System Identification, Lecture 4

Kristiaan Pelckmans (IT/UU, 2338)

Course code: 1RT880, Report code: 61800 - Spring 2011
F, FRI Uppsala University, Information Technology

1 Februari 2011

Lecture 4

- Stochastic Setup.
- Interpretation.
- Maximum Likelihood.
- Least Squares Revisited.
- Instrumental Variables.

Stochastic Setup

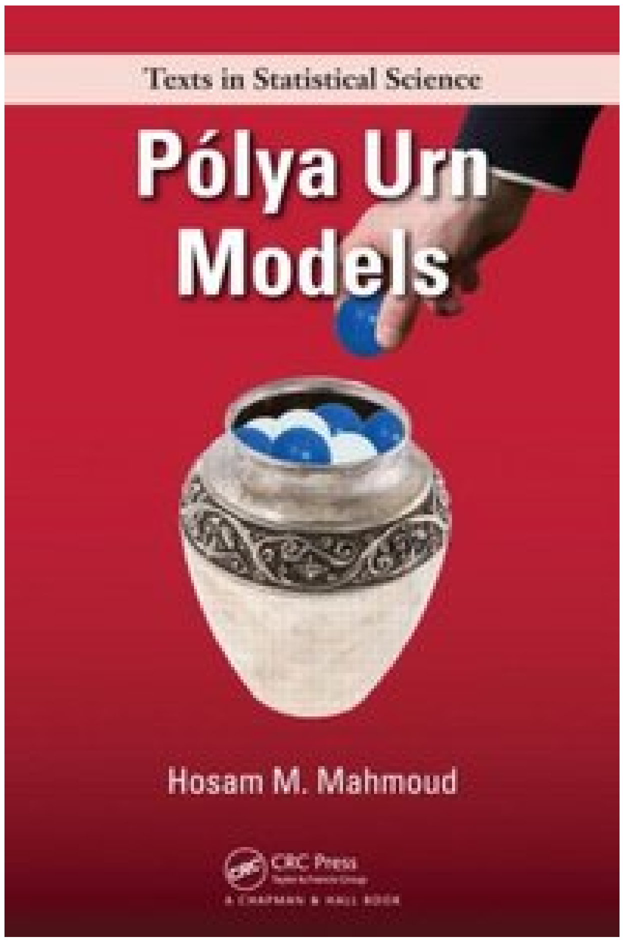
Why:

- Analysis (abstract away).
- Constructive.

Basics:

- Samples ω
- Sample space $\Omega = \{\omega\}$
- Event $A \subset \Omega$.

- Rules of probability $P : \{\Omega\} \to [0,1]$
  1. $P(\Omega) = 1$,
  2. $P(\{\}) = 0$,
  3. $P(A_1 \cup A_2) = P(A_1) + P(A_2)$ for $A_1 \cap A_2 = \{\}$



Random variables:

- Function $X : \Omega \to \mathbb{R}$.
- $P\left(\{\omega : X(\omega) = x\}\right) := P(X = x)$.
- CDF $F(x) := P(X \leq x)$.
- PDF $f(x) := \frac{dF(x)}{dx}$.
- Realizations vs. random variables.
- Expectation:

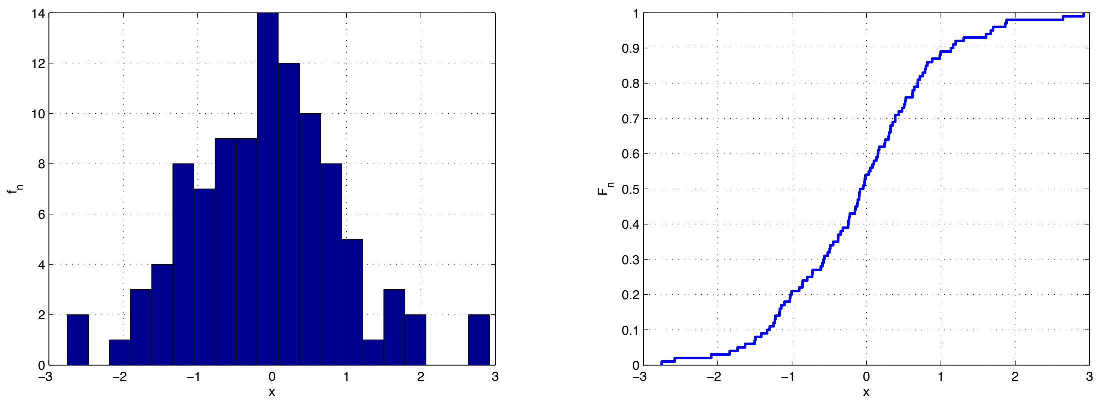
$$\mathbb{E}[X] := \int x \; dF(x) = \int x \; f(x) \; dx$$

