf{V}_1^* \\ \mathbf{V}_2^*\end{bmatrix}$$

- Optimal rank $s \leq r$ approximation:
$$\hat{\mathbf{B}} = \underset{\mathbf{B}\in\mathbb{R}^{m\times n}}{\operatorname{argmin}} \|\mathbf{A} - \mathbf{B}\|_F \quad \text{s.t.} \quad \operatorname{rank}(\mathbf{B}) = s$$

with $\|\mathbf{A}\|_F = \operatorname{tr} \mathbf{A}^T \mathbf{A}$ the Frobenius norm, is given by

$$\hat{\mathbf{B}} = \sum_{j=1}^{s} \sigma_j \mathbf{u}_i \mathbf{v}_j^T = \mathbf{U} \Sigma_{(s)} \mathbf{V}^T$$

Principal Component Analysis

Try to find 'hidden structure' in the data.

- Given $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$.

- Try to find $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subset \mathbb{R}^m$ such that $\mathbf{v}_i$ contains the same 'information' as $\mathbf{x}_i$.

- Optimization problem

$$\mathbf{w} = \underset{\mathbf{w} \in \mathbb{R}^n}{\operatorname{argmax}} \|\mathbf{w}^T \Phi\|_2 \text{ s.t. } \mathbf{w}^T \mathbf{w} = 1$$

or

$$\hat{\mathbf{V}} = \underset{\{\mathbf{V}_j, \mathbf{w}_j\}}{\operatorname{argmin}} \left\| \mathbf{X} - \sum_{j=1}^{m} \mathbf{V}_j \mathbf{w}_j \right\|_F .$$

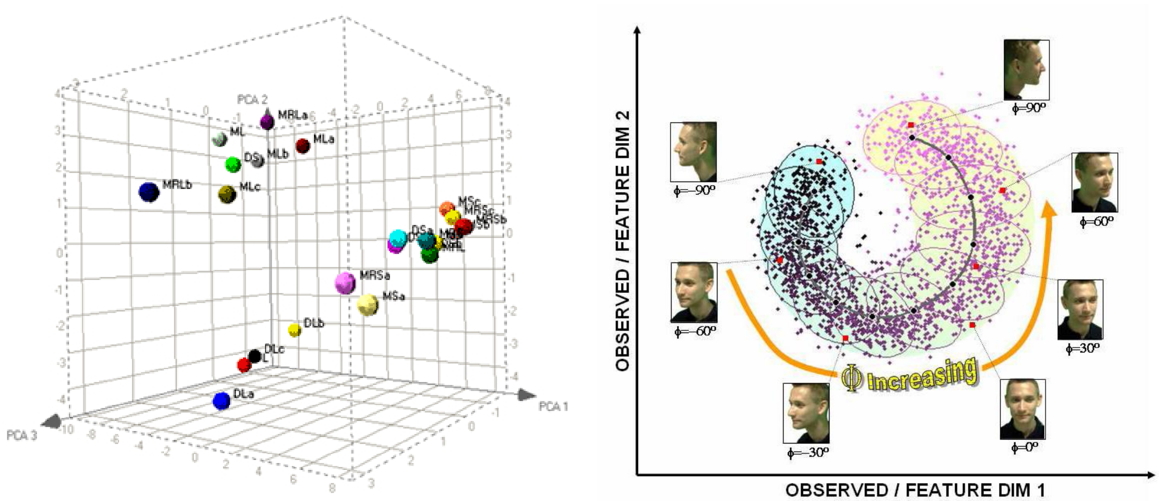

Figure 2: Examples of Principal Component Analysis.

Indirect Techniques

- Solve normal equations.
- Via SVD.

$$\theta_n = (\Phi^T\Phi)^{-1}(\Phi^T \mathbf{y})$$

or

$$\begin{aligned}(\mathbf{V}\Sigma^T\mathbf{U}^T\mathbf{U}\Sigma\mathbf{V}^T)^{-1}(\mathbf{V}\Sigma^T\mathbf{U}^T\mathbf{y})\\ = \mathbf{V}\Sigma_*^{-2}\mathbf{V}^T\mathbf{V}\Sigma^T\mathbf{U}^T\mathbf{y}\\ = \mathbf{V}\Sigma_*^{-T}\mathbf{U}^T\mathbf{y}\end{aligned}$$

- Via Pseudo-inverse.
- Via QR Decomposition
- In MATLAB

1. `>> theta = inv(X'*X) * (X'*Y)`
2. `>> theta =pinv(X) * Y`
3. `>> theta = X \ 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 Φ is deterministic. We run into problems when this matrix is a function of stochastic variables (ARX).