APTS High-dimensional statistics: Preliminary material

Yi Yu
University of Warwick
yi.yu.2@warwick.ac.uk

May 22, 2022

1 Introduction

J. E. Littlewood wrote in the *Lectures on the Theory of Functions* that

> *There are three principles, roughly expressible in the following terms: Every (measurable) set is nearly a finite sum of intervals; every function (of class L_p) is nearly continuous; every convergent sequence of functions is nearly uniformly convergent.*

These famous "nearly" principles provide us with some level of mathematical justifications on focusing on the linear regressions in many scenarios. In this APTS course, we start with the classical linear regression problems, then heading to the linear regression problems in high-dimensional settings, graphical settings, change point settings, as well as functional data analysis settings.

We will mostly focus on the theoretical justifications of the widely-adopted methods, aiming to lay the necessary foundation, no matter if you are going to pursue more theoretical or applied Ph.D. projects in the upcoming years. This course will cover a wide range of topics, but what is not covered in this course is way more than what is covered. Because of this, I will also spend significant amount of time to discuss what is not covered in each topic, as well as the topics which will not be covered at all.

1.1 Books

The material of this course comes from many different areas, but the following two are most important if you are new to this topic.

(HDP) Vershynin, Roman. High-dimensional probability: An introduction with applications in data science. Vol. 47. Cambridge university press, 2018.

(HDS) Wainwright, Martin J. High-dimensional statistics: A non-asymptotic viewpoint. Vol. 48. Cambridge University Press, 2019.

Both books are beyond excellent and both are worth being read thoroughly, despite your background and your future research interest. In the interest of time, you may want to read the first four chapters of HDP carefully before the start of this course. If you do not even have time for this task, you may want to read the following of this document, which is very nicely prepared by Rajen Shah (Cambridge).

1.2 Notation

Here we collect some matrix and vector notation we use in this preliminary material and throughout the course.

Given $A, B \subseteq \{1, \ldots, p\}$, and $\mathbf{x} \in \mathbb{R}^p$, we will write $\mathbf{x}_A$ for the sub-vector of $\mathbf{x}$ formed from those components of $\mathbf{x}$ indexed by A. Similarly, we will write $\mathbf{M}_A$ for the submatrix of $\mathbf{M}$ formed from those columns of $\mathbf{M}$ indexed by A. Further, $\mathbf{M}_{A,B}$ will be the submatrix of $\mathbf{M}$ formed from columns and rows indexed by A and B respectively. For example, $\mathbf{x}_{\{1,2\}} = (x_1, x_2)^T$, $\mathbf{M}_{\{1,2\}}$ is the matrix formed from the first two columns of $\mathbf{M}$, and $\mathbf{M}_{\{1,2\},\{1,2\}}$ is the top left 2×2 submatrix of $\mathbf{M}$.

In addition, when used in subscripts, we will use $-j$ and $-jk$ to denote $\{1, \ldots, p\} \setminus \{j\} := \{j\}^c$ and $\{1, \ldots, p\} \setminus \{j, k\} := \{j, k\}^c$ respectively. So for example, $\mathbf{M}_{-jk}$ is the submatrix of $\mathbf{M}$ that has columns j and k removed.

The matrix and vector subsetting operations will always occur first, so e.g. $\mathbf{M}_A^T = (\mathbf{M}_A)^T$.

2 Norms

For a d-dimensional vector $\mathbf{v} \in \mathbb{R}^d$, its ℓ_p-norm, where $p \in [1, \infty)$ is defined to be

$$\|\mathbf{v}\|_p = \left(\sum_{j=1}^{d} |v_j|^p \right)^{1/p}.$$

We also define the ℓ_∞-norm $\|\mathbf{v}\|_\infty = \max_j |v_j|$. We will primarily be interested in the cases $p = 1, 2, \infty$. One can show that

(i) for a scalar $t \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{R}^d$, $\|t\mathbf{v}\|_p = |t|\|\mathbf{v}\|_p$;

(ii) if $\|\mathbf{v}\|_p = 0$ then $\mathbf{v} = \mathbf{0}$;

(iii) for $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$, $\|\mathbf{u} + \mathbf{v}\|_p \leq \|\mathbf{u}\|_p + \|\mathbf{v}\|_p$.