- Expectation of $g : \mathbb{R} \to \mathbb{R}$.

$$\mathbb{E}[g(X)] = \int g(x) \; dF(x) = \int g(x) \; f(x) \; dx$$

- Independence:

$$\mathbb{E}[X_1 X_2] = \mathbb{E}[X_1]\ \mathbb{E}[X_2]$$



Examples:

- Urn. Sample= Ball. Random variable = Color ball.
- Sample = Image. Random variable = Color/black-white.
- Sample = Speech. Random variable = Words.
- Sample = text. Random variable = Length optimal compression.

- Sample = weather. Random variable = Temperature.

- Sample = External force. Random variable = Influence.

PDF:

- Gaussian PDF determined by 2 moment: mean μ and variance σ^2

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right)$$

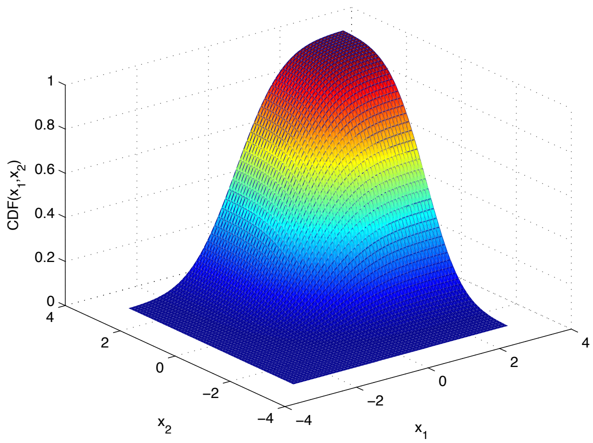
(or standard deviation σ).

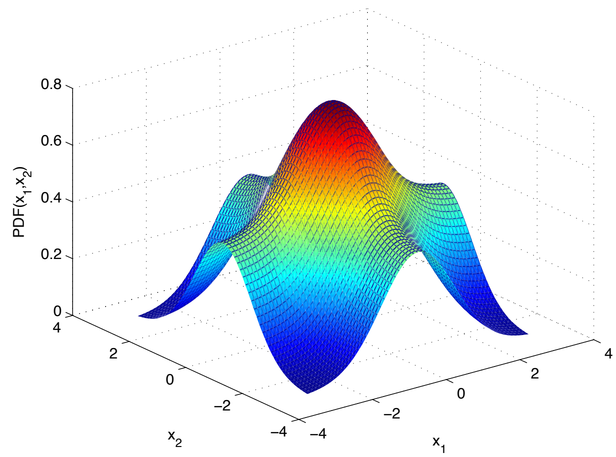
- Central Limit Theorem.

- Closed by convolution, conditioning and product.

- Multivariate Normal PDF for $\mathbf{x} \in \mathbb{R}^2$:

$$f(\mathbf{x}) = \mathcal{N}(\mu, \Sigma) = \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)\right)$$


CDF(x1,x2)
1
0.8
0.6
0.4
0.2
0
4
2
0
−2
−4
x2
−4
−2
0
2
4
x1


PDF(x1,x2)
0.8
0.6
0.4
0.2
0
4
2
0
−2
−4
x2
−4
−2
0
2
4
x1

- Let bivariate Gaussian with mean μ and variance matrix Σ as

$$f\left(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$$

Then conditional bivariate Gaussian $f(X_1 | X_2 = x_2) = \mathcal{N}(\bar{\mu}, \bar{\Sigma})$ where

$$\begin{cases} \bar{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \bar{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{cases}$$

Stochastic Processes

- Sequence of random variables indexed by time.
- Joint distribution function.
- Conditional PDF.
- Stationarity.
- Quasi-Stationary.
- Ergodic: different realizations vs. time averages.

