Advanced Econometric Methods EMET3011/8014

Lecture 7

John Stachurski

Semester 1, 2011

Announcements/Reminders

- Please get yourself fresh copy of the course notes PDF

Today's Lecture

- Empirical Risk Minimization
- ERM and LSQ
- Inference
- Start of Linear Algebra

Empirical Risk Minimization

An inductive principle

- Very general
- Includes important techniques as special cases

Example: The Regression Problem

We observe input x to a system, followed by output y

Aim: Predict output values from new input values

Strategy: choose f such that $f(x)$ a “good” prediction of y

Measuring “goodness”:

$$\text{Loss } L(y, f(x)) \text{ incurred on predicting } y \text{ with } f(x)$$

Function L is called a **loss function**

Common choices for the loss function include:

- quadratic loss: $L(y, f(x)) = (y - f(x))^2$
- absolute loss: $L(y, f(x)) = |y - f(x)|$
- discrete loss: $L(y, f(x)) = \mathbb{1}\{y \neq f(x)\}$

Quadratic loss popular for regression ($y \in \mathbb{R}$)

Discrete loss popular for classification (y discrete)

Let L be any loss function

Consider choosing f to minimize expected loss

$$R(f) := \mathbb{E}\left[L(y, f(x))\right] =: \int\int L(t, f(s))p(s,t)dsdt$$

- p is the joint density of (x, y)
- Expected loss given f is called the **risk** of f
- R is called the **risk function**

Example: If $L =$ quadratic loss, risk minimizer is $f^*(x) := \mathbb{E}\left[y \,|\, x\right]$

Proof: We'll do it later

Statistician's problem: Can't minimize

$$R(f) := \mathbb{E}\left[L(y, f(x))\right] =: \int\int L(t, f(s)) p(s,t) ds dt$$

because we don't know distribution p of (x, y)

However, we do observe pairs $(x_1, y_1), \ldots, (x_N, y_N) \overset{\text{IID}}{\sim} p$

The **empirical risk function** defined as

$$\hat{R}(f) := \frac{1}{N} \sum_{n=1}^{N} L(y_n, f(x_n))$$

When N large we have

$$R(f) := \mathbb{E}\left[L(y, f(x))\right] \approx \frac{1}{N} \sum_{n=1}^{N} L(y_n, f(x_n)) =: \hat{R}(f)$$

Empirical risk minimization: Inductive principle that attempts to minimize risk by minimizing the empirical risk

Under ERM principle, choose $\hat{f}$ by solving

$$\hat{f} := \operatorname*{argmin}_{f \in \mathcal{F}} \hat{R}(f)$$

$$=: \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{N} \sum_{n=1}^{N} L(y_n, f(x_n)) = \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{n=1}^{N} L(y_n, f(x_n))$$

Set of functions $\mathcal{F}$ is called the **hypothesis space**

- A class of candidate functions chosen by econometrician

Taking $\mathcal{F} =$ all $f : \mathbb{R} \to \mathbb{R}$ usually a bad idea—see below

Example: ERM and Linear Least Squares

Specializing ERM to quadratic loss gives **least squares problem**

$$\min_{f \in \mathcal{F}} \sum_{n=1}^{N} (y_n - f(x_n))^2$$

If, in addition, $\mathcal{F}$ is the set of affine functions

$$\mathcal{L} := \{ \text{ all functions of the form } \ell(x) = \alpha + \beta x \}$$

then we have **linear least squares problem**

$$\min_{\ell \in \mathcal{L}} \sum_{n=1}^{N} (y_n - \ell(x_n))^2 = \min_{\alpha, \beta} \sum_{n=1}^{N} (y_n - \alpha - \beta x_n)^2$$

Simple manipulations show minimizers are

$$\hat{\beta} = \frac{\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})}{\sum_{n=1}^{N}(x_n - \bar{x})^2} \quad \text{and} \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Linear least squares is empirical risk counterpart of best linear predictor problem

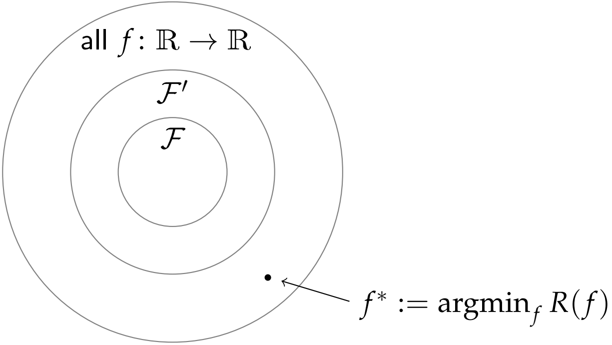
$$\min_{\ell \in \mathcal{L}} R(\ell) = \min_{\alpha, \beta \in \mathbb{R}} R(\alpha, \beta) = \min_{\alpha, \beta \in \mathbb{R}} \mathbb{E}\left[(y - \alpha - \beta x)^2\right]$$

Recall that solutions are

$$\beta^* := \frac{\mathrm{cov}[x, y]}{\mathrm{var}[x]} \quad \text{and} \quad \alpha^* := \mathbb{E}[y] - \beta^* \mathbb{E}[x]$$

Choosing the Hypothesis Space

Why minimize empirical risk over restricted space $\mathcal{F}$?



