THE COORDINATE-FREE APPROACH TO LINEAR MODELS

MICHAEL J. WICHURA

The University of Chicago
Chicago, Illinois

In memoriam

William H. Kruskal
1919–2005

PREFACE

When I was a graduate student in the mid 1960's I finally came to understand the mathematical theory underlying analysis of variance and regression after reading a draft of William Kruskal's monograph on the so-called coordinate-free, or geometric, approach to these subjects. Alas, with Kruskal's demise, this excellent treatise will never be published.

From time to time during the 1970's, 80's and early 90's, I had the good fortune of teaching the coordinate-free approach to linear models. While doing so, I evolved my own set of lecture notes, presented here. With regard to inspiration and content, my debt to Kruskal is clear. However, my notes are intended for a one, rather than three, quarter course, and are aimed at Statistics graduate students who are already fairly well versed in linear algebra. I have also included some of the highlights of the optimality theory for estimation and testing in linear models under the assumption of normality, feeling that the elegant setting provided by the coordinate-free approach is a natural one in which to place these jewels of mathematical statistics. Out of deference to Kruskal, who was my colleague here at the University of Chicago, I have not until now made my notes public. My hope is that readers will find the presentation both instructive and enjoyable.

Various graduate students, in particular Neal Thomas and Nathaniel Schenker have made many comments which have greatly improved these notes. My thanks go to all of them, and also to David Van Dyke and Peter Meyer for suggesting how easy/hard each of the some 200 exercises are. Thanks are also due to Mitzi Nakatsuka for her help in converting the notes to TeX, and to Persi Diaconis for his advice and encouragement.

Michael J. Wichura

University of Chicago
July 2005

CONTENTS

INTRODUCTION

CHAPTER 1

INTRODUCTION

In this chapter we introduce and contrast the matricial and geometric formulations of the so-called general linear model and describe the organization of the rest of these notes.

1. Orientation

Recall the classical framework of the *general linear model (GLM).* One is given an n-dimensional random vector $\boldsymbol{Y}^{n\times 1} = (Y_1 \dots, Y_n)^T$, perhaps multivariate normally distributed, with covariance matrix $(\mathrm{Cov}(Y_i, Y_j))^{n\times n} = \sigma^2 \boldsymbol{I}^{n\times n}$ and mean vector $\boldsymbol{\mu}^{n\times 1} = E(\boldsymbol{Y}) = (EY_1, \dots, EY_n)^T$ of the form

$$\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta},$$

where $\boldsymbol{X}^{n\times p}$ is known and σ^2 and $\boldsymbol{\beta}^{p\times 1} = (\beta_1, \dots, \beta_p)^T$ are unknown; in addition the β_i's may be subject to linear constraints $\boldsymbol{R}\boldsymbol{\beta} = 0$, where $\boldsymbol{R}^{c\times p}$ is known. $\boldsymbol{X}$ is called the *design,* or *regression, matrix,* and $\boldsymbol{\beta}$ is called the *parameter vector.*

1.1 Example. In the classical *two-sample problem,* one has

$$\boldsymbol{X}^T = \begin{pmatrix} \underbrace{\begin{matrix} 1 & 1 & \dots & 1 \\ 0 & 0 & \dots & 0 \end{matrix}}_{n_1 \text{ times}} & \underbrace{\begin{matrix} 0 & 0 & \dots & 0 \\ 1 & 1 & \dots & 1 \end{matrix}}_{n_2 \text{ times}} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix},$$

i.e.,

$$E(Y_i) = \begin{cases} \mu_1, & \text{if } 1 \le i \le n_1, \\ \mu_2, & \text{if } n_1 < i \le n_1 + n_2 = n \end{cases}.$$

•

1.2 Example. In *simple linear regression,* one has

$$\boldsymbol{X}^T = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} a \\ b \end{pmatrix},$$

i.e.,

$$E(Y_i) = a + bx_i \quad \text{for } i = 1, \dots, n.$$

•

Typical problems are the estimation of linear combinations of the β_i's, testing that some such linear combinations are 0 (or some other prescribed value), and estimation of σ^2.

1.3 Example. In the two-sample problem, one is often interested in estimating the difference $\mu_2 - \mu_1$ or in testing the null hypothesis that $\mu_1 = \mu_2$. •

1.4 Example. In simple linear regression, one seeks estimates of the intercept a and slope b and may want to test, e.g., the hypothesis that $b = 0$, or the hypothesis that $b = 1$. •

If you've had some prior statistical training, you may well have already encountered the resolution of these problems. You may know, for example, that provided $\boldsymbol{X}$ is of full rank and no linear constraints are imposed on $\boldsymbol{\beta}$, the *best (minimum variance) linear unbiased estimator (BLUE)* of $\sum_{1\le i\le p} c_i\beta_i$ is $\sum_{1\le i\le p} c_i\hat{\beta}_i$, where

$$(\hat{\beta}_1, \ldots, \hat{\beta}_p)^T = \boldsymbol{C}\boldsymbol{X}^T\boldsymbol{Y}, \quad \text{with} \quad \boldsymbol{C} = \boldsymbol{A}^{-1}, \quad \boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{X};$$

this is called the *Gauss-Markov theorem.*

In these notes we will be studying the GLM from a geometric point of view, using linear algebra in place of matrix algebra. Although we will not reach any conclusions that could not be obtained solely by matrix techniques, the basic ideas will emerge more clearly. With the added intuitive feeling and mathematical insight this provides, one will be better able to understand old results and formulate and prove new ones.

From a geometric perspective, the GLM may be described as follows, using some terms which will be defined in subsequent chapters. One is given a *random vector* Y taking values in some given *inner product space* $(V, \langle\cdot,\cdot\rangle)$. It is assumed that Y has a *weakly spherical covariance operator* and that the *mean* μ of Y lies in a given *manifold* M of V; for purposes of testing it is further assumed that Y is *normally distributed.* One desires to estimate μ (or *linear functionals* of μ) and to test hypotheses such as $\mu \in M_0$, where M_0 is a given *submanifold* of M. The Gauss-Markov theorem says that the BLUE of the linear functional $\psi(\mu)$ is $\psi(\hat{\mu})$, where $\hat{\mu}$ is the *orthogonal projection* of Y onto M. As we will see, this geometric description of the problem encompasses the matricial formulation of the GLM not only as it is set out above (take, e.g., $V = \mathbb{R}^n$, $\langle\cdot,\cdot\rangle$ = dot-product, $Y = \boldsymbol{Y}$, $\mu = \boldsymbol{\mu}$, and M = the subspace of $\mathbb{R}^n$ spanned by the columns of the design matrix $\boldsymbol{X}$), but also in cases where $\boldsymbol{X}$ is of less than full rank and/or linear constraints are imposed on the β_i's.

2. An illustrative example

To illustrate the differences between the matricial and geometric approaches, let's compare the ways in which one establishes the independence of

$$\bar{Y} = \hat{\mu} = \frac{\sum_{1\le i\le n} Y_i}{n} \quad \text{and} \quad s^2 = \hat{\sigma}^2 = \frac{1}{n-1} \sum_{1\le i\le n} (Y_i - \bar{Y})^2$$

in the one-sample problem

$$\boldsymbol{Y}^{n\times 1} \sim N(\mu\boldsymbol{e}, \sigma^2 \boldsymbol{I}^{n\times n}) \quad \text{with} \quad \boldsymbol{e} = (1, 1, \ldots, 1)^T. \tag{2.1}$$

(The vector $\boldsymbol{e}$ is called the *equiangular vector.*)

The classical matrix proof, which uses some facts about multivariate normal distributions, runs like this. Let $\boldsymbol{B}^{n\times n} = (b_{ij})$ be the matrix

$$\begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 & 0 & \cdots & 0 & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & 0 & \cdots & 0 & 0 \\ \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{-3}{\sqrt{12}} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdots & \frac{1}{\sqrt{n(n-1)}} & \frac{-(n-1)}{\sqrt{n(n-1)}} \end{pmatrix}.$$

Note that the rows (and columns) of $\boldsymbol{B}$ are orthonormal $\left(\sum_{1\le k\le n} b_{ik}b_{jk} = \delta_{ij} \equiv \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \ne j \end{cases}\right)$ and that the first row is $\frac{1}{\sqrt{n}}\,\boldsymbol{e}^T$. Set

$$\boldsymbol{Z} = \boldsymbol{BY}.$$

Then

$$\boldsymbol{Z} \sim N(\boldsymbol{\nu}, \boldsymbol{\Sigma})$$

with

$$\boldsymbol{\nu} = \boldsymbol{B}(\mu\boldsymbol{e}) = \mu\boldsymbol{Be} = (\sqrt{n}\,\mu, 0, \ldots, 0)^T$$

and

$$\boldsymbol{\Sigma} = \boldsymbol{B}(\sigma^2\boldsymbol{I})\boldsymbol{B}^T = \sigma^2\boldsymbol{BB}^T = \sigma^2\boldsymbol{I};$$

that is, $Z_1, Z_2, \ldots, Z_n$ are independent normal random variables, each with variance σ^2, $E(Z_1) = \sqrt{n}\,\mu$, and $E(Z_j) = 0$ for $2 \le j \le n$. Moreover

$$Z_1 = \frac{1}{\sqrt{n}} \sum_{1\le i\le n} Y_i = \sqrt{n}\,\bar{Y}, \text{ or } \bar{Y} = \frac{Z_1}{\sqrt{n}},$$

while

$$\begin{aligned} (n-1)s^2 &= \sum_{1\le i\le n} (Y_i - \bar{Y})^2 = \sum_{1\le i\le n} Y_i^2 - n\bar{Y}^2 \\ &= \sum_{1\le i\le n} Y_i^2 - Z_1^2 = \sum_{1\le i\le n} Z_i^2 - Z_1^2 = \sum_{2\le i\le n} Z_i^2 \end{aligned} \tag{2.2}$$

because

$$\sum_{1\le i\le n} Z_i^2 = \boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{Y}^T\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{Y} = \sum_{1\le i\le n} Y_i^2.$$

This gives the independence of $\bar{Y}$ and s^2 and it's an easy step to get the marginal distributions: $\bar{Y} \sim N(\mu, \sigma^2/n)$ and $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$.

What is the nature of the transformation $\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{Y}$? Let $\boldsymbol{b}_1 = \boldsymbol{e}/\sqrt{n}$, $\boldsymbol{b}_2, \ldots, \boldsymbol{b}_n$ denote the transposes of the rows of $\boldsymbol{B}$. The coordinates of $\boldsymbol{Y} = \sum_{1\le j\le n} C_j \boldsymbol{b}_j$ with respect to this new orthonormal basis for $\mathbb{R}^n$ are given by

$$C_i = \boldsymbol{b}_i^T\boldsymbol{Y} = Z_i, \quad i = 1, \ldots, n.$$

The effect of the change of coordinates $\boldsymbol{Y} \to \boldsymbol{Z}$ is to split $\boldsymbol{Y}$ into its components along, and orthogonal to, the equiangular vector $\boldsymbol{e}$.