Maximum Likelihood

- If the *true* PDF of X were f, then the probability of occurence of a realisation x of X were $f(x)$.
- Consider a class of such PDFs $\{f_\theta : \theta \in \Theta\}$
- Likelihood function: this PDF as a function of an unknown parameter θ
$$L(\theta; x) = f_\theta(x)$$
- Idea: given an observation x_n, find a model under which this sample was most likely to occur.
$$\theta_n = \operatorname*{argmax}_{\theta \in \Theta} L(\theta; x_n)$$
- Equivalent to
$$\theta_n = \operatorname*{argmax}_{\theta \in \Theta} \log L(\theta; x_n)$$

Properties of θ_n

- Assume that x_n contains n independent samples.
- Consistent, i.e. $\theta_n \to \theta_0$ with rate $\sqrt{1/n}$.
- Asymptotic normal: if n large

$$f\left(\sqrt{n}(\theta_n - \theta_0)\right) = \mathcal{N}(0,1)$$

- Sufficient.
- Efficient.
- Regularity conditions: identifiability:

$$\exists x : L(x,\theta) \neq L(x,\theta') \Leftrightarrow \theta \neq \theta'$$

Interpretations

"The metallurgist told his friend the statistician how he planned to test the effect of heat on the strength of a metal bar by sawing the bar into six pieces. The first two would go into the hot oven, the next two into the medium oven and the last two into the cool oven. The statistician, horrified, explained how he should randomize in order to avoid the effect of a possible gradient of strength in the metal bar. The method of randomization was applied, and it turned out that the randomized experiment called for putting the first two into the hot oven, the next two into the medium oven and the last two into the cool oven. "Obviously, we can't do that," said the metallurgist. "On the contrary, you have to do that," said the statistician."

Least Squares Revisited

- Observations of $\{y_i\}_{i=1}^n$.
- Model as realizations of random variable Y_i with

$$Y_i = \mathbf{x}_i^T \theta + V_i,\ V_i \sim \mathcal{N}(0,1)$$

with $\{e_i\}$ realizations of $\{V_i\}_i$ *i.i.d.* .

- $\{\mathbf{x}_i\}_{i=1}^n$ and θ deterministic.
- Equivalently

$$f_i(y) = \mathcal{N}(y; \mathbf{x}_i^T \theta, \sigma)$$

- Since $\{V_i\}$ indendent

$$f(y_1, \ldots, y_n) = \prod_{i=1}^n \mathcal{N}(y_i; \mathbf{x}_i^T \theta, \sigma)$$

- Maximum Likelihood

$$\theta_n = \underset{\theta}{\operatorname{argmax}} \log L(\theta, y_1, \ldots, y_n) = \log \prod_{i=1}^{n} \mathcal{N}(y; \mathbf{x}_i^T \theta, \sigma)$$

- $=$ Ordinary Least Squares (OLS).

$$\theta_n = \underset{\theta}{\operatorname{argmin}}\, c \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T \theta\right)^2$$

- Also when zero mean, uncorrelated errors with equal, bounded variance (Gauss-Markov).

- If noise not independent, but

$$\begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} \sim \mathcal{N}\left(0_n, \Sigma\right)$$

  Then ML leads to BLUE

$$\theta_n = \underset{\theta}{\operatorname{argmin}}\, \mathbf{e}^T \Sigma^{-1} \mathbf{e} = (\mathbf{y} - X\theta)^T \Sigma^{-1} (\mathbf{y} - X\theta)$$

Analysis OLS

Model with $\{e_i\}$ zero mean white noise with $\mathbb{E}[V_i^2] = \lambda^2$ and $\{\theta, \mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$

$$y_i = \mathbf{x}_i \theta_0 + e_i.$$

OLS: $\theta_n = \mathrm{argmin}_\theta \left\| \mathbf{y} - \mathbf{X}\theta \right\|_2^2.$