Bigger $\mathcal{F}$ means smaller empirical risk—is this not good?

Not necessarily: We actually want small risk

Large $\mathcal{F}$ means

- $\hat{R}(\hat{f})$ is small
- $R(\hat{f})$ may or may not be small

We are trying to minimize risk based on a sample, rather than the actual distribution

Don't want to read "too much" into this particular sample

Example: Minimizing empirical risk over progressively larger $\mathcal{F}$

Suppose system is defined by

$$x \sim U[-1,1] \text{ and then } y = \cos(\pi x) + u \text{ where } u \sim N(0,1)$$

Implies a joint density p for (x, y)

For $L =$ quadratic, the risk is then

$$R(f) = \mathbb{E}\left[(y - f(x))^2\right] = \int \int (t - f(s))^2 p(s,t) ds dt \tag{1}$$

The minimizer of the risk is $f^*(x) := \cos(\pi x)$

We generate $N = 25$ data points (x_n, y_n) from the model (i.e., p)

ERM problem is

$$\min_{f \in \mathcal{P}_d} \hat{R}(f)$$

where

- $\hat{R}(f) = \frac{1}{N} \sum_{n=1}^{N} (y_n - f(x_n))^2$
- $\mathcal{P}_d :=$ all polynomials of degree d

Note: $f \in \mathcal{P}_d$ means $f(x) = c_0 + c_1 x^1 + \cdots c_d x^d$ with $c_i \in \mathbb{R}$

Note that the hypothesis spaces are increasing in d

In particular, $\mathcal{P}_d \subset \mathcal{P}_{d+1}$ for all d

Proof for $d = 1$:

$$\mathcal{P}_1 \ni f(x) := c_0 + c_1 x = c_0 + c_1 x + 0x^2 =: f(x) \in \mathcal{P}_2$$

$$\therefore \quad \mathcal{L} = \mathcal{P}_1 \subset \mathcal{P}_2 \subset \mathcal{P}_3 \subset \cdots$$

Let $\hat{f}_d := \operatorname{argmin}_{f \in \mathcal{P}_d} \hat{R}(f)$

What happens as we increase d?

- We know that $\hat{R}(\hat{f}_d)$ is (at least weakly) decreasing in d
- How about the risk $R(\hat{f}_d)$?

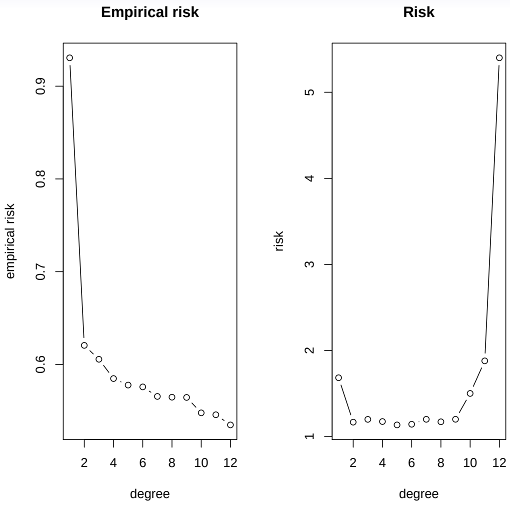


Figure: Risk and empirical risk as a function of d

Empirical risk falls monotonically with d

But risk decreases slightly and then increases rapidly

- Small d: high empirical risk and high risk
- Medium d: risk is minimized
- Large d: small empirical risk and high risk

High risk means large expected loss

Let's plot functions to learn more

In the plots:

- The N data points are plotted as circles
- Risk minimizer $f^*(x) = \cos(\pi x)$ plotted in black
- Estimate $\hat{f}_d$ plotted in red