Now I'll show you the geometric proof, which uses some properties of (weakly) spherical normal random vectors taking values in an inner product space $(V, \langle\cdot,\cdot\rangle)$, here $(\mathbb{R}^n$, dot-product). The assumptions imply that $\boldsymbol{Y}$ is spherical normally distributed about its mean $E(\boldsymbol{Y})$ and that $E(\boldsymbol{Y})$ lies in the manifold M spanned by $\boldsymbol{e}$. Let P_M denote orthogonal projection onto M, Q_M orthogonal projection onto the orthogonal complement $M^\perp$ of M. Basic distribution theory says that $P_M\boldsymbol{Y}$ and $Q_M\boldsymbol{Y}$ are independent. But

$$P_M\boldsymbol{Y} = \frac{\langle \boldsymbol{e}, \boldsymbol{Y}\rangle}{\langle \boldsymbol{e}, \boldsymbol{e}\rangle}\,\boldsymbol{e} = \bar{Y}\boldsymbol{e} \tag{2.3}$$

and

$$Q_M\boldsymbol{Y} = \boldsymbol{Y} - P_M\boldsymbol{Y} = \boldsymbol{Y} - \bar{Y}\boldsymbol{e} = (Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y})^T;$$

it follows that $\bar{Y}$ and $(n-1)s^2 = \sum_{1\le i\le n}(Y_i - \bar{Y})^2 = \|Q_M\boldsymbol{Y}\|^2$ are independent. Again it is an easy matter to get the marginal distributions.

To my way of thinking, granted the technical apparatus, the second proof is clearer, being more to the point. The first proof does the same things, but (to the uninitiated) in an obscure manner.

3. Overview

Chapter 2 covers fundamental concepts from linear algebra, such as the notion of orthogonal projection. Basic distribution theory for random vectors taking values in inner product spaces is developed in Chapter 3. The "geometric" version of the Gauss-Markov theorem is discussed in Chapter 4, and optimal properties of Gauss-Markov estimation under the assumption of normality are considered in Chapter 5. F-testing of null hypotheses and the related issue of interval estimation are taken up in Chapter 6. Chapters 7 and 8 deal with the analysis of covariance and missing observations respectively.

From the perspective of mathematical statistics, there are some very elegant results, and some notable surprises, connected with the optimality theory for Gauss-Markov estimation / F-testing under the assumption of normality. In the sections that deal with these matters — in particular Sections 5.4, 5.5, 6.3, 6.4, and 6.6 — the mathematics is somewhat harder than elsewhere, corresponding to the greater depth of the theory.

These notes are aimed at students who have already had some exposure to the matricial formulation of the GLM, perhaps through a methodology course, and are interested in the underlying theory. Each of the following chapters contains numerous exercises, along with a problem set which develops some topic complementing the material in that chapter. Most of the exercises are easy but, I hope, instructive. I typically devote a substantial amount of class time to having students present solutions to the exercises. Some exercises foreshadow what's to come, by covering a special case of material that will be presented in full generality later on. Moreover, the assertions of some exercises are appealed to later on the text. If you are working through the book on your own, you should at least read over each exercise, even if you don't work things out. Each exercise is assigned a difficulty level using the syntax "Exercise [d]", where d is an integer in the range 1 to 5 — the larger is d, the harder the exercise. The value of d depends both on the intrinsic difficulty of the exercise and the length of time needed to write up the solution. The problem sets are somewhat harder than the exercises and require a sustained effort for their completion.

Chapters are organized into sections. Within each section of the current chapter, enumerated items are numbered consecutively in the form (*section number*.*item number*). References to items in a different chapter take the expanded form (*chapter number*.*section number*.*item number*). For example, (2.4) refers to the 4^{th} numbered item (which may be an example, exercise, theorem, formula, or whatever) in the 2^{nd} section of the current chapter, while (6.1.3) refers to the 3^{rd} numbered item in the 1^{st} section of the 6^{th} chapter.

The end of: a proof is marked by a ▪; of an example, by a •; of an exercise, by a ⋄; of a part of problem set, by a ∘.

To help distinguish between the matricial and geometric points of view, matrices, including row and column vectors, are written in ***italic boldface*** type while linear transformations and elements of abstract vector spaces are written simply in *italic* type. We speak, for example, of the design matrix $\boldsymbol{X}$, but of vectors v and w in an inner product space V.

CHAPTER 2

TOPICS IN LINEAR ALGEBRA

In this chapter we discuss some topics from linear algebra which play a central role in the geometrical analysis of the GLM. The notion of orthogonal projection in an inner product space is introduced in Section 2.1 and studied in Section 2.2. A class of orthogonal decompositions that are useful in the design of experiments is studied in Section 2.3. The spectral representation of self-adjoint transformations is developed in Section 2.4. Linear and bilinear functionals are discussed in Section 2.5. The chapter closes with a problem set in Section 2.6, followed by an Appendix containing a brief review of the basic definitions and facts from linear algebra with which we presume the reader is already familiar.

1. Orthogonal projections

Throughout these notes we operate in the context of a finite-dimensional *inner product space* $(V, \langle\cdot,\cdot\rangle)$ — V is a finite-dimensional vector space and $\langle\cdot,\cdot\rangle\colon V \times V \to \mathbb{R}$ an *inner product:*

(i) *(positive definiteness)* $\langle x, x\rangle \geq 0$ for all $x \in V$ and $\langle x, x\rangle = 0$ if and only if $x = 0$,

(ii) *(symmetry)* $\langle x, y\rangle = \langle y, x\rangle$ for all $x, y \in V$,

(iii) *(bilinearity)* For all $c_1, c_2 \in \mathbb{R}$ and $x_1, x_2, x, y_1, y_2, y \in V$, one has

$$\langle c_1x_1 + c_2x_2, y\rangle = c_1\langle x_1, y\rangle + c_2\langle x_2, y\rangle$$
$$\langle x, c_1y_1 + c_2y_2\rangle = c_1\langle x, y_1\rangle + c_2\langle x, y_2\rangle.$$

The canonical example is $V = \mathbb{R}^n$ endowed with the dot-product:

$$\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \boldsymbol{x} \cdot \boldsymbol{y} = \sum\nolimits_{1\leq i\leq n} x_iy_i = \boldsymbol{x}^T\boldsymbol{y}$$

for $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ and $\boldsymbol{y} = (y_1, \ldots, y_n)^T$. Unless specifically stated to the contrary, we always view $\mathbb{R}^n$ as endowed with the dot-product.

Two vectors x and y in V are said to be *perpendicular,* or *orthogonal* (with respect to $\langle\cdot,\cdot\rangle$), if

$$\langle x, y\rangle = 0;$$

one writes

$$x \perp y.$$

The quantity

$$\|x\| = \sqrt{\langle x, x\rangle}$$

is called the *length*, or *norm* of x. The squared length of the sum of two vectors is given by

$$\begin{aligned}\|x+y\|^2 &= \langle x+y, x+y\rangle = \langle x, x+y\rangle + \langle y, x+y\rangle \\ &= \|x\|^2 + 2\langle x, y\rangle + \|y\|^2,\end{aligned}$$

which reduces to the *Pythagorean theorem*

$$\|x+y\|^2 = \|x\|^2 + \|y\|^2 \tag{1.1}$$

when $x \perp y$.

1.2 Exercise [1]. Let $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ be two non-zero vectors in $\mathbb{R}^2$ and let θ be the angle between them, measured counter-clockwise from v_1 to v_2. Show that

$$\cos(\theta) = \frac{\langle v_1, v_2\rangle}{\|v_1\|\,\|v_2\|}$$

and deduce that $\boldsymbol{v}_1 \perp \boldsymbol{v}_2$ if and only if $\theta = 90^\circ$ or 270°.
[Hint: Use the identity $\cos(\theta_2 - \theta_1) = \cos(\theta_1)\cos(\theta_2) + \sin(\theta_1)\sin(\theta_2)$.] ◇

1.3 Exercise [3]. Let $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$ be the vectors in $\mathbb{R}^n$ given by

$$\boldsymbol{x}_i = 1,\ \boldsymbol{y}_i = i - \tfrac{n+1}{2},\ \text{and } \boldsymbol{z}_i = \big(i - \tfrac{n+1}{2}\big)^2 - \tfrac{n^2-1}{12} \quad \text{for } 1 \le i \le n. \tag{1.4}$$

Show that $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ are mutually orthogonal and span the same subspace as do $(1,1,\ldots,1)^T$, $(1,2,\ldots,n)^T$, and $(1,4,\ldots,n^2)^T$. Exhibit $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ explicitly for $n = 5$. ◇

1.5 Exercise [1]. Let $\mathcal{O}: V \to V$ be a linear transformation. Show that $\mathcal{O}$ preserves lengths:

$$\|\mathcal{O}x\| = \|x\| \quad \text{for all } x \in V$$

if and only if it preserves inner products:

$$\langle \mathcal{O}x, \mathcal{O}y\rangle = \langle x, y\rangle \quad \text{for all } x, y \in V.$$

Such a transformation is said to be *orthogonal.*
[Hint: Observe

$$\langle x, y\rangle = \frac{\|x+y\|^2 - \|x\|^2 - \|y\|^2}{2} \tag{1.6}$$

for $x, y \in V$.] ◇

1.7 Exercise [2]. (1) Let v_1 and v_2 be elements of V. Show that $v_1 = v_2$ if and only if $\langle v_1, w\rangle = \langle v_2, w\rangle$ for each $w \in V$, or just for each w in some basis for V. (2) Let T_1 and T_2 be two linear transformations of V. Show that $T_1 = T_2$ if and only if $\langle v, T_1 w\rangle = \langle v, T_2 w\rangle$ for each v and w in V, or just for each v and w in some basis for V. ◇

As intimated in the introduction, the notion of orthogonal projection onto subspaces of V plays a key role in the study of the GLM. We begin our study of projections with the following seminal result.