Properties (i) and (ii) are rather clear from the definition, but showing (iii), which is known as the *triangle inequality*, is more involved.

Exercise 2.1. Show that we also have what is sometimes known as the reverse triangle inequality, that

$$\|\mathbf{u} - \mathbf{v}\|_p \geq \|\mathbf{u}\|_p - \|\mathbf{v}\|_p.$$

Hölder's inequality states that when $p, q \in [1, \infty]$ are such that $p^{-1} + q^{-1} = 1$, where $1/\infty$ is understood to be 0,

$$|\mathbf{v}^T\mathbf{u}| \leq \|\mathbf{v}\|_p \|\mathbf{u}\|_q.$$

The case where $p = q = 2$ is known as the *Cauchy–Schwarz inequality.*

Exercise 2.2. Show that $\|\mathbf{u} + \mathbf{v}\|_2^2 = \|\mathbf{u}\|_2^2 + 2\mathbf{u}^T\mathbf{v} + \|\mathbf{v}\|_2^2$. Further show property (iii) above for the ℓ_2-norm using the Cauchy–Schwarz inequality.

Exercise 2.3. Prove Hölder's inequality when $p = 1$, $q = \infty$ i.e. show that

$$\|\mathbf{u}\|_\infty \|\mathbf{v}\|_1 = \max_j |u_j| \sum_k |v_k| \geq |\mathbf{u}^T\mathbf{v}|.$$

For $u \in \mathbb{R}$ let us define

$$\operatorname{sgn}(u) = \begin{cases} 1 & \text{if } u > 0 \\ 0 & \text{if } u = 0 \\ -1 & \text{if } u < 0. \end{cases}$$

With a slight abuse of notation, for $\mathbf{v} \in \mathbb{R}^d$, also define $\operatorname{sgn}(\mathbf{v}) = (\operatorname{sgn}(v_1), \ldots, \operatorname{sgn}(v_d))^T$. Note that $\operatorname{sgn}(\mathbf{v})^T\mathbf{v} = \|\mathbf{v}\|_1$.

Exercise 2.4. Show that $\|\mathbf{v}\|_1 \leq \sqrt{d}\|\mathbf{v}\|_2$ when $\mathbf{v} \in \mathbb{R}^d$.

3 Matrix algebra

The course will assume you are already familiar with the APTS Statistical Computing module and have a thorough understanding of linear algebra. We briefly review some key elements of this here, as well as adding some more material that will be useful for our developments.

Any symmetric matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$ may be expressed in its *eigendecomposition*:

$$\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{U}^T$$

where $\mathbf{U} \in \mathbb{R}^{d \times d}$ is an orthogonal matrix whose columns are eigenvectors of $\mathbf{M}$ (so $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}$) and $\mathbf{D}$ is diagonal with $D_{11} \geq D_{22} \geq \cdots \geq D_{dd}$ being the corresponding eigenvalues of $\mathbf{M}$. We say such an $\mathbf{M}$ is *positive semi-definite* if $\mathbf{u}^T\mathbf{M}\mathbf{u} \geq 0$ for all $\mathbf{u} \in \mathbb{R}^d$. It is *positive definite* if $\mathbf{u}^T\mathbf{M}\mathbf{u} > 0$ for all $\mathbf{u} \neq 0$.

Exercise 3.1. Check that $\|\mathbf{U}\mathbf{v}\|_2 = \|\mathbf{v}\|_2$ for orthogonal $\mathbf{U}$.

Exercise 3.2. Show that a symmetric matrix is positive definite if and only if all its eigenvalues are positive. Argue that a positive definite matrix is invertible.

Exercise 3.3. Show that if $\mathbf{A} \in \mathbb{R}^{d \times p}$ then $\mathbf{A}^T\mathbf{A}$ is positive semi-definite.