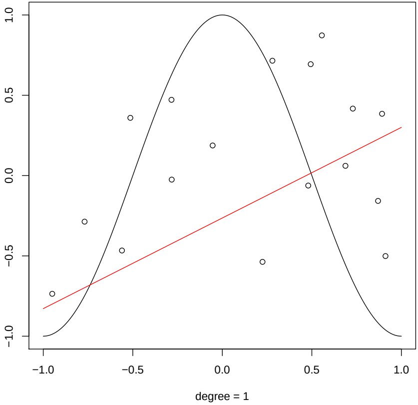


Figure: Fitted polynomial, $d = 1$

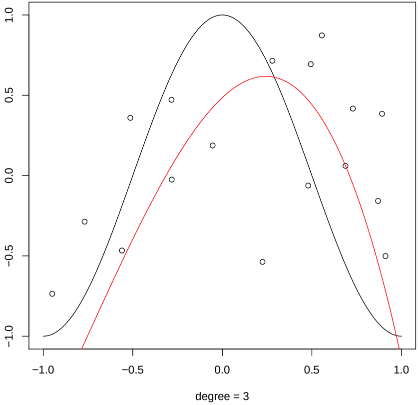


Figure: Fitted polynomial, $d = 3$

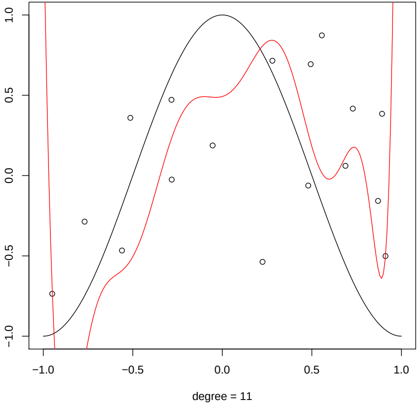


Figure: Fitted polynomial, $d = 11$

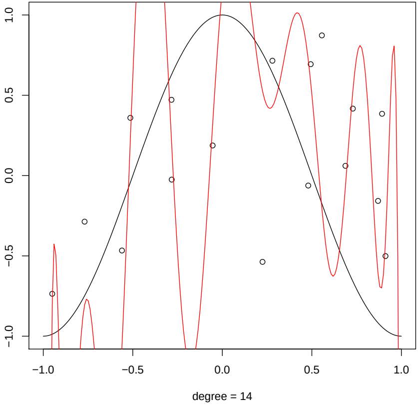


Figure: Fitted polynomial, $d = 14$

Summary

Choice of the hypothesis space $\mathcal{F}$ is crucial

$\mathcal{F}$ too big means "overfitting" data—high risk

In real statistical applications, can't see black line (true model)

Can't use this information to choose $\mathcal{F}$

Many people choose $\mathcal{F} = \mathcal{L}$, but may not be good choice

Ideally, should choose $\mathcal{F}$ on the basis of economic theory

Message: Statistical learning equals prior knowledge plus data

Other Applications of ERM

Many techniques can be recovered as special cases of ERM...

Example: Sample mean as estimator of mean

Want to predict x given IID sample $x_1, \ldots, x_N$

Letting θ be our prediction, risk is

$$R(\theta) = \mathbb{E}\left[L(x, \theta)\right]$$

Empirical risk is

$$\hat{R}(\theta) = \frac{1}{N} \sum_{n=1}^{N} L(x_n, \theta)$$

Exercise: If L is quadratic, then

- minimizer of risk is mean
- minimizer of empirical risk is sample mean

The ERM principle is essentially nonparametric in nature

Empirical risk determined by

1. The loss function
2. The empirical distribution

But some parametric techniques can be recovered as special case

Example: Maximum likelihood from ERM

- Data $x_1, \ldots, x_N \overset{\text{IID}}{\sim} p(\cdot;\theta)$
- Loss function $L(\theta, x) := -\ln p(x;\theta)$

Applying ERM,

$$\begin{aligned}
\hat{\theta} &= \underset{\theta}{\operatorname{argmin}} \sum_{n=1}^{N} L(\theta, x_n) \\
&= \underset{\theta}{\operatorname{argmin}} \left\{ -\sum_{n=1}^{N} \ln p(x_n;\theta) \right\} \\
&= \underset{\theta}{\operatorname{argmax}} \left\{ \sum_{n=1}^{N} \ln p(x_n;\theta) \right\} = \underset{\theta}{\operatorname{argmax}} \ \ell(\theta)
\end{aligned}$$

Methods of Inference

Until now we've studied estimating and predicting.

A different problem:

- We hold a belief or theory concerning the probabilities generating data
- Are interested in whether the data provides evidence for/against that theory

Example:

- $x_1, \ldots, x_N \overset{\text{IID}}{\sim} \mathcal{N}(\theta, \sigma^2)$ with θ and σ unknown
- $\hat{\theta} := \bar{x}$, an estimator of θ

Suppose theory implies specific value θ_0 for θ

- Prices should be equal to marginal cost
- Excess profits should be equal to zero, etc.

What light does realization $\hat{\theta}$ shed on our theory?

Naive answer: $\hat{\theta}$ contradicts our theory when it's a long way from our hypothesized value θ_0

But what is "a long way"?

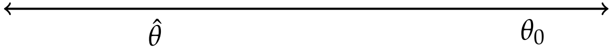


Figure: Theoretical and realized values

To determine what "a long way" means:

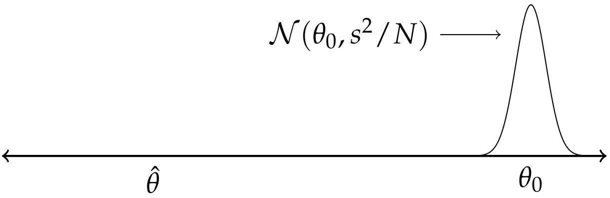
Look at the distribution of $\hat{\theta}$

In the present case, $\hat{\theta} := \bar{x} \sim \mathcal{N}(\theta, \sigma^2 / N)$