1.8 Theorem. *Suppose that M is a subspace of V and that $x \in V$. There is exactly one vector $m \in M$ such that the residual $x - m$ is orthogonal to M:*

$$(x - m) \perp y \quad \text{for all } y \in M, \tag{1.9}$$

or, equivalently, such that m is closest to x:

$$\|x - m\| = \inf\{\, \|x - y\| : y \in M \,\}. \tag{1.10}$$

The proof will be given shortly. The unique $m \in M$ such that (1.9) and (1.10) hold is called the *orthogonal projection of x onto M*, written $P_M x$, and the mapping P_M which sends $x \in V$ to $P_M x$ is called *orthogonal projection onto M*. In the context of $\mathbb{R}^n$ with $\boldsymbol{x} = (x_i)$ and $\boldsymbol{m} = (m_i)$, (1.9) reads

$$\sum\nolimits_{1\le i\le n} (x_i - m_i) y_i = 0 \quad \text{for all } \boldsymbol{y} = (y_i) \in M,$$

while (1.10) is the *least squares characterization* of $\boldsymbol{m}$:

$$\sum\nolimits_{1\le i\le n} (x_i - m_i)^2 = \inf\Big\{ \sum\nolimits_{1\le i\le n} (x_i - y_i)^2 : \boldsymbol{y} \in M \Big\}.$$

1.11 Exercise [1]. Let M be a subspace of V. (a) Show that

$$\|x - P_M x\|^2 = \|x\|^2 - \|P_M x\|^2 \tag{1.12}$$

for all $x \in V$. (b) Deduce that

$$\|P_M x\| \le \|x\| \tag{1.13}$$

for all $x \in V$, with equality holding if and only if $x \in M$. ◇

Proof of Theorem 1.8. (1.9) *implies* (1.10): Suppose $m \in M$ satisfies (1.9). Then for all $y \in M$ the Pythagorean theorem gives

$$\|x - y\|^2 = \|(x - m) + (m - y)\|^2 = \|x - m\|^2 + \|m - y\|^2,$$

so m satisfies (1.10).

(1.10) *implies* (1.9): Suppose $m \in M$ satisfies (1.10). Then for any $0 \ne y \in M$ and any $\delta \in \mathbb{R}$,

$$\|x - m\|^2 \le \|x - m + \delta y\|^2 = \|x - m\|^2 + 2\delta\langle x - m, y\rangle + \delta^2\|y\|^2;$$

this relation forces $\langle x - m, y\rangle = 0$.

Uniqueness: If

$$m_1 + y_1 = x = m_2 + y_2$$

with $m_i \in M$ and $y_i \perp M$ for $i = 1$ and 2, then the vector

$$m_1 - m_2 = y_2 - y_1$$

lies in M and is perpendicular to M, and so is perpendicular to itself:

$$0 = \langle m_1 - m_2, m_1 - m_2 \rangle = \|m_1 - m_2\|^2,$$

whence $m_1 = m_2$ by the positive-definiteness of $\langle \cdot, \cdot \rangle$.

Existence: We will show in a moment that M has a basis $m_1, \ldots, m_k$ consisting of mutually orthogonal vectors. For any such basis the generic $m = \sum_{1 \le j \le k} c_j m_j$ in M satisfies

$$(x - m) \perp M$$

if and only if $x - m$ is orthogonal to each m_i, i.e., if and only if

$$\langle m_i, x \rangle = \langle m_i, \sum\nolimits_{1 \le j \le k} c_j m_j \rangle = \sum\nolimits_{1 \le j \le k} \langle m_i, m_j \rangle c_j = \langle m_i, m_i \rangle c_i$$

for $1 \le i \le k$. It follows that we can take

$$P_M x = \sum\nolimits_{1 \le i \le k} \frac{\langle m_i, x \rangle}{\langle m_i, m_i \rangle} \, m_i. \tag{1.14}$$

To produce an orthogonal basis for M, let $m_1^*, \ldots, m_k^*$ be any basis for M and inductively define new basis vectors $m_1, \ldots, m_k$ by the recipe $m_1 = m_1^*$ and

$$\begin{aligned} m_j &= m_j^* - P_{[m_1^*, \ldots, m_{j-1}^*]} m_j^* = m_j^* - P_{[m_1, \ldots, m_{j-1}]} m_j^* \\ &= m_j^* - \sum\nolimits_{1 \le i \le j-1} \frac{\langle m_j^*, m_i \rangle}{\langle m_i, m_i \rangle} \, m_i \end{aligned} \tag{1.15}$$

for $j = 2, \ldots, k$; here $[m_1^*, \ldots, m_{j-1}^*]$ denotes the span of $m_1^*, \ldots, m_{j-1}^*$ and $[m_1, \ldots, m_{j-1}]$ denotes the (identical) span of $m_1, \ldots, m_{j-1}$ (see Subsection 2.6.1). ∎

The recursive scheme for cranking out the m_j's above is called *Gram-Schmidt orthogonalization.* As a special case of (1.14) we have the following simple, yet key, formula for projecting onto a one-dimensional space:

$$P_{[m]} x = \frac{\langle x, m \rangle}{\langle m, m \rangle} \, m \qquad \text{for } m \ne 0. \tag{1.16}$$

1.17 Example. In the context of $\mathbb{R}^n$ take $\boldsymbol{m} = \boldsymbol{e} \equiv (1, \ldots, 1)^T$. Formula (1.16) then reads

$$P_{[\boldsymbol{e}]} \boldsymbol{x} = \frac{\langle \boldsymbol{x}, \boldsymbol{e} \rangle}{\langle \boldsymbol{e}, \boldsymbol{e} \rangle} \, \boldsymbol{e} = \frac{\sum_i x_i}{\sum_i 1} \, \boldsymbol{e} = \bar{x} \boldsymbol{e} = (\bar{x}, \ldots, \bar{x})^T;$$

we used this result in the introduction (see (1.2.3)). Formula (1.12) reads

$$\sum\nolimits_i (x_i - \bar{x})^2 = \|\boldsymbol{x} - \bar{x}\boldsymbol{e}\|^2 = \|\boldsymbol{x}\|^2 - \|\bar{x}\boldsymbol{e}\|^2 = \sum\nolimits_i x_i^2 - n\bar{x}^2;$$

this is just the computing formula for $(n-1)s^2$ used at (1.2.2). According to (1.10), $c = \bar{x}$ minimizes the sum of squares $\sum_i (x_i - c)^2$. •

1.18 Exercise [2]. Use the preceding techniques to compute $P_M\boldsymbol{y}$ for $\boldsymbol{y} \in \mathbb{R}^n$ and $M = [\boldsymbol{e}, (x_1, \ldots, x_n)^T]$, the manifold spanned by the columns of the design matrix for simple linear regression. ◇

1.19 Exercise [3]. Show that the transposes of the rows of the matrix $\boldsymbol{B}$ of Section 1.2 result from first applying the Gram-Schmidt orthogonalization scheme to the vectors $\boldsymbol{e}$, $(1, -1, 0, \ldots, 0)^T$, $(0, 1, -1, 0, \ldots, 0)^T$, $\ldots$, $(0, \ldots, 0, 1, -1)^T$ in $\mathbb{R}^n$ and then normalizing to unit length. ◇

1.20 Exercise [2]. Suppose $x, y \in V$. Prove the *Cauchy-Schwarz inequality*:

$$|\langle x, y\rangle| \le \|x\|\,\|y\|, \tag{1.21}$$

with equality if and only if x and y are linearly dependent. Deduce *Minkowski's inequality*:

$$\|x + y\| \le \|x\| + \|y\|. \tag{1.22}$$

[Hint: For (1.21), take $m = y$ in (1.16) and use part (b) of Exercise 1.11.] ◇

1.23 Exercise [4]. Define $d: V \times V \to \mathbb{R}$ by $d(x, y) = \|y - x\|$. Show that d is a metric on V such that the set $\{\, x \in V : \|x\| = 1 \,\}$ is compact. ◇

1.24 Exercise [3]. Show that for any two subspaces M and N of V,

$$\sup\{\, \|P_M x\| : x \in N \text{ and } \|x\| = 1 \,\} \le 1, \tag{1.25}$$

with equality holding if and only if M and N have a non-zero vector in common. [Hint: A continuous real-valued function on a compact set attains its maximum.] ◇

To close this section we generalize (1.14) to cover the case of an arbitrary basis for M. Suppose then that the basis vectors $m_1, \ldots, m_k$ are not necessarily orthogonal and let $x \in V$. As in the derivation of (1.14),

$$P_M x = \sum\nolimits_j c_j m_j$$

is determined by the condition

$$m_i \perp (x - P_M x) \quad \text{for } i = 1, \ldots, k,$$

i.e., by the so-called *normal equations*

$$\sum\nolimits_{1 \le j \le k} \langle m_i, m_j\rangle c_j = \langle m_i, x\rangle, \quad i = 1, \ldots, k. \tag{1.26}$$

In matrix notation (1.26) reads

$$\langle m, m\rangle \boldsymbol{c} = \langle m, x\rangle$$

so that

$$\boldsymbol{c} = \langle m, m \rangle^{-1} \langle m, x \rangle,$$

the $k \times k$ matrix $\langle m, m \rangle$ and the $k \times 1$ column vectors $\langle m, x \rangle$ and $\boldsymbol{c}$ being given by

$$\big(\langle m, m \rangle\big)_{ij} = \langle m_i, m_j \rangle, \quad \big(\langle m, x \rangle\big)_i = \langle m_i, x \rangle, \quad \text{and} \quad (\boldsymbol{c})_i = c_i$$

for $i, j = 1, \ldots, k$. If the m_i's are mutually orthogonal, $\langle m, m \rangle$ is diagonal and (1.14) follows immediately.

1.27 Exercise [3]. Show that the matrix $\langle m, m \rangle$ above is in fact nonsingular, so that the inverse $\langle m, m \rangle^{-1}$ exists.
[Hint: Show that $\sum_j \langle m_i, m_j \rangle c_j = 0$ for each i implies $c_1 = \cdots = c_k = 0$.] ◇

1.28 Exercise [2]. Redo Exercise 1.18 using (1.26). Check that the two ways of calculating $P_M y$ do in fact give the same result. ◇

1.29 Exercise [2]. Let M be the subspace of $\mathbb{R}^n$ spanned by the columns of an $n \times k$ matrix $\boldsymbol{X}$. Supposing these columns to be linearly independent, use (1.26) to show that the matrix representing P_M with respect to the usual coordinate basis of $\mathbb{R}^n$ is $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Exhibit this matrix explicitly in the case $\boldsymbol{X} = \boldsymbol{e}$. ◇

1.30 Exercise [4]. Suppose $X_0, X_1, \ldots, X_k$ are square integrable random variables defined on a common probability space. Consider using $X_1, \ldots, X_k$ to predict X_0. Show that among predictors of the form

$$\hat{X}_0 = c_0 + \sum\nolimits_{1 \le i \le k} c_i X_i,$$

the c's being constants, the one minimizing the *mean square error of prediction*

$$E(\hat{X}_0 - X_0)^2$$

is

$$\hat{X}_0 = \mu_0 + \sum\nolimits_{1 \le i \le k} \Big(\sum\nolimits_{1 \le j \le k} \sigma^{ij} \sigma_{j0} \Big)(X_i - \mu_i),$$

where

$$\begin{aligned} \mu_i &= E(X_i), & 0 \le i \le k, \\ \sigma_{ij} &= \operatorname{Cov}(X_i, X_j), & 0 \le i, j \le k, \end{aligned}$$

and the $k \times k$ matrix $(\sigma^{ij})_{1 \le i,j \le k}$ is the inverse of the matrix $(\sigma_{ij})_{1 \le i,j \le k}$, the latter assumed to be nonsingular.
[Hint: This is just a matter of projecting $X_0 - \mu_0$ onto the subspace spanned by 1, $X_1 - \mu_1, \ldots, X_k - \mu_k$ in the space $\mathcal{L}_2$ of square integrable random variables on the given probability space, the inner product between two variables A and B being $E(AB)$.] ◇

2. Properties of orthogonal projections

This section develops properties of orthogonal projections. The results will prove to be useful in analyzing the GLM.