The maximum and minimum eigenvalues, $c_{\min}(\mathbf{M}), c_{\max}(\mathbf{M})$, of a symmetric matrix $\mathbf{M}$ obey the following.

$$c_{\max}(\mathbf{M}) = \sup_{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\|_2 = 1} \|\mathbf{M}\mathbf{v}\|_2, \qquad c_{\min}(\mathbf{M}) = \inf_{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\|_2 = 1} \|\mathbf{M}\mathbf{v}\|_2.$$

Indeed,

$$\begin{aligned}
\sup_{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\|_2 = 1} \|\mathbf{M}\mathbf{v}\|_2 &= \sup_{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\|_2 = 1} \sqrt{\mathbf{v}^T\mathbf{U}\mathbf{D}^2\mathbf{U}^T\mathbf{v}} \\
&= \sup_{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\|_2 = 1} \sqrt{\mathbf{u}^T\mathbf{D}^2\mathbf{u}} \quad \text{making the substitution } \mathbf{u} = \mathbf{U}^T\mathbf{v} \\
&= \sup_{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\|_2 = 1} \left(\sum_{j=1}^{d} D_{jj}^2 u_j^2 \right)^{1/2} \\
&\leq \left\{ \sup_{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\|_2 = 1} \big(\max_{j=1,\ldots,d} D_{jj}^2 \|\mathbf{u}\|_2^2 \big) \right\}^{1/2} \quad \text{by exercise 2.3} \\
&= D_{11} = c_{\max}(\mathbf{M}).
\end{aligned}$$

The inequality above is an equality when $\mathbf{u}$ has $u_1 = 1$, $u_j = 0$ for all $j > 1$.

Exercise 3.4. Write out the argument for the corresponding result for the minimum eigenvalue. Show further that for any $A \subseteq \{1, \ldots, d\}$, $c_{\min}(\mathbf{M}) \leq c_{\min}(\mathbf{M}_{A,A}) \leq c_{\max}(\mathbf{M}_{A,A}) \leq c_{\max}(\mathbf{M})$.

The trace $\mathrm{tr}(\mathbf{M})$ of a square matrix is the sum of its diagonal entries:

$$\mathrm{tr}(\mathbf{M}) = \sum_{j=1}^{d} M_{jj}.$$

If matrices $\mathbf{A}$ and $\mathbf{B}$ have dimensions such that $\mathbf{AB}$ and $\mathbf{BA}$ are valid matrix multiplications, then $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$.

Exercise 3.5. Show that the trace of a symmetric matrix is the sum of its eigenvalues.

The *singular value decomposition* (SVD) is a generalisation of an eigendecomposition of a square matrix. We can factorise any $\mathbf{X} \in \mathbb{R}^{n \times p}$ into its SVD

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T.$$

Here the $\mathbf{U} \in \mathbb{R}^{n\times n}$ and $\mathbf{V} \in \mathbb{R}^{p\times p}$ are orthogonal matrices and $\mathbf{D} \in \mathbb{R}^{n\times p}$ has $D_{11} \geq D_{22} \geq \cdots \geq D_{mm} \geq 0$ where $m = \min(n, p)$ and all other entries of $\mathbf{D}$ are zero. To compute such a decomposition typically requires $O(np\min(n, p))$ operations. The rth columns of $\mathbf{U}$ and $\mathbf{V}$ are known as the rth left and right singular vectors of $\mathbf{X}$ respectively, and D_{rr} is the rth singular value.

When $n > p$, we can replace $\mathbf{U}$ by its first p columns and $\mathbf{D}$ by its first p rows to produce another version of the SVD (sometimes known as the thin SVD). Then $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ where $\mathbf{U} \in \mathbb{R}^{n\times p}$ has orthonormal columns (but is no longer square) and $\mathbf{D}$ is square and diagonal. There is an analogous version for when $p > n$.