Our theory specifies that θ should be equal to θ_0

Parameter σ^2 can be estimated consistently by sample variance s^2

Thus, $\mathcal{N}(\theta_0, s^2 / N) =$ hypothesized density for $\hat{\theta}$ given beliefs



We see that $\hat{\theta}$ is a long way from θ_0

Meaning: If our theory correct, then $\hat{\theta}$ a realization from way out in the tail of its own distribution

Thus, realization $\hat{\theta}$ "unlikely" when our theory is true

Can be construed as evidence against the theory

Confidence Sets

We observe sample $\mathbf{x} := (x_1, \ldots, x_N)$ generated by model M_θ

- $\mathbb{P}_\theta\{\mathbf{x} \in B\} :=$ prob $\mathbf{x} \in B$ when $\mathbf{x}$ generated by M_θ

Here $\theta \in \Theta$ an abstract index—model not necessarily parametric

Confidence set:

- Set $C \subset \Theta$ of parameters that are "plausible" given $\mathbf{x}$
- Translation: Models $\{M_\theta\}_{\theta \in C}$ that are "plausible" given $\mathbf{x}$

Let $\alpha \in [0,1]$

Random set $C(\mathbf{x}) \subset \Theta$ a $1-\alpha$ **confidence set** if

$$\mathbb{P}_\theta\{\theta \in C(\mathbf{x})\} \geq 1-\alpha \quad \text{for all} \quad \theta \in \Theta$$

Remarks:

- It's the set that's random here, not the parameter θ
- If set is an interval, then also called a **confidence interval**

Sequence $C_N(\mathbf{x})$ an **asymptotic** $1-\alpha$ **confidence set** if

$$\lim_{N\to\infty} \mathbb{P}_\theta\{\theta \in C_N(\mathbf{x})\} \geq 1-\alpha \quad \text{for all} \quad \theta \in \Theta$$

Example: Confidence Sets for the ecdf

FTS: If $x_1, \ldots, x_N \overset{\text{IID}}{\sim} F$, then $\sup_{s \in \mathbb{R}} |F_N(s) - F(s)| \xrightarrow{p} 0$

In 1933, A.N. Kolmogorov derived the asymptotic distribution

$$\sqrt{N} \sup_{s \in \mathbb{R}} |F_N(s) - F(s)| \xrightarrow{d} K$$

where K is the **Kolmogorov** distribution

$$K(s) := \frac{\sqrt{2\pi}}{s} \sum_{i=1}^{\infty} \exp\left[-\frac{(2i-1)^2 \pi^2}{8s^2}\right] \qquad (s \geq 0)$$

Can be used to form asymptotic confidence set for F

Let

- $\mathfrak{F} :=$ the set of all cdfs on $\mathbb{R}$
- $k_{1-\alpha} := K^{-1}(1-\alpha)$

Define $C_N(\mathbf{x})$ to be the set of all $G \in \mathfrak{F}$ such that

$$F_N(s) - \frac{k_{1-\alpha}}{\sqrt{N}} \leq G(s) \leq F_N(s) + \frac{k_{1-\alpha}}{\sqrt{N}} \text{ for all } s \in \mathbb{R}$$

Claim: $C_N(\mathbf{x}) \subset \mathfrak{F}$ is an asymptotic $1-\alpha$ confidence set for F

That is, $\lim_{N\to\infty} \mathbb{P}\{F \in C_N(\mathbf{x})\} \geq 1-\alpha$

Proof: Next slide

By definition,

$$F \in C_N(\mathbf{x}) \iff F_N(s) - \frac{k_{1-\alpha}}{\sqrt{N}} \leq F(s) \leq F_N(s) + \frac{k_{1-\alpha}}{\sqrt{N}} \quad \text{for all } s$$

Alternatively,

$$\begin{aligned} F \in C_N(\mathbf{x}) &\iff -k_{1-\alpha} \leq \sqrt{N}(F_N(s) - F(s)) \leq k_{1-\alpha} \quad \text{for all } s \\ &\iff \sqrt{N}|F_N(s) - F(s)| \leq k_{1-\alpha} \quad \text{for all } s \\ &\iff \sup_s \sqrt{N}|F_N(s) - F(s)| \leq k_{1-\alpha} \end{aligned}$$

$$\begin{aligned} \therefore \quad \mathbb{P}\{F \in C_N(\mathbf{x})\} &= \mathbb{P}\left\{\sup_s \sqrt{N}|F_N(s) - F(s)| \leq k_{1-\alpha}\right\} \\ &\to K(k_{1-\alpha}) = 1 - \alpha \end{aligned}$$