Let $(V, \langle\cdot,\cdot\rangle)$ be an inner product space, let M be a subspace of V, and let

$$M^\perp = \{\, x \in V : x \perp M \,\} \equiv \{\, x \in V : x \perp y \text{ for all } y \in M \,\} \tag{2.1}$$

be the *orthogonal complement* of M. Note that $M^\perp$ is a subspace of V. Let P_M and Q_M denote orthogonal projection onto M and $M^\perp$ respectively and let $x \in V$. By Theorem 1.8,

$$x = P_M x + Q_M x$$

is the unique representation of x as the sum of an element of M and an element of $M^\perp$.

2.2 Exercise [2]. Let M and $M_1, \ldots, M_k$ be subspaces of V. Show that

$$\begin{aligned} d(M^\perp) &= d(V) - d(M), \\ (M^\perp)^\perp &= M, \\ M_1 \subset M_2 &\iff M_1^\perp \supset M_2^\perp, \\ \Bigl(\sum\nolimits_{1\le i\le k} M_i\Bigr)^\perp &= \bigcap\nolimits_{1\le i\le k} M_i^\perp, \\ \Bigl(\bigcap\nolimits_{1\le i\le k} M_i\Bigr)^\perp &= \sum\nolimits_{1\le i\le k} M_i^\perp. \end{aligned}$$

◇

2A. Characterization of orthogonal projections

A linear transformation $T\colon V \to V$ is said to be *idempotent* if $T^2 = T$ and said to be *self-adjoint* if

$$\langle Tx, y\rangle = \langle x, Ty\rangle \tag{2.3}$$

for all x, y in V.

2.4 Proposition. (i) *P_M is an idempotent, self-adjoint, linear transformation with range M and null space $M^\perp$.*
(ii) *Conversely, an idempotent, self-adjoint, linear transformation T mapping V into V is the orthogonal projection onto its range.*

Proof. (i): Write P for P_M and Q for Q_M. The properties of P_M are all immediate consequences of the orthogonality of M and $M^\perp$ and the uniqueness of the decompositions

$$\begin{aligned} x &= Px + Qx \qquad (Px \in M, \quad Qx \in M^\perp) \\ y &= Py + Qy \qquad (Py \in M, \quad Qy \in M^\perp), \end{aligned}$$

to wit:

(a) P is linear, since $x + y = Px + Qx + Py + Qy$ implies $P(x + y) = Px + Py$ and $cx = cPx + cQx$ implies $P(cx) = c(Px)$;

(b) P is idempotent, since $Px = Px + 0$ implies $P^2x = Px$;

(c) P is self-adjoint, since $\langle Px, y\rangle = \langle Px, Py + Qy\rangle = \langle Px, Py\rangle = \langle Px + Qx, Py\rangle = \langle x, Py\rangle$ for all $x, y \in V$;

(d) M is the range of P, since on the one hand $x \in V$ implies $Px \in M$ and on the other hand $x \in M$ implies $x = Px$;

(e) $M^\perp$ is the null space of P, since $x \in M^\perp \iff x = Qx \iff Px = 0$.

(ii): For $x \in V$ write

$$x = Tx + (x - Tx).$$

Since $Tx \in \mathcal{R}(T)$, we need only show $x - Tx \in \big(\mathcal{R}(T)\big)^\perp$. For this observe that for all $y \in V$, one has $\langle x - Tx, Ty\rangle = \langle T(x - Tx), y\rangle = \langle Tx - T^2x, y\rangle = \langle Tx - Tx, y\rangle = 0$. ∎

2.5 Exercise [2]. Given $0 \neq m \in M$, show that the transformation

$$T: v \to \frac{\langle v, m\rangle}{\langle m, m\rangle}\, m$$

has all the properties required to characterize it as orthogonal projection onto $[m]$. ◇

2.6 Exercise [2]. Let M be a subspace of V. Show that the orthogonal linear transformation $T: V \to V$ which reflects each $x \in M$ through the origin and leaves each $x \in M^\perp$ fixed is $T = I - 2P_M$. ◇

2.7 Exercise [2]. Let $T: V \to V$ be self-adjoint. Show that

$$\big(\mathcal{R}(T)\big)^\perp = \mathcal{N}(T) \tag{2.8}$$

and deduce

$$\mathcal{R}(T^2) = \mathcal{R}(T). \tag{2.9}$$

[Hint: To get (2.9), first show $\mathcal{N}(T^2) \subset \mathcal{N}(T)$.] ◇

2B. Differences of orthogonal projections

If M and N are two subspaces of V with $M \subset N$, the subspace

$$N - M = \{\, x \in N : x \perp M \,\} = N \cap M^\perp \tag{2.10}$$

is called the *(relative) orthogonal complement* of M in N.

2.11 Exercise [2]. Let M and N be subspaces of V with $M \subset N$. Show that if the vectors $v_1, \ldots, v_k$ span N, then the vectors $Q_M v_1, \ldots, Q_M v_k$ span $N - M$. ◇

For self-adjoint transformations S and T mapping V to V, the notation

$$S \le T \tag{2.12}$$

means $\langle x, Sx\rangle \le \langle x, Tx\rangle$ for all $x \in V$.

2.13 Proposition. *Let M and N be subspaces of V. The following are equivalent:*

(i) $P_N - P_M$ *is an orthogonal projection,*

(ii) $M \subset N$,

(iii) $P_M \leq P_N$,

(iv) $\|P_N x\|^2 \geq \|P_M x\|^2$ *for all* $x \in V$,

(v) $P_M = P_M P_N$,

and

(vi) $P_N - P_M = P_{N-M}$.

Proof. (ii) *implies* (i) *and* (vi): Suppose $M \subset N$. We will show

$$P_N = P_M + P_{N-M}$$

so that (i) and (vi) hold. $P_M + P_{N-M}$ is a self-adjoint, linear transformation which is idempotent $\big((P_M + P_{N-M})^2 = P_M^2 + P_M P_{N-M} + P_{N-M} P_M + P_{N-M}^2 = P_M + 0 + 0 + P_{N-M}\big)$; its range is contained in N, and yet contains N because it contains both M and $N - M$. According to Proposition 2.4, these properties characterize $P_M + P_{N-M}$ as P_N.

(i) *implies* (iii): Write P for $P_N - P_M$. For all $x \in V$ one has

$$\begin{aligned}\langle P_N x, x\rangle - \langle P_M x, x\rangle &= \big\langle (P_N - P_M)x, x\big\rangle \\ &= \langle Px, x\rangle = \langle P^2 x, x\rangle = \langle Px, Px\rangle = \|Px\|^2 \geq 0.\end{aligned}$$

(iii) *implies* (iv): For all $x \in V$, $\|P_N x\|^2 = \langle P_N x, P_N x\rangle = \langle x, P_N x\rangle \geq \langle x, P_M x\rangle = \|P_M x\|^2$.

(iv) *implies* (v): For all $x \in V$,

$$\begin{aligned}\|P_M x\|^2 &= \|P_N(P_M x)\|^2 + \|Q_N(P_M x)\|^2 \\ &\geq \|P_N(P_M x)\|^2 \geq \|P_M(P_M x)\|^2 = \|P_M x\|^2;\end{aligned}$$

this implies $Q_N P_M = 0$. Thus for all $x, y \in V$,

$$\langle y, P_M Q_N x\rangle = \langle P_M y, Q_N x\rangle = \langle Q_N P_M y, x\rangle = 0$$

so $P_M Q_N = 0$ also. Now use

$$P_M = P_M(P_N + Q_N) = P_M P_N + P_M Q_N$$

to conclude $P_M = P_M P_N$.

(v) *implies* (ii): Since $P_M = P_M(P_N + Q_N) = P_M P_N + P_M Q_N$, $P_M = P_M P_N$ entails $P_M Q_N = 0 \Longrightarrow N^\perp \subset M^\perp \Longrightarrow M \subset N$.

(vi) *implies* (i): Trivial. ∎

2.14 Exercise [3]. Let p and $n_1, \dots, n_p$ be given and let $V \equiv \{\, (x_{ij})_{1\le j\le n_i, 1\le i\le p} : x_{ij} \in \mathbb{R} \,\}$ be endowed with the dot-product

$$\langle (x_{ij}), (y_{ij}) \rangle = \sum\nolimits_{1\le i\le p} \sum\nolimits_{1\le j\le n_i} x_{ij} y_{ij} \,.$$

Put

$$M = \{\, (x_{ij}) \in V : \text{for some } \beta_1,\, \dots,\, \beta_p \in \mathbb{R},\, x_{ij} = \beta_i \text{ for all } i,j \,\} = [\boldsymbol{v}_1, \dots, \boldsymbol{v}_p]$$

and

$$M_0 = \{\, (x_{ij}) \in V : \text{for some } \beta \in \mathbb{R},\, x_{ij} = \beta \text{ for all } i,j \,\} = [\sum\nolimits_{1\le i\le p} \boldsymbol{v}_i],$$

where $(\boldsymbol{v}_i)_{i'j'} = \delta_{ii'}$. Show that for each $x \in V$,

$$\begin{aligned} (P_M x)_{ij} &= \bar{x}_i \equiv \frac{1}{n_i} \sum\nolimits_{1\le j\le n_i} x_{ij} \,, \\ (P_{M_0} x)_{ij} &= \bar{x} \equiv \frac{1}{n} \sum\nolimits_i n_i \bar{x}_{i\cdot} = \frac{1}{n} \sum\nolimits_{ij} x_{ij} \end{aligned} \tag{2.15}$$

with $n = \sum_i n_i$; deduce that

$$(P_{M-M_0} x)_{ij} = \bar{x}_i - \bar{x} \,. \tag{2.16}$$

[Hint: $v_1,\, \dots,\, v_p$ is an orthogonal basis for M.] ◇

2.17 Exercise [1]. Let M and N be subspaces of an inner product space V. Put $L = M \cap N$. Show that $P_{M-L} P_{N-L} = P_M P_N - P_L$. ◇