4 Multivariate calculus

Given a function $f : \mathbb{R}^d \to \mathbb{R}$, we will denote the column vector of partial derivatives or gradient vector by

$$\frac{\partial f}{\partial \mathbf{x}} = \frac{\partial}{\partial \mathbf{x}} f = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_d} \right)^T.$$

You may be more familiar with the alternative notation ∇f. Check that you are happy with the following derivatives of common functions:

$$\frac{\partial(\mathbf{c}^T\mathbf{x})}{\partial \mathbf{x}} = \mathbf{c}$$
$$\frac{\partial(\mathbf{x}^T\mathbf{A}\mathbf{x})}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}.$$

It is straightforward (but slightly tedious) to show these results by e.g. expressing $\mathbf{x}^T\mathbf{A}\mathbf{x} = \sum_{i,j} x_i A_{ij} x_j$ and differentiating this with respect to x_k.

Exercise 4.1. Compute

$$\frac{\partial}{\partial \boldsymbol{\beta}} \|\boldsymbol{\beta}\|_2^2/2$$
$$\frac{\partial}{\partial \boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/2.$$

Of course the chain rule then also gives, for example

$$\frac{\partial(g(\mathbf{x}^T\mathbf{A}\mathbf{x}))}{\partial \mathbf{x}} = g'(\mathbf{x}^T\mathbf{A}\mathbf{x})(\mathbf{A} + \mathbf{A}^T)\mathbf{x}.$$

Exercise 4.2. Compute

$$\frac{\partial \|\boldsymbol{\beta}\|_2}{\partial \boldsymbol{\beta}}$$

when $\boldsymbol{\beta} \neq 0$.

5 Convexity

In recent years the fields of optimisation and statistics have grown much closer. Researchers in many areas of statistics are now expected to have a good grasp of basic topics in convex optimisation in particular. High-dimensional statistics is one such area, with convexity playing a crucial role in the formulation of key methods such as the Lasso, which we will study in detail in the course.

Here we review some basic facts about convex sets and functions, which will provide a foundation for the more detailed treatment of convex analysis and optimisation in the course.

A set $A \subseteq \mathbb{R}^d$ is *convex* if

$$\mathbf{x}, \mathbf{y} \in A \Rightarrow (1-t)\mathbf{x} + t\mathbf{y} \in A \qquad \text{for all } t \in (0,1).$$

In words, given any two points in A, the line segment between them is contained in A.

Exercise 5.1. Show that the set of symmetric $d \times d$ positive definite matrices is a convex subset of $\mathbb{R}^{d\times d}$.

A function $f : \mathbb{R}^d \to \mathbb{R}$ is *convex* if

$$f\big((1-t)\mathbf{x} + t\mathbf{y}\big) \leq (1-t)f(\mathbf{x}) + tf(\mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $t \in (0,1)$. It is *strictly convex* if the inequality is strict for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $\mathbf{x} \neq \mathbf{y}$ and $t \in (0,1)$.

Exercise 5.2. Let $f_1, \dots, f_m : \mathbb{R}^d \to \mathbb{R}$ be convex functions. Show that if $c_1, \dots, c_m \geq 0$, $c_1 f_1 + \cdots c_m f_m$ is a convex function. Show furthermore that if one of the functions f_j is strictly convex, then the sum above is a strictly convex function.

Exercise 5.3. Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is convex and $\mathbf{A} \in \mathbb{R}^{d\times m}$. Show that $g : \mathbb{R}^m \to \mathbb{R}$ defined by $g(\mathbf{x}) = f(\mathbf{A}\mathbf{x})$ is convex.

Exercise 5.4. Show that if a strictly convex function f has a minimiser, then it must be unique.

Proposition 1. *If $f : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable then*

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{0} \ \textit{ implies that } \ \mathbf{x} \ \textit{minimises } f.$$

Proposition 2. *If $f : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable then*

(i) f is convex iff. its Hessian $\mathbf{H}(\mathbf{x})$ is positive semi-definite for all $\mathbf{x} \in \mathbb{R}^d$,

(ii) f is strictly convex if $\mathbf{H}(\mathbf{x})$ is positive definite for all $\mathbf{x} \in \mathbb{R}^d$.

Exercise 5.5. Explain why $\boldsymbol{\beta} \mapsto \|\boldsymbol{\beta}\|_2^2$ is strictly convex.

Exercise 5.6. Show that if

$$\hat{\boldsymbol{\beta}} = \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p} \{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2\}$$

then $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$.