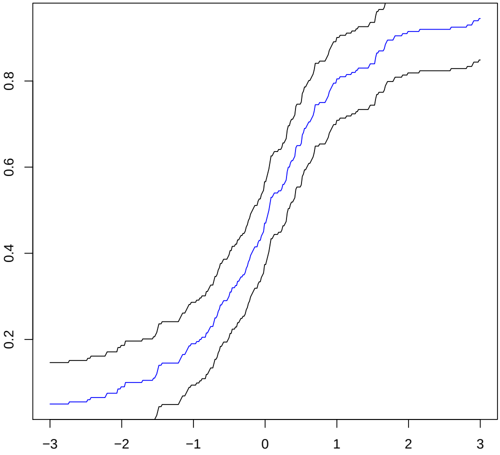


Figure: $N = 200$

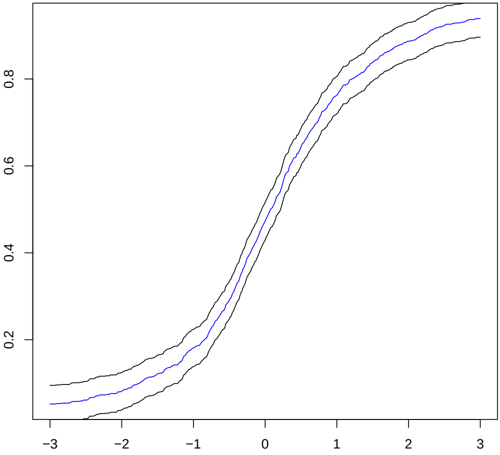


Figure: $N = 10^3$

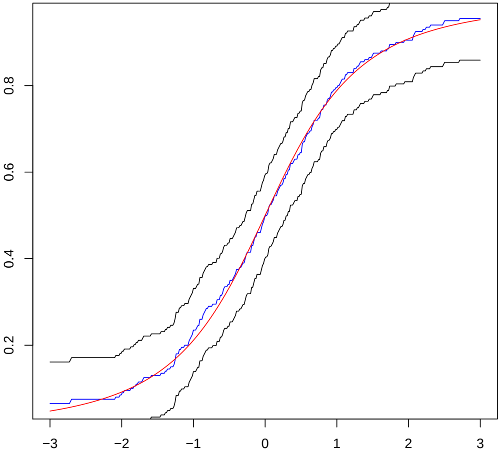


Figure: $N = 200$, true F in red

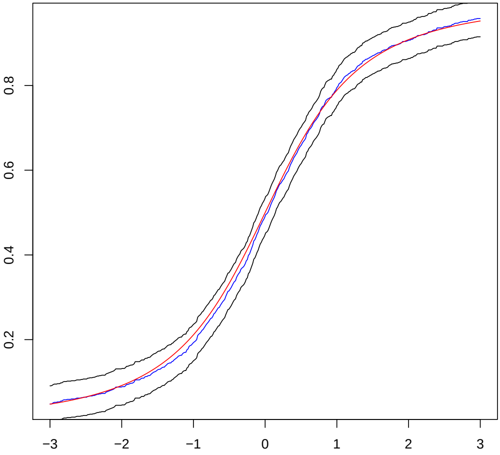


Figure: $N = 10^3$, true F in red

Hypothesis Tests

Begins with a specification of the **null hypothesis**: Data is generated by a given model/class of models

Test of the null hypothesis: An attempt to reject it

Why try to reject rather than confirm?

Example: Consider testing theory that all swans are white

- No amount of white swan sightings proves theory true
- On the other hand, a single black swan proves theory false

Highlights fundamental asymmetry in testing theory by observation

Convention: Only reject null if find strong evidence against it

How did this convention come about?

Suppose we have a collection of theories about how economy works

Now step through theories,

- taking validity as the null, and
- attempting to reject

If the theory rejected then we discard it—process of elimination

However, don't want to mistakenly discard a good theory

So don't reject unless find strong evidence against null

Two mistakes we can make in our test of null hypothesis:

$$\textbf{type I error} \leftrightarrow \text{ reject true null}$$

$$\textbf{type II error} \leftrightarrow \text{ fail to reject false null}$$

Aim: To be conservative when it comes to rejecting null

Method: Design test so that probability of type I error small

Implementation

Null hypothesis H_0:

$$\mathbf{x} := (x_1, \ldots, x_N) \text{ generated by } M_\theta \text{ where } \theta \in \Theta_0$$

A **test** is a binary function ϕ mapping $\mathbf{x}$ into $\{0, 1\}$

The decision rule is

$$\begin{aligned} &\text{if } \phi(\mathbf{x}) = 1, \text{ then reject } H_0 \\ &\text{if } \phi(\mathbf{x}) = 0, \text{ then do not reject } H_0 \end{aligned}$$

"Do not reject" $\neq$ "accept" — see "Argument from ignorance"

Power Function

The **power function** associated with test ϕ is the function

$$\beta(\theta) := \mathbb{P}_\theta\{\phi(\mathbf{x}) = 1\}$$

Ideally

- $\beta(\theta) = 1$ when $\theta \notin \Theta_0$
- $\beta(\theta) = 0$ when $\theta \in \Theta_0$

In practice, this is usually difficult to achieve

Note that $\sup_{\theta \in \Theta_0} \beta(\theta) =$ max probability of type I error