2.18 Exercise [2]. Suppose L and M are subspaces of V. Put

$$K = P_L(M) \equiv \{\, P_L m : m \in M \,\}.$$

Show that

$$P_{V-M} - P_{L-K} = P_{(V-M)-(L-K)}. \tag{2.19}$$

Solve this exercise twice: (1) by arguing that $L - K \subset M^\perp$ (in fact, $L - K = L \cap M^\perp$); and (2) by arguing (independently of (1)) that $\|Q_M x\| \ge \|P_{L-K} x\|$ for each $x \in V$. ◇

2.20 Exercise [1]. Suppose $x_1, \dots, x_\ell$ is a basis for a subspace N of V. Let $M = [x_1, \dots, x_k]$ be the subspace spanned by the first k of the x_i's, where $1 \le k < \ell$. For an element v of V write

$$P_M v = \sum\nolimits_{i=1}^{k} a_i x_i \text{ and } P_N v = \sum\nolimits_{j=1}^{\ell} b_j x_j, \tag{2.21}$$

and for $j = k+1, \dots, \ell$, write

$$P_M x_j = \sum\nolimits_{i=1}^{k} c_{i,j} x_i; \tag{2.22}$$

the coefficients a_i, b_j, and $c_{i,j}$ in these sums are of course unique. Show that for $i = 1, \dots, k$,

$$a_i = b_i + \sum\nolimits_{j=k+1}^{\ell} c_{i,j} b_j. \tag{2.23}$$ ◇

2C. Sums of orthogonal projections

Subspaces M and N of V are said to be *orthogonal* (or *perpendicular*), written $M \perp N$, if $m \perp n$ for each $m \in M$ and $n \in N$.

2.24 Proposition. *Let P_1, P_2, …, P_k be orthogonal projections onto M_1, M_2, …, M_k respectively. Set $P = P_1 + P_2 + \cdots + P_k$. The following are equivalent:*

(i) *P is an orthogonal projection,*

(ii) *$P_i P_j = 0$ for all $i \neq j$,*

(iii) *the M_i's are mutually orthogonal,*

and

(iv) *P is orthogonal projection onto $\sum_i M_i$.*

Proof. (i) *implies* (ii): For each $x \in V$, one has

$$\|x\|^2 \geq \|Px\|^2 = \langle Px, x\rangle = \Big\langle \sum_i P_i x, x \Big\rangle = \sum_i \langle P_i x, x\rangle = \sum_i \|P_i x\|^2.$$

Taking $x = P_j y$ gives

$$\|P_j y\|^2 \geq \|P_j y\|^2 + \sum_{i \neq j} \|P_i P_j y\|^2.$$

As $y \in V$ is arbitrary, we get $P_i P_j = 0$ for $i \neq j$.

(ii) *implies* (iii): $P_i P_j = 0 \Longrightarrow M_j \subset M_i^\perp \Longrightarrow M_i \perp M_j$.

(iii) *implies* (i), (iv): For any two orthogonal subspaces K and L of V, part (vi) of Proposition 2.13 gives $P_{K+L} = P_K + P_L$. By induction

$$P = \sum_i P_i = P_{\sum_i M_i}.$$

(iv) *implies* (i): Trivial. ∎

The most important thing here is that (iii) implies (iv): given mutually orthogonal subspaces $M_1, \ldots, M_k$,

$$P_{\sum_i M_i} = \sum_i P_{M_i}. \tag{2.25}$$

Notice how this generalizes (1.14).

2.26 Exercise [3]. Let I and J be positive integers and let the space V of $I \times J$ matrices $(x_{ij})_{1 \leq i \leq I, 1 \leq j \leq J}$ be endowed with the dot-product

$$\langle (x_{ij}), (y_{ij}) \rangle = \sum_i \sum_j x_{ij} y_{ij}\,.$$

Let M_R be the subspace of row-wise constant matrices ($x_{ij} = x_{ij'}$, for all i, j, j'), M_C the subspace of column-wise constant matrices ($x_{ij} = x_{i'j}$, for all i, i', j), and M_G the subspace of constant matrices ($x_{ij} = x_{i'j'}$, for all i, i', j, j'). Show that

$$(P_{M_R}x)_{ij} = \bar{x}_{i\cdot} = \frac{1}{J}\sum_j x_{ij}\,, \tag{2.27$_R$}$$

$$(P_{M_C}x)_{ij} = \bar{x}_{\cdot j} \equiv \frac{1}{I}\sum_i x_{ij}\,, \tag{2.27$_C$}$$

$$(P_{M_G}x)_{ij} = \bar{x}_{\cdot\cdot} \equiv \frac{1}{IJ}\sum_{ij} x_{ij} \tag{2.27$_G$}$$

for each $x \in V$. Show further that

$$M_G + (M_R - M_G) + (M_C - M_G)$$

is an orthogonal decomposition of

$$M = M_R + M_C,$$

and deduce

$$(P_M x)_{ij} = \bar{x}_{i\cdot} + \bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot}\,. \tag{2.28}$$

[Hint: The $I \times J$ matrices $r_1, \ldots, r_I$ defined by $(r_i)_{i'j'} = \delta_{ii'}$ are an orthogonal basis for M_R.] ◇

2D. Products of orthogonal projections

Two subspaces M and N of V are said to be *book orthogonal*, written $M \perp_B N$, if $(M - L) \perp (N - L)$ with $L = M \cap N$. The imagery is that of two consecutive pages of a book which has been opened in such a way that the pages are at right angles to one another. Note that $M \perp_B N$ if $M \perp N$, or if $M \subset N$, or if $M \supset N$. In particular, V and the trivial subspace 0 are each book orthogonal to every subspace of V.

2.29 Proposition. *Let M and N be two subspaces of V. The following are equivalent:*

(i) $P_M P_N$ *is an orthogonal projection,*

(ii) P_M *and* P_N *commute,*

(iii) $M \perp_B N$,

and

(iv) $P_M P_N = P_L = P_N P_M$ *with* $L = M \cap N$.

Proof. (iv) *implies* (i): Trivial.

(i) *implies* (ii): Given that $P_M P_N$ is an orthogonal projection, one has

$$\langle P_M P_N x, y\rangle = \langle x, P_M P_N y\rangle = \langle P_M x, P_N y\rangle = \langle P_N P_M x, y\rangle$$

for all $x, y \in V$, so $P_M P_N = P_N P_M$.

(ii) *implies* (iv): Set $P = P_M P_N$. Using (ii), we have that P is linear, self-adjoint, and idempotent ($P^2 = P_M P_N P_M P_N = P_M^2 P_N^2 = P_M P_N = P$); moreover $\mathcal{R}(P) = L$, so $P = P_L$ by Proposition 2.4.

(iv) *is equivalent to* (iii): On the one hand, (iv) means $P_M P_N - P_L = 0$ and on the other hand, (iii) is equivalent to $P_{M-L} P_{N-L} = 0$ (use (ii) $\Longleftrightarrow$ (iii) in Proposition 2.24). That (iii) and (iv) are equivalent follows from the identity $P_{M-L} P_{N-L} = P_M P_N - P_L$ of Exercise 2.17. ∎

2.30 Exercise [1]. Show that the subspaces M_R and M_C of Exercise 2.26 are book orthogonal. ◇

2.31 Exercise [1]. In $V = \mathbb{R}^3$, let $\boldsymbol{x} = (1,0,0)^T$, $\boldsymbol{y} = (0,1,0)^T$, and $\boldsymbol{z} = (0,0,1)^T$. Put $L = [\boldsymbol{x}+\boldsymbol{y}]$, $M = [\boldsymbol{x},\boldsymbol{y}]$, and $N = [\boldsymbol{y},\boldsymbol{z}]$. Show that $L \subset M$ and $M \perp_B N$, but it is not the case that $L \perp_B N$. ◇

2.32 Exercise [2]. Suppose that $L_1, \ldots, L_k$ are mutually orthogonal subspaces of V. Show that the subspaces

$$M_J = \sum\nolimits_{j \in J} L_j \tag{2.33}$$

of V with $J \subset \{1, 2, \ldots, k\}$ are mutually book orthogonal. ◇

2.34 Exercise [2]. Show that if M and N are book orthogonal subspaces of V, then so are M and $N^\perp$, and so are $M^\perp$ and $N^\perp$. ◇

2.35 Exercise [2]. Suppose that $M_1, \ldots, M_k$ are mutually book orthogonal subspaces of V. Show that $\prod_{1 \le j \le k} P_{M_j}$ is orthogonal projection onto $\bigcap_{1 \le j \le k} M_j$ and that $\prod_{1 \le j \le k} Q_{M_j}$ is orthogonal projection onto $\left(\sum_{1 \le j \le k} M_j\right)^\perp$. ◇

2.36 Exercise [2]. Show that if $M_1, \ldots, M_k$ are mutually book orthogonal subspaces of V, then so are the subspaces of V of the form $\bigcap_{j \in J} M_j$, where $J \subset \{1, 2, \ldots, k\}$. ◇

2.37 Exercise [3]. Prove the following converse to Exercise 2.32: if $M_1, \ldots, M_n$ are mutually book orthogonal subspaces of V, then there exist mutually orthogonal subspaces $L_1, \ldots, L_k$ of V such that each M_m can be written in the form (2.33).
[Hint: Multiply the identity $I_V = \prod_{1 \le \ell \le n} (P_{M_\ell} + Q_{M_\ell})$ by P_{M_m} after expanding out the product.] ◇

2.38 Exercise [4]. Let M and N be arbitrary subspaces of V and put $L = M \cap N$. Let the successive products P_M, $P_N P_M$, $P_M P_N P_M$, $P_N P_M P_N P_M$, ... of P_M and P_N in alternating order be denoted by $T_1, T_2, T_3, T_4, \ldots$. Show that for each $x \in V$,

$$T_j x \to P_L x \quad \text{as} \quad j \to \infty.$$

[Hint: First observe that $T_j - P_L$ is the j-fold product of P_{M-L} and P_{N-L} in alternating order and then make use of Exercise 1.24.] ◇

2E. An algebraic form of Cochran's theorem

We begin with an elementary lemma:

2.39 Lemma. *Suppose M, $M_1, M_2, \ldots, M_k$ are subspaces of V with $M = M_1 + M_2 + \cdots + M_k$. The following are equivalent:*

(i) *every vector $x \in M$ has a unique representation of the form*

$$x = \sum_{i=1}^{k} x_i$$

with $x_i \in M_i$ for each i,

(ii) *for $1 \le i < k$, M_i and $\sum_{j>i} M_j$ are disjoint,*

(iii) *$d(M) = \sum_{i=1}^{k} d(M_i)$ ("d" $\equiv$ dimension).*

Proof. (i) *is equivalent to* (ii): Each of (i) and (ii) is easily seen to be equivalent to the following property:

$$\sum_{i=1}^{k} z_i = 0 \text{ with } z_i \in M_i \text{ for each } i \text{ implies } z_i = 0 \text{ for each } i.$$

(ii) *is equivalent to* (iii): The identity

$$d(K + L) = d(K) + d(L) - d(K \cap L),$$

holding for arbitrary subspaces K and L of V, gives

$$d\Big(M_i + \big(\sum_{j>i} M_j\big)\Big) = d(M_i) + d\big(\sum_{j>i} M_j\big) - d\Big(M_i \cap \big(\sum_{j>i} M_j\big)\Big)$$

for $1 \le i < k$. Combining these relations, we find

$$d(M) = d\big(\sum_{j=1}^{k} M_j\big) = \sum_{j=1}^{k} d(M_j) - \sum_{i=1}^{k-1} d\Big(M_i \cap \big(\sum_{j>i} M_j\big)\Big).$$

So $d(M) = \sum_{i=1}^{k} d(M_i)$ if and only if $d\big(M_i \cap \big(\sum_{j>i} M_j\big)\big) = 0$ for $1 \le i < k$. ∎

When the equivalent conditions of the lemma are met, one says that M is the *direct sum* of the M_i's, written $M = \oplus_{i=1}^{k} M_i$. The sum M of mutually orthogonal subspaces M_i is necessarily direct and in this situation one says that the M_i's form an *orthogonal decomposition* of M.