6 Basic tail bounds

Tail bounds are vital for the study of many modern statistical algorithms. Here we will review the most basic of these. We begin our discussion with the simplest tail bound, *Markov's inequality*. This states that given a non-negative random variable W,

$$\mathbb{P}(W \geq t) \leq \frac{\mathbb{E}(W)}{t}.$$

It follows from taking expectations of both sides of the inequality $t\mathbb{1}_{\{W \geq t\}} \leq W$. This immediately implies that given a strictly increasing function $\varphi : \mathbb{R} \to [0, \infty)$ and any random variable W,

$$\mathbb{P}(W \geq t) = \mathbb{P}\{\varphi(W) \geq \varphi(t)\} \leq \frac{\mathbb{E}(\varphi(W))}{\varphi(t)},$$

provided $\varphi(t) > 0$. Applying this with $\varphi(t) = e^{\alpha t}$ ($\alpha > 0$) yields the so-called *Chernoff bound*:

$$\mathbb{P}(W \geq t) \leq \inf_{\alpha > 0} e^{-\alpha t}\mathbb{E}e^{\alpha W}.$$

Consider the case when $W \sim \mathcal{N}(0, \sigma^2)$. Recall that the moment generating function (mgf) of W is

$$\mathbb{E}e^{\alpha W} = e^{\alpha^2\sigma^2/2}. \tag{6.1}$$

Thus

$$\mathbb{P}(W \geq t) \leq \inf_{\alpha > 0} e^{\alpha^2\sigma^2/2 - \alpha t} = e^{-t^2/(2\sigma^2)}.$$

Note that to arrive at this bound, all we required was (an upper bound on) mgf of W (6.1). This motivates the following definition.

Definition 1. We say a random variable W with mean $\mu = \mathbb{E}(W)$ is *sub-Gaussian* if there exists $\sigma > 0$ such that

$$\mathbb{E}e^{\alpha(W-\mu)} \leq e^{\alpha^2\sigma^2/2}$$

for all $\alpha \in \mathbb{R}$. We then say that W is *sub-Gaussian with parameter* σ.

The normal example above immediately gives the following result.

Proposition 3 (Sub-Gaussian tail bound)**.** *If W is sub-Gaussian with parameter σ and $\mathbb{E}(W) = \mu$, then*

$$\mathbb{P}(W - \mu \geq t) \leq e^{-t^2/(2\sigma^2)}.$$

It is often helpful to have a tail bound on the maximum of a collection of random variables. A simple *union bound* can be helpful in this regard. This states that given events $\Omega_1, \ldots, \Omega_m$,

$$\mathbb{P}(\cup_m \Omega_j) \leq \sum_j \mathbb{P}(\Omega_j).$$

Exercise 6.1. Show that if $W_1, \ldots, W_m$ are all mean-zero sub-Gaussian random variables with common parameter σ, then

$$\mathbb{P}(\max_j |W_j| \leq 2A\sigma\sqrt{\log(m)}) \leq 2m^{-(2A^2-1)}.$$

7 Linear regression

Imagine data are available in the form of observations $(Y_i, \mathbf{x}_i) \in \mathbb{R} \times \mathbb{R}^p$, $i = 1, \ldots, n$, and the aim is to infer a simple *regression function* relating the average value of a *response*, Y_i, and a collection of *predictors* or *variables*, $\mathbf{x}_i$. This is an example of regression analysis, one of the most important tasks in statistics.

A *linear model* for the data assumes that it is generated according to

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\varepsilon}, \tag{7.1}$$

where $\mathbf{Y} \in \mathbb{R}^n$ is the vector of responses; $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the predictor matrix (or design matrix) with ith row $\mathbf{x}_i^T$; $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ represents random error; and $\boldsymbol{\beta}^0 \in \mathbb{R}^p$ is the unknown vector of coefficients.

Provided $p \ll n$, a sensible way to estimate $\boldsymbol{\beta}^0$ is by ordinary least squares (OLS). This yields an estimator $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ with

$$\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} := \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{7.2}$$

provided $\mathbf{X}$ has full column rank (i.e. the columns of $\mathbf{X}$ are linearly independent so $\mathbf{X}\mathbf{z} = \mathbf{0}$ if and only if $\mathbf{z} = \mathbf{0}$).