We want to keep probability of type I error small

Standard procedure: Choose small number α, adjust test such that

$$\beta(\theta) \leq \alpha \quad \text{for all } \theta \in \Theta_0 \tag{2}$$

If (2) holds, then ϕ said to be of **size α**

Writing $\beta_N := \beta$ where N is the sample size, suppose

$$\lim_{N \to \infty} \beta_N(\theta) \leq \alpha \quad \text{for all } \theta \in \Theta_0 \tag{3}$$

If (3) holds, then $\phi_N = \phi$ called **asymptotically of size α**

Example: Asset Price Returns

Standardized daily returns on Nikkei 225, Jan 1984 – May 2009

Let $x_n :=$ return on n-th day

Standardized return is

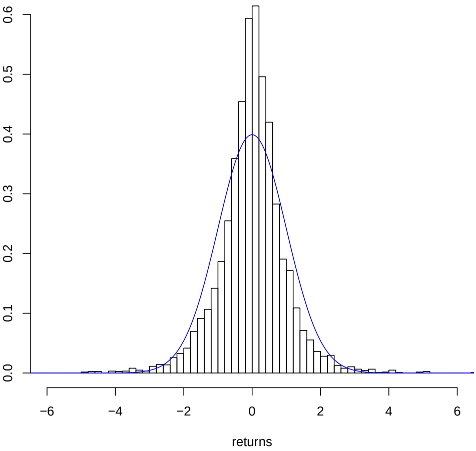
$$\tilde{x}_n := \frac{x_n - \bar{x}}{s}$$

Suppose $x_n \sim \mathcal{N}(\mu, \sigma^2)$ for some μ, σ

Then

$$\tilde{x}_n := \frac{x_n - \bar{x}}{s} \approx \frac{x_n - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

Histogram of standardized returns, $\mathcal{N}(0,1)$ superimposed in blue



Fit to standard normal is not particularly good

Histogram suggests that the density of returns is

- more peaked
- has heavier tails

This is a common observation for asset price returns

Can we test this more formally?

Let

- $\Phi :=$ standard normal cdf
- $F_N :=$ ecdf of standardized returns

Null hypothesis: Standardized returns are IID draws from Φ

Under the null,

$$\sqrt{N} \sup_{s \in \mathbb{R}} |F_N(s) - \Phi(s)| \xrightarrow{d} K$$

We can use this information to construct a test of size α

Let α be given, and let $k_{1-\alpha} = K^{-1}(1-\alpha)$ as before

Consider the test

$$\phi_N(\mathbf{x}) := \mathbb{1}\left\{\sqrt{N}\sup_{s\in\mathbb{R}}|F_N(s) - \Phi(s)| > k_{1-\alpha}\right\}$$

Let $\beta_N(\Phi)$ be the value of the power function when H_0 true

We have

$$\lim_{N\to\infty}\beta_N(\Phi) = \lim_{N\to\infty}\mathbb{P}\left\{\sqrt{N}\sup_{s\in\mathbb{R}}|F_N(s) - \Phi(s)| > k_{1-\alpha}\right\} = \alpha$$

Hence test is asymptotically of size α

The value of the statistic $\sqrt{N} \sup_{s \in \mathbb{R}} |F_N(s) - \Phi(s)|$ is 5.67

If $\alpha = 0.05$, then $k_{1-\alpha}$ is 1.36, so reject null

Looking back, there's an obvious problem in our approach

We were interested in testing normality

Our null hypothesis was that standardized returns are $\overset{\text{IID}}{\sim} \Phi$

Rejection of null may be due to IID assumption

Extensions to non-IID case beyond scope of this course

See course notes for references

Linear Algebra

Review of

- Vectors and vector operations
- Matrices and matrix operations
- Linear mappings
- Systems of equations

Vectors

N-vector is a sequence of N numbers:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} \quad \text{where } x_n \in \mathbb{R} \text{ for each } n$$

Can also write $\mathbf{x}$ horizontally, like so: $\mathbf{x} = (x_1, \ldots, x_N)$

$\mathbb{R}^N :=$ set of all N-vectors

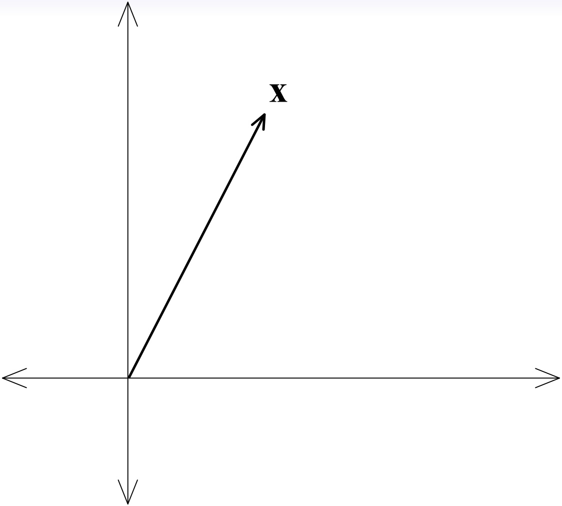


Figure: Vector $\mathbf{x} = (x_1, x_2)$ in $\mathbb{R}^2$

The vector of ones will be denoted $\mathbf{1}$

$$\mathbf{1} := \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

Vector of zeros will be denoted $\mathbf{0}$

$$\mathbf{0} := \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

Two fundamental algebraic operations:

- vector addition
- scalar multiplication

1. **Sum** of $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{y} \in \mathbb{R}^N$ defined by

$$\mathbf{x} + \mathbf{y} =: \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} := \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_N + y_N \end{pmatrix}$$

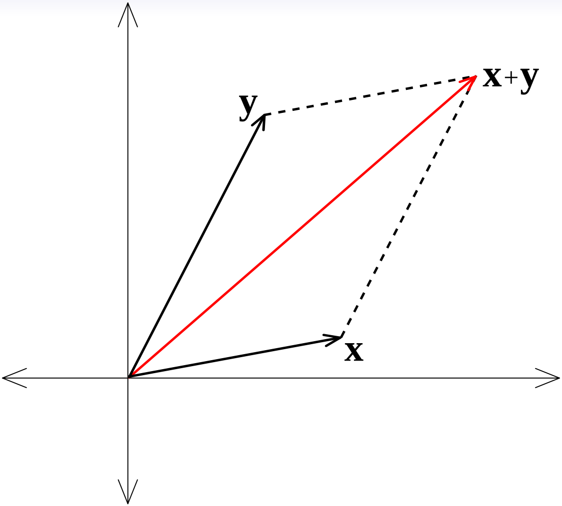


Figure: Vector addition

2. **Scalar product** of $\alpha \in \mathbb{R}$ and $\mathbf{x}$ defined by

$$\alpha\mathbf{x} := \begin{pmatrix} \alpha x_1 \\ \alpha x_2 \\ \vdots \\ \alpha x_N \end{pmatrix}$$

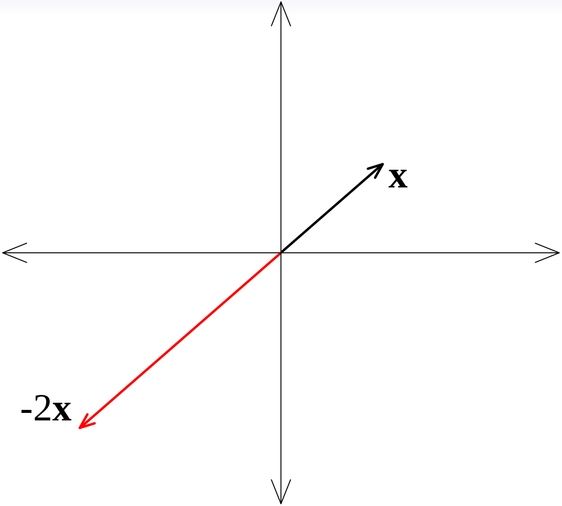


Figure: Scalar multiplication

Subtraction performed element by element, analogous to addition

$$\mathbf{x} - \mathbf{y} := \begin{pmatrix} x_1 - y_1 \\ x_2 - y_2 \\ \vdots \\ x_N - y_N \end{pmatrix}$$

Def can be given in terms of addition and scalar multiplication:

$$\mathbf{x} - \mathbf{y} := \mathbf{x} + (-1)\mathbf{y}$$

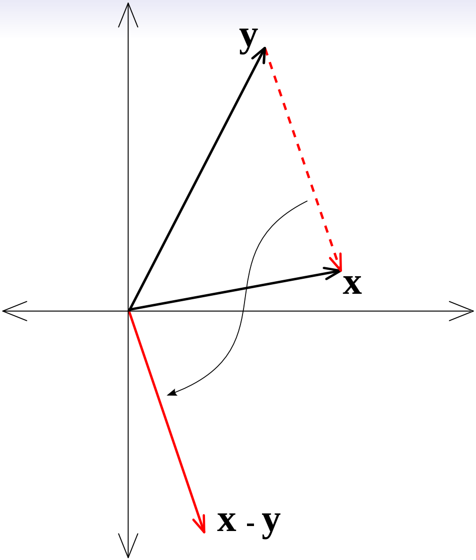


Figure: Difference between vectors

Inner product of two vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^N$ is

$$\mathbf{x}'\mathbf{y} := \sum_{n=1}^{N} x_n y_n = \mathbf{y}'\mathbf{x}$$

The (euclidean) **norm** of $\mathbf{x} \in \mathbb{R}^N$ is defined as

$$\|\mathbf{x}\| := \sqrt{\mathbf{x}'\mathbf{x}} = \left(\sum_{n=1}^{N} x_n^2 \right)^{1/2}$$

Interpretations:

- $\|\mathbf{x}\|$ represents the "length" of $\mathbf{x}$
- $\|\mathbf{x} - \mathbf{y}\|$ represents distance between $\mathbf{x}$ and $\mathbf{y}$

For any $\alpha \in \mathbb{R}$ and any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$, following properties are satisfied:

1. $\|\mathbf{x}\| \geq 0$ and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$
2. $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$