We are now ready for the following result, which forms the algebraic basis for Cochran's theorem (see Section 3.9) on the distribution of quadratic forms in normally distributed random variables. There is evidently considerable overlap between the result here and Proposition 2.24. Note though that here it is the *sum* of the operators involved that is assumed at the outset to be an orthogonal projection, whereas in Proposition 2.24 it is the individual *summands* that are postulated to be orthogonal projections.

2.40 Proposition (Algebraic form of Cochran's theorem). *Let $T_1, \dots, T_k$ be self-adjoint linear transformations of V into V and suppose that*

$$P = \sum\nolimits_{1 \le i \le k} T_i$$

is an orthogonal projection. Put $M_i = \mathcal{R}(T_i)$, $1 \le i \le k$, and $M = \mathcal{R}(P)$. The following are equivalent:

(i) *each T_i is idempotent,*

(ii) *$T_i T_j = 0$ for $i \neq j$,*

(iii) *$\sum_{i=1}^k d(M_i) = d(M)$,*

and

(iv) *$M_1, \dots, M_k$ form an orthogonal decomposition of M.*

Proof. We treat the case where P is the identity transformation I on V, leaving the general case to the reader as an exercise.

(i) *implies* (iv): Each T_i is an orthogonal projection by (i) and $\sum_i T_i$ an orthogonal projection by the umbrella assumption $\sum_i T_i = I$. Proposition 2.24 says that $\sum_i T_i$ is orthogonal projection onto the orthogonal direct sum $\oplus_i M_i$. But $\sum_i T_i = I$ is trivially orthogonal projection onto V. Thus the M_i's form an orthogonal decomposition of V.

(iv) *implies* (iii): Trivial.

(iii) *implies* (ii): For any $x \in V$,

$$x = Ix = \sum\nolimits_i T_i x$$

so $V = \sum_{i=1}^k M_i$. Taking $x = T_j y$ and using Lemma 2.39 with $M = V$, we get

$$T_i T_j y = 0 \quad \text{for } i = 1, \dots, j-1, j+1, \dots, k$$

(as well as $T_j y = T_j^2 y$—but we don't need this now). Since j and y are arbitrary, $T_i T_j = 0$ for $i \neq j$.

(ii) *implies* (i): One has, using (ii),

$$T_i - T_i^2 = T_i(I - T_i) = T_i\big(\sum\nolimits_{j \neq i} T_j\big) = \sum\nolimits_{j \neq i} T_i T_j = 0$$

for all i. ∎

2.41 Exercise [3]. Prove Proposition 2.40 for the general orthogonal projection P.

[Hint: The case of general P can be reduced to the case $P = I$ treated in the preceding proof by introducing $T_0 = Q_M$. Specifically, let conditions (i′), ..., (iv′) be defined like (i), ..., (iv), but with the indices ranging from 0 to k instead of from 1 to k, and with M replaced by V. The idea is to show that condition (c') is equivalent to condition (c), for c = i, ..., iv. In arguing that (ii) implies (ii′), make use of Exercise 2.7.] ◇

2.42 Exercise [3]. In the context of $\mathbb{R}^n$, suppose $I = T_1 + T_2$ with T_1 being the transformation having matrix

$$(t_{ij}) = \begin{pmatrix} 1/n & 1/n & \dots & 1/n \\ 1/n & 1/n & \dots & 1/n \\ \vdots & \vdots & & \vdots \\ 1/n & 1/n & \dots & 1/n \end{pmatrix}$$

with respect to the usual coordinate basis, so that

$$[T_1((x_j))]_i = \sum\nolimits_{1 \le j \le n} t_{ij} x_j$$

for $1 \le i \le n$. In regard to Proposition 2.40, show that T_1 and T_2 are self-adjoint and that some one (and therefore all) of conditions (i)–(iv) is satisfied. ◇

2.43 Exercise [2]. In the context of $\mathbb{R}^2$, let T_1 and T_2 be the transformations having matrices

$$[T_1] = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad [T_2] = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

with respect to the usual coordinate basis. Show that even though T_1 and T_2 are self-adjoint and sum to the identity, some one (and therefore each) of conditions (i)–(iv) of Proposition 2.40 is not satisfied. ◇

3. Tjur's theorem

Let V be an inner product space. A finite collection $\mathcal{L}$ of distinct subspaces of V such that

$$\begin{aligned} &\text{(T1)}\ L_1, L_2 \in \mathcal{L} \text{ implies } L_1 \text{ and } L_2 \text{ are book orthogonal,} \\ &\text{(T2)}\ L_1, L_2 \in \mathcal{L} \text{ implies } L_1 \cap L_2 \in \mathcal{L}, \text{ and} \\ &\text{(T3)}\ V \in \mathcal{L} \end{aligned} \tag{3.1}$$

is called a *Tjur system* (after the Danish statistician Tue Tjur). For example, the subspaces V, M, and M_0 of Exercise 2.14 constitute a Tjur system, as do the subspaces V, M_R, M_C, and M_G of Exercise 2.26. Since intersections of mutually book orthogonal spaces are themselves mutually book orthogonal (see Exercise 2.36), any finite collection of mutually book orthogonal subspaces of V can be augmented to a Tjur system.

The elements of a Tjur system $\mathcal{L}$ are partially ordered by inclusion. For $L \in \mathcal{L}$ we write $K \le L$ to mean that $K \in \mathcal{L}$ and K is a subset of L, and write $K < L$ to mean that $K \le L$ and $K \ne L$. This notation is used in the following key theorem, which shows that the elements of $\mathcal{L}$ can be represented neatly and simply in terms of an explicit orthogonal decomposition of V.

3.2 Theorem (Tjur's theorem). *Let $\mathcal{L}$ be a Tjur system of subspaces of V. For each $L \in \mathcal{L}$, put*

$$L^\circ = L - \sum\nolimits_{K<L} K. \tag{3.3}$$

Then

(i) *the subspaces L° for $L \in \mathcal{L}$ are mutually orthogonal,*

(ii) $V = \sum_{L\in\mathcal{L}} L^\circ$,

(iii) *for each $L \in \mathcal{L}$, one has $L = \sum_{K\le L} K^\circ$.*

Moreover, the L°'s are uniquely determined by these conditions.

Proof. (i) *and* (iii) *imply* (3.3): Suppose that $\{ L^* : L \in \mathcal{L} \}$ is a collection of mutually orthogonal subspaces of V such that $L = \sum_{K\le L} K^*$ for each $L \in \mathcal{L}$; we claim $L^* = L^\circ$ for each L. For this note that $L^* + \bigl(\sum_{K<L} K^*\bigr)$ is an orthogonal decomposition of L, so

$$L^* = L - \bigl(\sum\nolimits_{K<L} K^*\bigr) = L - \bigl(\sum\nolimits_{K<L} K\bigr) = L^\circ.$$

(3.3) *implies* (i)–(iii): We first use induction to show that

$$L = \sum\nolimits_{K\le L} K^\circ \tag{3.4}$$

for each $L \in \mathcal{L}$. Condition (T2) implies that $\mathcal{L}$ has a smallest element, say L_0; (3.4) holds for L_0 because $L_0^\circ = L_0$. Suppose now

$$K = \sum\nolimits_{J\le K} J^\circ \quad \text{for all } K \in \mathcal{L} \text{ with } K < L. \tag{3.5}$$

Then

$$\begin{aligned} L &= L^\circ + \sum\nolimits_{K<L} K && \text{(by the definition (3.3) of } L^\circ\text{)} \\ &= L^\circ + \sum\nolimits_{K<L}\bigl(\sum\nolimits_{J\le K} J^\circ\bigr) && \text{(by the induction hypothesis (3.5))} \\ &= L^\circ + \sum\nolimits_{J<L} J^\circ = \sum\nolimits_{K\le L} K^\circ, \end{aligned}$$

so (3.4) holds for L. (iii) now follows from Exercise 3.7 below, and (ii) holds because $V \in \mathcal{L}$ by (T3).