Exercise 7.1. Show that if $\mathbf{X}$ has full column rank then $\mathbf{X}^T\mathbf{X}$ is invertible.

Recall that for a random vector $\mathbf{Z} \in \mathbb{R}^d$ and $\mathbf{m} \in \mathbb{R}^k$ and $\mathbf{A} \in \mathbb{R}^{k \times d}$,

$$\mathbb{E}(\mathbf{m} + \mathbf{A}\mathbf{Z}) = \mathbf{m} + \mathbf{A}\mathbb{E}(\mathbf{Z})$$

and

$$\begin{aligned}
\mathrm{Var}(\mathbf{m} + \mathbf{A}\mathbf{Z}) &= \mathbb{E}[\{\mathbf{m} + \mathbf{A}\mathbf{Z} - \mathbb{E}(\mathbf{m} + \mathbf{A}\mathbf{Z})\}\{\mathbf{m} + \mathbf{A}\mathbf{Z} - \mathbb{E}(\mathbf{m} + \mathbf{A}\mathbf{Z})\}^T] \\
&= \mathbb{E}\{\mathbf{A}(\mathbf{Z} - \mathbb{E}\mathbf{Z})(\mathbf{Z} - \mathbb{E}\mathbf{Z})^T\mathbf{A}^T\} \\
&= \mathbf{A}\mathbb{E}\{(\mathbf{Z} - \mathbb{E}\mathbf{Z})(\mathbf{Z} - \mathbb{E}\mathbf{Z})^T\}\mathbf{A}^T \\
&= \mathbf{A}\mathrm{Var}(\mathbf{Z})\mathbf{A}^T.
\end{aligned}$$

Exercise 7.2. Show that when $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$, we have $\mathbb{E}_{\boldsymbol{\beta}^0,\sigma^2}(\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}) = \boldsymbol{\beta}^0$ and $\mathrm{Var}_{\boldsymbol{\beta}^0,\sigma^2}(\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

8 The multivariate normal distribution

You should already know what a univariate normal distribution is: the density is given by

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

where $\mu \in \mathbb{R}$ is the mean and $\sigma^2 > 0$ is the variance.

We say a random variable $\mathbf{Z} \in \mathbb{R}^d$ has a d-variate normal distribution if for every $\mathbf{t} \in \mathbb{R}^d$, $\mathbf{t}^T\mathbf{Z}$ has a univariate normal distribution. The multivariate normal distribution is uniquely characterised by its mean and variance. Thus we can write $\mathbf{Z} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ when $\mathbb{E}(\mathbf{Z}) = \mu$ and $\text{Var}(\mathbf{Z}) = \boldsymbol{\Sigma}$. As a further consequence, we have that for $A, B \subseteq \{1, \ldots, d\}$, $\mathbf{Z}_A$ is independent of $\mathbf{Z}_B$ if and only if $\text{Cov}(\mathbf{Z}_A, \mathbf{Z}_B)$. When $\boldsymbol{\Sigma}$ is positive definite, the density of $\mathbf{Z}$ is

$$f(\mathbf{z}) = \frac{1}{(2\pi)^{p/2}\text{det}(\boldsymbol{\Sigma})^{1/2}} \exp\Big(-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{z} - \boldsymbol{\mu})\Big).$$

Exercise 8.1. Show that affine transformations of a multivariate normal $\mathbf{Z}$ are also normal, that is show that for any $\mathbf{m} \in \mathbb{R}^k$ and $\mathbf{A} \in \mathbb{R}^{k\times d}$, $\mathbf{m} + \mathbf{A}\mathbf{Z} \sim \mathcal{N}_k(\mathbf{m} + \mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$ is multivariate normal.