Third property called the **triangle inequality**

Matrices

Typical $N \times K$ **matrix**:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ a_{21} & a_{22} & \cdots & a_{2K} \\ \vdots & \vdots & & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NK} \end{pmatrix}$$

Symbol a_{nk} stands for element in the n-th row of the k-th column

$N \times K$ matrix also called a

- **row vector** if $N = 1$
- **column vector** if $K = 1$

If $N = K$, then $\mathbf{A}$ called **square**

If square and $a_{nk} = a_{kn}$ for every k and n, then called **symmetric**

For square $\mathbf{A}$, elements a_{nn} called the **principal diagonal**:

$$\begin{pmatrix} \mathbf{a_{11}} & a_{12} & \cdots & a_{1N} \\ a_{21} & \mathbf{a_{22}} & \cdots & a_{2N} \\ \vdots & \vdots & & \vdots \\ a_{N1} & a_{N2} & \cdots & \mathbf{a_{NN}} \end{pmatrix}$$

Identity matrix:

$$\mathbf{I} := \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

Algebraic Operations for Matrices

Addition and scalar multiplication are also defined for matrices

There is also a new operation: Matrix multiplication

Scalar multiplication is element by element, as in the vector case:

$$\gamma \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ a_{21} & a_{22} & \cdots & a_{2K} \\ \vdots & \vdots & & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NK} \end{pmatrix} := \begin{pmatrix} \gamma a_{11} & \gamma a_{12} & \cdots & \gamma a_{1K} \\ \gamma a_{21} & \gamma a_{22} & \cdots & \gamma a_{2K} \\ \vdots & \vdots & & \vdots \\ \gamma a_{N1} & \gamma a_{N2} & \cdots & \gamma a_{NK} \end{pmatrix}$$

Addition also element by element:

$$
\begin{pmatrix}
a_{11} & \cdots & a_{1K} \\
a_{21} & \cdots & a_{2K} \\
\vdots & \vdots & \vdots \\
a_{N1} & \cdots & a_{NK}
\end{pmatrix}
+
\begin{pmatrix}
b_{11} & \cdots & b_{1K} \\
b_{21} & \cdots & b_{2K} \\
\vdots & \vdots & \vdots \\
b_{N1} & \cdots & b_{NK}
\end{pmatrix}
:=
\begin{pmatrix}
a_{11}+b_{11} & \cdots & a_{1K}+b_{1K} \\
a_{21}+b_{21} & \cdots & a_{2K}+b_{2K} \\
\vdots & \vdots & \vdots \\
a_{N1}+b_{N1} & \cdots & a_{NK}+b_{NK}
\end{pmatrix}
$$

Note that matrices must be same dimension

Multiplication of matrices:

Product $\mathbf{AB}$: i, j-th element is inner product of i-th row of $\mathbf{A}$ and j-th column of $\mathbf{B}$

$$\begin{pmatrix} \mathbf{a_{11}} & \cdots & \mathbf{a_{1K}} \\ a_{21} & \cdots & a_{2K} \\ \vdots & \vdots & \vdots \\ a_{N1} & \cdots & a_{NK} \end{pmatrix} \begin{pmatrix} \mathbf{b_{11}} & \cdots & b_{1J} \\ \mathbf{b_{21}} & \cdots & b_{2J} \\ \vdots & \vdots & \vdots \\ \mathbf{b_{K1}} & \cdots & b_{KJ} \end{pmatrix} = \begin{pmatrix} \mathbf{c_{11}} & \cdots & c_{1J} \\ c_{21} & \cdots & c_{2J} \\ \vdots & \vdots & \vdots \\ c_{N1} & \cdots & c_{NJ} \end{pmatrix}$$

In this display,

$$c_{11} = \text{row}_1(\mathbf{A})' \text{col}_1(\mathbf{B}) = \sum_{k=1}^{K} a_{1k} b_{k1}$$

Suppose $\mathbf{A}$ is $N \times K$ and $\mathbf{B}$ is $J \times M$

- $\mathbf{AB}$ defined only if $K = J$
- Resulting matrix $\mathbf{AB}$ is $N \times M$

The rule to remember:

$$\text{product of } N \times K \text{ and } K \times M \text{ is } N \times M$$

Multiplication is not commutative: $\mathbf{AB} \neq \mathbf{BA}$

In fact $\mathbf{BA}$ is not well-defined unless $N = M$ also holds

Rules for multiplication:

For conformable matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, we have

- $\mathbf{A}(\mathbf{B}\mathbf{C}) = (\mathbf{A}\mathbf{B})\mathbf{C}$
- $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{A}\mathbf{B} + \mathbf{A}\mathbf{C}$
- $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{A}\mathbf{C} + \mathbf{B}\mathbf{C}$

(Here "conformable" means operation makes sense)

Comments

Assignment due next Thursday (14th April)

I'll be around over teaching break

Pleeease prepare for mid term over teaching break

We'll finish chapter 6 (and 7?) next lecture (April 28)