To complete the proof we show that (i) holds. Let $L_1, L_2 \in \mathcal{L}$; we have to show

$$L_1^\circ \perp L_2^\circ. \tag{3.6}$$

By (T2) $L = L_1 \cap L_2 \in \mathcal{L}$. There are three cases to consider: (1) $L < L_1$ and $L < L_2$; (2) $L = L_1$; and (3) $L = L_2$. In case (1), $L_1^\circ \subset L_1 - L$ and $L_2^\circ \subset L_2 - L$, so (3.6) holds because L_1 and L_2 are book orthogonal by (T1). In case (2) $L_1 < L_2$, so (3.6) holds because $L_1^\circ \subset L_1$ and $L_2^\circ \subset L_2 - L_1$. Case (3) is like case (2). ∎

3.7 Exercise *(The principle of induction for a partially ordered set)* [4]. Let I be a finite set and let "$\leq$" be a *partial order* on I:

$$i \leq i \quad \text{for all } i \in I, \tag{3.8$_1$}$$

$$i \leq j \text{ and } j \leq k \quad \text{implies} \quad i \leq k, \tag{3.8$_2$}$$

$$i \leq j \text{ and } j \leq i \quad \text{implies} \quad i = j. \tag{3.8$_3$}$$

For $i, j \in I$, write $i < j$ to mean $i \leq j$ and $i \neq j$. Say that an element j of I is *minimal* if there doesn't exist an $i \in I$ such that $i < j$. For each $i \in I$, let $S(i)$ be a statement involving i. Prove the *principle of induction for* I: if $S(i)$ is valid for each minimal element i and if for each non-minimal element j, $S(j)$ is valid whenever $S(i)$ is valid for all $i < j$, then $S(j)$ is valid for all $j \in I$.
[Hint: If $i < j$ and $j < k$, then i, j, and k must be distinct elements of I.] ◇

3.9 Exercise [2]. Show that the subspace L° defined by (3.3) can be written as $L \cap \left(\bigcap_{K<L}(L - K)\right)$.
[Hint: See Exercise 2.2.] ◇

3.10 Exercise [2]. Show by example that the conclusion (i) of Tjur's theorem would not follow if condition (T2) were dropped from (3.1). ◇

3.11 Exercise [1]. Use Tjur's Theorem to solve Exercise 2.37. ◇

Tjur's theorem states in part that $\sum_{L\in\mathcal{L}} L^\circ$ is an orthogonal decomposition of V. In applications one needs to know how to project onto the subspaces L°, what dimensions these subspaces have, and how long the components $P_{L^\circ}v$ of a vector $v \in V$ are. This information is readily obtained from corresponding information about the original spaces $L \in \mathcal{L}$. Indeed, conclusion (iii) in Tjur's theorem implies that

$$P_L = \sum\nolimits_{K \leq L} P_{K^\circ} \quad \text{for } L \in \mathcal{L}. \tag{3.12}$$

These equations can be solved recursively for the P_{L°'s, working upwards from the minimal element L_0 of $\mathcal{L}$ to the maximal element V:

$$\begin{aligned} P_{L_0^\circ} &= P_{L_0}, \\ P_{L^\circ} &= P_L - \sum\nolimits_{K<L} P_{K^\circ} \quad \text{for } L > L_0. \end{aligned} \tag{3.13}$$

Similarly, the dimensions of the L° spaces can be found by solving the equations

$$d(L) = \sum\nolimits_{K \leq L} d(K^\circ), \quad L \in \mathcal{L}, \tag{3.14}$$

and for $v \in V$ the squared lengths

$$\ell_{L^\circ}^2 \equiv \ell_{L^\circ}^2 v = \|P_{L^\circ}v\|^2 \tag{3.15}$$

can be found from the corresponding squared lengths

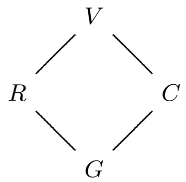
$$\ell_L^2 \equiv \ell_L^2(v) = \|P_L v\|^2 \tag{3.16}$$

by solving the equations

$$\ell_L^2 = \sum\nolimits_{K \le L} \ell_{K^\circ}^2, \quad L \in \mathcal{L}. \tag{3.17}$$

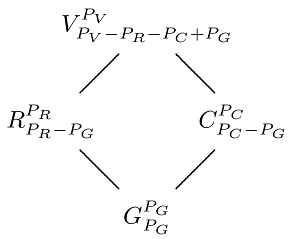
In specific situations the computations are facilitated by referring to a *structure diagram* which displays the ordering of the elements of $\mathcal{L}$. The following example illustrates the method.

3.18 Example. Consider the Tjur system $\{V, M_R, M_C, M_G\}$ of Exercise 2.26. For notational simplicity write R for M_R, C for M_C, and G for M_G. The structure diagram is

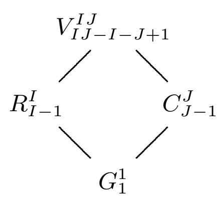


A line between two elements of $\mathcal{L}$ indicates that the lower subspace is included in the upper one. No line needs to be drawn for the inclusion $G \subset V$, since that relation is readily inferred from the diagram as it stands.

Formulas for the projections P_L were given in Exercise 2.26 (see (2.27)). To calculate the P_{L°'s from the P_L's using (3.13), begin by superscripting each L in the structure diagram by P_L. Then, working upwards from the bottom of the diagram, subscript each L by P_{L°, calculated as the difference between the corresponding superscript and the sum of all subscripts for spaces K strictly below L (i.e., for all K such that $K < L$). For example, the subscript on V is calculated as $P_V - (P_{R^\circ} + P_{C^\circ} + P_{G^\circ})$. The result is



The method of su$^{\text{per}}_{\text{b}}$scripts can also be used to organize the computation of the $d(L^\circ)$'s and $\ell_{L^\circ}^2$'s. For example, since $d(V) = IJ$, $d(R) = I$, $d(C) = J$, and $d(G) = 1$, the annotated structure diagram for the calculation of the $d(L^\circ)$'s is



•

The *analysis of variance table* for a Tjur system $\mathcal{L}$ lists for each $L \in \mathcal{L}$ the quantities L°, $d(L^\circ)$, and $\ell^2_{L^\circ} = \|P_{L^\circ} v\|^2$ appearing in the decompositions

$$V = \sum\nolimits_{L \in \mathcal{L}} L^\circ, \quad d(V) = \sum\nolimits_{L \in \mathcal{L}} d(L^\circ), \quad \|v\|^2 = \sum\nolimits_{L \in \mathcal{L}} \ell^2_{L^\circ}(v).$$

Following conventional statistical practice, rows for smaller L's (used in building simpler models) appear above rows for larger L's (used in building more complicated models). The analysis of variance table for the Tjur system in the preceding Example is given in Table 3.19.

3.19 Table. *Analysis of variance table for the Tjur system* $\{V, R, C, G\}$ *of Example* 3.18. *The quantities* $\ell^2_L = \|P_L v\|^2$ *appearing in the "squared length" column are given by*

$$\ell^2_V = \sum\nolimits_{ij} v^2_{ij}, \quad \ell^2_R = J \sum\nolimits_i \bar{v}^2_{i\cdot}, \quad \ell^2_C = I \sum\nolimits_j \bar{v}^2_{\cdot j}, \quad \ell^2_G = IJ\bar{v}^2_{\cdot\cdot}\,.$$

Label	*Subspace*	*Dimension*	*Squared length*
G	$G^\circ = G$	1	$\ell^2_G = IJ\bar{v}^2_{\cdot\cdot}$
R	$R^\circ = R - G$	$I - 1$	$\ell^2_R - \ell^2_G = J\sum_i(\bar{v}_{i\cdot} - \bar{v}_{\cdot\cdot})^2$
C	$C^\circ = C - G$	$J - 1$	$\ell^2_C - \ell^2_G = I\sum_j(\bar{v}_{\cdot j} - \bar{v}_{\cdot\cdot})^2$
V	$V^\circ = V - (R + C + G)$	$IJ - I - J + 1$	$\ell^2_V - \ell^2_R - \ell^2_C + \ell^2_G$ $= \sum_{ij}(v_{ij} - \bar{v}_{i\cdot} - \bar{v}_{\cdot j} + \bar{v}_{\cdot\cdot})^2$
Sum	V	IJ	$\ell^2_V = \sum_{ij} v^2_{ij}$

3.20 Exercise [2]. Verify the entries in the "squared lengths" column of Table 3.19. To get the expressions to the left of the "=" signs, use the method of su$^{\text{per}}_{\text{b}}$scripts. To get the expressions to the right of the "=" signs, use the formulae for the P_{L°'s derived in Example 3.18. ◇

3.21 Exercise [2]. Draw the structure diagram for the Tjur system $\{V, M, M_0\}$ of Exercise 2.14, and use the method of su$^{\text{per}}_{\text{b}}$scripts to construct the corresponding analysis of variance table. ◇

3.22 Exercise *(The Möbius inversion formula)* [4]. As in Exercise 3.7, let "$\leq$" be a partial order on a finite set I. (1) The *Möbius function* of $(I, \leq)$ is the matrix $\mu = (\mu_{ik})_{i \in I, k \in I}$ such that $\mu_{ik} = 0$ for $i \not\leq k$ and such that

$$\sum\nolimits_{j\,:\,i \leq j \leq k} \mu_{ij} = \delta_{ik} \quad \text{for } i \leq k \tag{3.23}$$

or, equivalently,

$$\sum\nolimits_{j\,:\,i \leq j \leq k} \mu_{jk} = \delta_{ik} \quad \text{for } i \leq k. \tag{3.24}$$

Prove that such a matrix μ exists, and is inverse to the matrix $\zeta = (\zeta_{ik})_{i\in I,k\in I}$ with

$$\zeta_{ik} = \begin{cases} 1, & \text{if } i \le k, \\ 0, & \text{otherwise;} \end{cases}$$

ζ is called the *zeta function* of $(I, \le)$. (2) Prove of functions f and g mapping I into some vector space W that

$$g(j) = \sum\nolimits_{i\le j} f(i) \quad \text{for all } j \in I \tag{3.25}$$

if and only if

$$f(j) = \sum\nolimits_{i\le j} g(i)\mu_{ij} \quad \text{for } j \in I; \tag{3.26}$$

this result is called the *Möbius inversion formula*.
[Hint: For each fixed $i \in I$, equation (3.23) can be solved recursively for μ_{ik} with $i \le k$.] ◇

3.27 Exercise [3]. Recall that the elements of a Tjur system $\mathcal{L}$ of subspaces of V are partially ordered by inclusion ("$\le$"). Let μ be the Möbius function for $(\mathcal{L}, \le)$. Use the Möbius inversion formula from the preceding exercise to show that

$$P_{L^\circ} = \sum P_K \mu_{KL}, \quad d(L^\circ) = \sum d(K)\mu_{KL}, \quad \ell^2_{L^\circ} = \sum \ell^2_K \mu_{KL} \tag{3.28}$$

for all $L \in \mathcal{L}$, the summations in each case extending over $K \le L$, or even over all $K \in \mathcal{L}$. Check (3.28) against the entries in Table 3.19. ◇

As Example 3.18 suggests, Tjur systems arise naturally in the theory of the design of experiments. The rest of this section elaborates on this idea. In particular, we will see how the notions of book orthogonality, intersection, and inclusion of subspaces correspond to certain simple relationships among the levels of the treatments assigned to the experimental units.

The design of an experiment often calls for the available experimental units (e.g., plots of land) to be divided into groups, with each unit in a group receiving the same level of some experimental treatment (e.g., type of seed). To abstract this idea, let X be a finite set whose elements are called *experimental units*. A (treatment) *factor* F is a partition of X into nonempty disjoint subsets called *blocks*, or *levels of* F. The factor U whose blocks are single units is called the *units factor*, while the factor T having only one block is called the *trivial factor*.