9 Normal conditionals

Definition 2. If $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ are random vectors with a joint density $f_{\mathbf{XYZ}}$ then we say $\mathbf{X}$ is conditionally independent of $\mathbf{Y}$ given $\mathbf{Z}$, and write

$$\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$$

if

$$f_{\mathbf{XY}|\mathbf{Z}}(\mathbf{x}, \mathbf{y}|\mathbf{z}) = f_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}|\mathbf{z})f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}|\mathbf{z}).$$

Here $f_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}|\mathbf{z})$ for example is the conditional density of $\mathbf{X}$ given $\mathbf{Z}$. Equivalently

$$\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z} \iff f_{\mathbf{X}|\mathbf{YZ}}(\mathbf{x}|\mathbf{y}, \mathbf{z}) = f_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}|\mathbf{z}).$$

Now let $\mathbf{Z} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ positive definite. Note $\boldsymbol{\Sigma}_{A,A}$ is also positive definite for any A.

Proposition 4.

$$\mathbf{Z}_A|\mathbf{Z}_B = \mathbf{z}_B \sim \mathcal{N}_{|A|}(\boldsymbol{\mu}_A + \boldsymbol{\Sigma}_{A,B}\boldsymbol{\Sigma}_{B,B}^{-1}(\mathbf{z}_B - \boldsymbol{\mu}_B),\ \boldsymbol{\Sigma}_{A,A} - \boldsymbol{\Sigma}_{A,B}\boldsymbol{\Sigma}_{B,B}^{-1}\boldsymbol{\Sigma}_{B,A})$$

Proof. Let us write $\mathbf{Z}_A = \mathbf{M}\mathbf{Z}_B + (\mathbf{Z}_A - \mathbf{M}\mathbf{Z}_B)$ with matrix $\mathbf{M} \in \mathbb{R}^{|A|\times|B|}$ such that $\mathbf{Z}_A - \mathbf{M}\mathbf{Z}_B$ and $\mathbf{Z}_B$ are independent, i.e. such that

$$\text{Cov}(\mathbf{Z}_B,\ \mathbf{Z}_A - \mathbf{M}\mathbf{Z}_B) = \boldsymbol{\Sigma}_{B,A} - \boldsymbol{\Sigma}_{B,B}\mathbf{M}^T = \mathbf{0}.$$

This occurs when we take $\mathbf{M}^T = \boldsymbol{\Sigma}_{B,B}^{-1}\boldsymbol{\Sigma}_{B,A}$. Because $\mathbf{Z}_A - \mathbf{M}\mathbf{Z}_B$ and $\mathbf{Z}_B$ are independent, the distribution of $\mathbf{Z}_A - \mathbf{M}\mathbf{Z}_B$ conditional on $\mathbf{Z}_B = \mathbf{z}_B$ is equal to its unconditional distribution. Now

$$\begin{aligned}
\mathbb{E}(\mathbf{Z}_A - \mathbf{M}\mathbf{Z}_B) &= \boldsymbol{\mu}_A - \boldsymbol{\Sigma}_{A,B}\boldsymbol{\Sigma}_{B,B}^{-1}\boldsymbol{\mu}_B \\
\text{Var}(\mathbf{Z}_A - \mathbf{M}\mathbf{Z}_B) &= \boldsymbol{\Sigma}_{A,A} + \boldsymbol{\Sigma}_{A,B}\boldsymbol{\Sigma}_{B,B}^{-1}\boldsymbol{\Sigma}_{B,B}\boldsymbol{\Sigma}_{B,B}^{-1}\boldsymbol{\Sigma}_{B,A} - 2\boldsymbol{\Sigma}_{A,B}\boldsymbol{\Sigma}_{B,B}^{-1}\boldsymbol{\Sigma}_{B,A} \\
&= \boldsymbol{\Sigma}_{A,A} - \boldsymbol{\Sigma}_{A,B}\boldsymbol{\Sigma}_{B,B}^{-1}\boldsymbol{\Sigma}_{B,A}.
\end{aligned}$$

Since $\mathbf{M}\mathbf{Z}_B$ is a function of $\mathbf{Z}_B$, conditional on $\mathbf{Z}_B = \mathbf{z}_B$, it equals $\mathbf{M}\mathbf{z}_B$. Then as $\mathbf{Z}_A - \mathbf{M}\mathbf{Z}_B$ is normally distributed, we have the result. □