Normal Equations: $(\mathbf{X}^T\mathbf{X})\theta = \mathbf{X}^T\mathbf{y}.$

- Unbiased:

$$\mathbb{E}[\theta_n] = \theta_0.$$

- Covariance:

$$\mathbb{E}[(\theta_n - \theta_0)^T(\theta_n - \theta_0)] = \lambda^2(\mathbf{X}^T\mathbf{X})^{-1} = \frac{\lambda^2}{n}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T\right)^{-1}.$$

- Objective:

$$\mathbb{E}\|\mathbf{y} - \mathbf{X}\theta_n\|_2^2 = \lambda^2(n-d).$$

- Efficient. $\sqrt{n}(\theta_n - \theta_0) \to \mathcal{N}(0, I^{-1})$ for $n \to \infty$.

Cramér-Rao Bound

- Lowerbound to the variance of any unbiased estimator θ_n to θ.

- Discrete setting: D is set of samples, with $\{D_n\}$ denoting the observed datasets.

- Given PDFs $f_\theta > 0$ over all possible datasets D_n such that

$$\sum_{D_n} f_\theta(D_n) = 1$$

- Given estimator $\theta_n(D_n)$ of $\theta \in \Theta$ such that $\forall \theta \in \Theta$

$$\mathbb{E}[\theta_n(D_n)] = \sum_{D_n} \theta_n(D_n) f_\theta(D_n) = \theta$$

- Taking derivative w.r.t. θ gives

$$\begin{cases} \sum_{D_n} \frac{df_\theta(D_n)}{d\theta} = 0 \\ \sum_{D_n} \theta_n(D_n) \frac{df_\theta(D_n)}{d\theta} = 1 \end{cases}$$

and hence

$$1 = \sum_{D_n} (\theta - \theta_n(D_n)) \frac{df_\theta(D_n)}{d\theta}$$

- Combining:

$$1 = \sum_{D_n} (\theta - \theta_n(D_n)) \left(\frac{f_\theta(D_n)}{f_\theta(D_n)} \right) \left(\frac{df_\theta(D_n)}{d\theta} \right)$$

$$1 = \sum_{D_n} (\theta - \theta_n(D_n)) \sqrt{f_\theta(D_n)} \ \left(\frac{\sqrt{f_\theta(D_n)}}{f_\theta(D_n)} \frac{df_\theta(D_n)}{d\theta} \right)$$

- Cauchy-Schwarz ($\mathbf{a}^T \mathbf{b} \leq \|\mathbf{a}\|_2 \|\mathbf{b}\|_2$) gives

$$1 \leq \sum_{D_n} (\theta - \theta_n(D_n))^2 f_\theta(D_n) \sum_{D_n} \left(\frac{\sqrt{f_\theta(D_n)}}{f_\theta(D_n)} \frac{df_\theta(D_n)}{d\theta} \right)^2$$

Or

$$\boxed{\mathbb{E}_\theta[\theta - \theta_n(D_n)]^2 \geq \frac{1}{I(\theta)}}$$

with

$$I(\theta) = \mathbb{E}_\theta \left[\frac{df_\theta(D_n)}{d\theta} \frac{1}{f_\theta(D_n)} \right]^2$$

Dynamical System

Given the ARX(1,1) system

$$y_t + a y_{t-1} = b u_{t-1} + e_t, \ \forall t.$$

where $\{e_t\}$ zero mean and unit variance white noise. Then application of OLS gives

- Normal equations of $\theta = (-a, b)^T$

$$\left(\begin{bmatrix} \sum_{t=1}^n y_{t-1}^2 & \sum_{t=1}^n y_{t-1} u_{t-1} \\ \sum_{t=1}^n u_{t-1} y_{t-1} & \sum_{t=1}^n u_{t-1}^2 \end{bmatrix} \right) \theta = \begin{bmatrix} \sum_{t=1}^n y_{t-1} y_t \\ \sum_{t=1}^n u_{t-1} y_t \end{bmatrix}.$$

- Taking Expectations

$$\mathbb{E} \begin{bmatrix} \sum_{t=1}^n y_{t-1}^2 & \sum_{t=1}^n y_{t-1} u_{t-1} \\ \sum_{t=1}^n u_{t-1} y_{t-1} & \sum_{t=1}^n u_{t-1}^2 \end{bmatrix} \theta = \mathbb{E} \begin{bmatrix} \sum_{t=1}^n y_{t-1} y_t \\ \sum_{t=1}^n u_{t-1} y_t \end{bmatrix}.$$

- Approximatively

$$\theta \approx \begin{bmatrix} \mathbb{E}[y_t^2] & \mathbb{E}[y_t u_t] \\ \mathbb{E}[u_t y_t] & \mathbb{E}[u_t^2] \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{E}[y_t y_{t-1}] \\ \mathbb{E}[y_t u_{t-1}] \end{bmatrix} = \mathbf{R}^{-1}\mathbf{r}.$$

- Ill-conditioning.

Instrumental Variables

- LS: $\min \sum_{i=1}^n (y_i - \mathbf{x}_i^T\theta)^2$
- Normal equations

$$\left(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T\right)\theta = \left(\sum_{i=1}^n \mathbf{x}_i y_i\right)$$

- Or

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

- Interpretation as orthogonal projection.
- Idea:

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

- Choose $\{Z_t\}$ taking values in $\mathbb{R}^d$ such that
  - $\{Z_t\}$ independent from $\{V_t\}$
  - $\mathbf{R}$ full rank.

- Example. Past input signals.

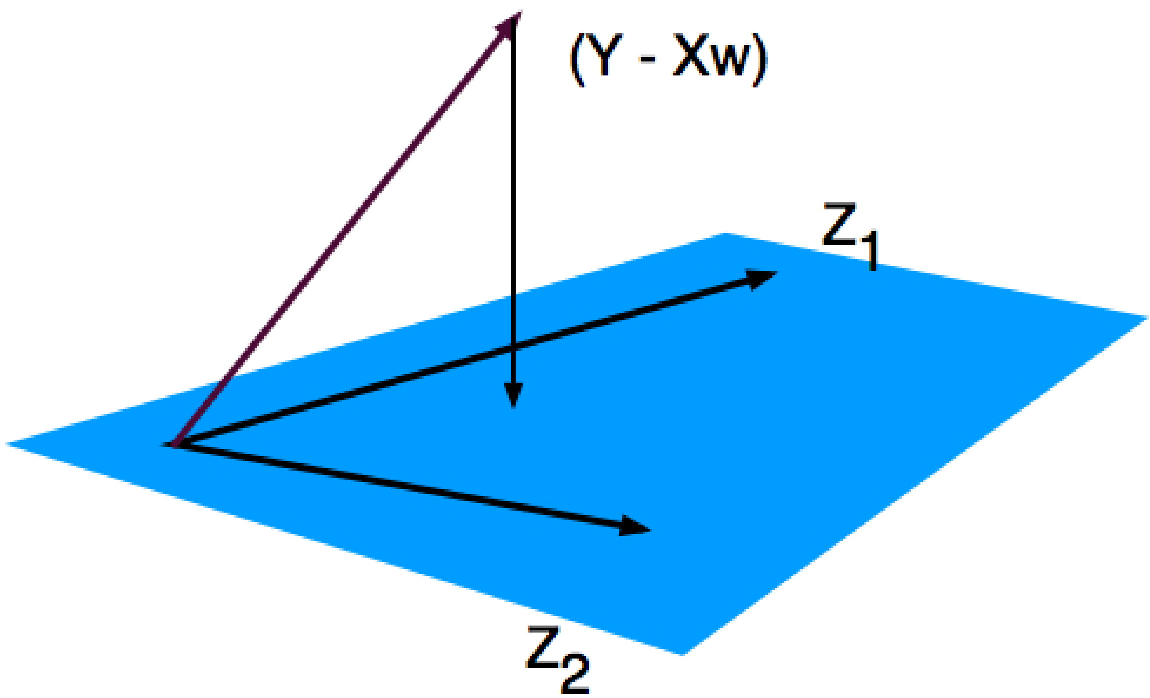


Figure 1: Instrumental Variable as Modified Projection.

Conclusions

- Stochastic (Theory) and Statistical (Data).
- Maximum Likelihood.
- Least Squares.
- Instrumental Variables.