For a subset f of X put

$$|f| = \text{cardinality}(f), \tag{3.29}$$

the number of experimental units in f. A factor F is said to be *balanced* if each of its levels is assigned the same number of units, i.e., if

$$|f| = |g| \quad \text{for all blocks } f \text{ and } g \text{ of } F. \tag{3.30}$$

For example, if $X = \{\,(i,j) : 1 \le i \le I,\ 1 \le j \le J\,\}$, then the row factor R with blocks $r_i = \{\,(i,j) : 1 \le j \le J\,\}$ for $1 \le i \le I$ is balanced because $|r_i| = J$ for each i.

If F is a factor and x and y are experimental units, the notation

$$x \sim_F y \tag{3.31}$$

means that x and y are assigned the same level of F, i.e., belong to the same block of F. The relation "$\sim_F$" is an equivalence relation:

$$x \sim_F x \quad \text{for each } x \in X, \tag{3.32$_1$}$$
$$x \sim_F y \quad \text{implies} \quad y \sim_F x, \tag{3.32$_2$}$$
$$x \sim_F y \text{ and } y \sim_F z \quad \text{implies} \quad x \sim_F z. \tag{3.32$_3$}$$

Conversely, if "$\sim$" is an arbitrary equivalence relation on X, then its equivalence classes constitute a factor. The correspondence between factors and equivalence relations is one-to-one.

Let F and G be factors. One says F *is nested in* G if each block of G is a union of blocks of F, or, equivalently, if

$$\text{for } x, y \in X,\ \ x \sim_F y \text{ implies } x \sim_G y; \tag{3.33}$$

this situation is written as

$$G \le F \tag{3.34}$$

because it implies that F has at least as many levels as G. The units factor U is nested in each factor, and each factor is nested in the trivial factor T. The relation "$\le$" is a partial order on the set of factors of X:

$$F \le F \quad \text{for all factors } F, \tag{3.35$_1$}$$
$$H \le G \text{ and } G \le F \quad \text{implies} \quad H \le F, \tag{3.35$_2$}$$
$$G \le F \text{ and } F \le G \quad \text{implies} \quad F = G. \tag{3.35$_3$}$$

Interesting factors can be built up from given ones by means of their equivalence relations. For example, suppose F and G are factors. Write

$$x \sim_\times y \tag{3.36}$$

for $x, y \in X$ to mean $x \sim_F y$ and $x \sim_G y$. "$\sim_\times$" is an equivalence relation; the factor $F \times G$ whose blocks are the $\sim_\times$-equivalence classes is called the *cross-classification induced by F and G*. It turns out that $F \times G$ is the least upper bound of F and G with respect to the "nested in" ordering, so $F \times G$ is also called the *maximum of F and G* and written as $F \vee G$.

3.37 Exercise [2]. Let F and G be factors. Show that $F \times G$ is indeed the *least upper bound* of F and G with respect to the ordering (3.34):

$$F \le F \times G \quad \text{and} \quad G \le F \times G, \tag{3.38$_1$}$$
$$F \le H \text{ and } G \le H \quad \text{implies} \quad F \times G \le H. \tag{3.38$_2$}$$ ◇

Let again F and G be factors. Write

$$x \sim_\wedge y \tag{3.39}$$

for $x, y \in X$ to mean that there exists a finite sequence $z_0, \ldots, z_k$ of experimental units such that

$$z_0 = x \quad \text{and} \quad z_k = y \tag{3.40$_1$}$$

and

$$z_{j-1} \sim_F z_j \quad \text{and/or} \quad z_{j-1} \sim_G z_j \quad \text{for } 1 \le j \le k. \tag{3.40$_2$}$$

Loosely speaking, $x \sim_\wedge y$ if it is possible to move from x to y through X in a finite number of steps, each of which takes place within a block of F (an F-step) or a block of G (a G-step); an efficient move would alternate F-steps with G-steps, but this is not required. "$\sim_\wedge$" is an equivalence relation; the factor whose blocks are the $\sim_\wedge$-equivalence classes turns out to be the greatest lower bound of F and G, and so is called the *minimum of* F *and* G and written as $F \wedge G$.

3.41 Exercise [3]. Let F and G be factors. Verify that the relation "$\sim_\wedge$" is indeed an equivalence relation and that the factor $F \wedge G$ is the *greatest lower bound* of F and G:

$$F \wedge G \le F \quad \text{and} \quad F \wedge G \le G, \tag{3.42$_1$}$$

$$H \le F \text{ and } H \le G \quad \text{implies} \quad H \le F \wedge G. \tag{3.42$_2$}$$ ◇

3.43 Example. Suppose $X = \{1,2,3,4,5,6,7,8,9\}$ and the experimental units are allocated to the 3 levels r_1, r_2, and r_3 of a row factor R and the 3 levels c_1, c_2, and c_3 of a column factor C as indicated below:

	Columns		
Rows	c_1	c_2	c_3
r_1	1,2	3	
r_2	4,5	6	9
r_3			7,8

For example, units 1, 2, and 3 are assigned to level r_1 of R, and units 3 and 6 are assigned to level c_2 of C. The blocks of $R \times C$ are $\{1,2\}$, $\{3\}$, $\{4,5\}$, $\{6\}$, $\{7,8\}$, and $\{9\}$. $R \wedge C$ is the trivial factor because each unit can be reached from each other unit by a succession of row and column steps. If the last experimental unit 9 were to be removed from X, then $R \wedge C$ would have 2 blocks, $\{1,2,3,4,5,6\}$ and $\{7,8\}$. •

3.44 Exercise [4]. Let $F_1, \ldots, F_k$ be factors such that $f_1 \cap \cdots \cap f_k \ne \emptyset$ for all $f_1 \in F_1, \ldots, f_k \in F_k$. For subsets J of $K \equiv \{1, \ldots, k\}$, let

$$F_J = \prod\nolimits_{j \in J} F_j \tag{3.45}$$

denote the cross-classification induced by factors F_j with $j \in J$ — $x \sim_{F_J} y$ if and only if $x \sim_{F_j} y$ for all $j \in J$. (Take $F_\emptyset$ to be trivial factor T.) Show that

$$F_{J_1} \wedge F_{J_2} = F_{J_1 \cap J_2} \tag{3.46}$$

for all subsets J_1 and J_2 of K.
[Hint First do the case where $k = 3$, $J_1 = \{1, 2\}$, and $J_2 = \{2, 3\}$.] ◇

Let now V be the vector space of real-valued functions on X. For each subset f of X, let $I_f \in V$ be the *indicator function of* f:

$$I_f(x) = \begin{cases} 1, & \text{if } x \in f, \\ 0, & \text{if } x \notin f. \end{cases} \tag{3.47}$$

For each factor F, let

$$L_F = [\, I_f : f \in F \,] = \{\, v \in V : x \sim_F y \text{ implies } v(x) = v(y) \,\} \tag{3.48}$$

be the subspace of V consisting of functions that are constant on each block of F. Note that L_F has dimension

$$d(L_F) = |F|, \ \text{ the number of blocks of } F. \tag{3.49}$$

3.50 Proposition. *Let F and G be factors. Then*

$$G \le F \quad \textit{if and only if} \quad L_G \subset L_F \tag{3.51}$$

and

$$L_{F \wedge G} = L_F \cap L_G. \tag{3.52}$$

Proof. *$G \le F$ implies $L_G \subset L_F$*: $G \le F$ means that each block of G is a union of blocks of F, so constancy on blocks of G trivially implies constancy on blocks of F.

$L_G \subset L_F$ implies $G \le F$: Suppose $L_G \subset L_F$. Let g be a block of G and f a block of F such that $f \cap g \neq \emptyset$; we have to show that $f \subset g$. For this, note that the function $I_g \in L_G \subset L_F$ has the value 1 at some point of f and so is identically 1 on f.

$L_{F \wedge G} \subset L_F \cap L_G$: This follows from (3.51) applied to $F \wedge G \le F$ and $F \wedge G \le G$.

$L_F \cap L_G \subset L_{F \wedge G}$: Suppose $v \in V$ is constant over blocks of both F and G; we have to show v is constant over blocks of $F \wedge G$. For this suppose $x, y \in X$ with $x \sim_{F \wedge G} y$. Choose z_0, …, z_k satisfying (3.40). Then $v(z_{j-1}) = v(z_j)$ for each j, so $v(x) = v(z_0) = v(z_k) = v(y)$. ∎

Let now V be endowed with the dot-product

$$\langle u, v \rangle = \sum\nolimits_{x \in X} u(x) v(x). \tag{3.53}$$

Let P_F denote orthogonal projection onto L_F. Since the functions I_f for $f \in F$ are an orthogonal basis for L_F, we have

$$P_F v = \sum\nolimits_{f \in F} \bar{v}_f I_f \tag{3.54}$$

where

$$\bar{v}_f = \frac{\langle I_f, v\rangle}{\langle I_f, I_f\rangle} = \frac{1}{|f|}\sum\nolimits_{x\in F} v(x); \tag{3.55}$$

in other words, the value of $P_F v$ at an experimental unit y is the average value of $v \in V$ over all units x that are in the same block of F as y. In situations where the symbol v is adorned with subscripts and/or superscripts, we will sometimes write $A_f(v)$ in place of $\bar{v}_f$. Note that

$$\ell^2_{L_F}(v) \equiv \|P_F v\|^2 = \sum\nolimits_{x\in X}\big((P_F v)(x)\big)^2 = \sum\nolimits_{f\in F}|f|\,\bar{v}_f^2 \equiv SS_F, \tag{3.56}$$

the symbol "*SS*" denoting *sum of squares*.

Factors F and G are said be *orthogonal* if L_F and L_G are book orthogonal. By Proposition 2.29, this is equivalent to $P_F P_G = P_G P_F$, and also to $P_F P_G = P_{L_F \cap L_G}$. In view of (3.52), this last condition is the same as $P_F P_G = P_{F\wedge G}$. The following proposition gives a useful characterization of orthogonality in terms of the sizes of the blocks of F, G, and $F \wedge G$. If f and h are subsets of X, say that f *is nested in* h is f is a subset of h.

3.57 Proposition. *Factors F and G are orthogonal if and only if*

$$\frac{|f\cap g|}{|h|} = \frac{|f|}{|h|}\times\frac{|g|}{|h|} \tag{3.58}$$

for all blocks $f \in F$, $g \in G$, and $h \in H \equiv F \wedge G$ such that f and g are nested in h.

Proof. The functions $\delta_x = I_{\{x\}}$, $x \in X$, form a basis for V, so F and G are orthogonal if and only if $\langle \delta_x, P_H\delta_y\rangle = \langle \delta_x, P_F P_G \delta_y\rangle$, and hence if and only if

$$\langle \delta_x, P_H \delta_y\rangle \equiv \langle P_F\delta_x, P_G\delta_y\rangle \tag{3.59}$$

for all $x, y \in X$. The left- and right-hand sides of (3.59) are 0 unless x and y both belong to the same block h of H, in which case the left-hand side is

$$\Big\langle \delta_x, \sum\nolimits_{\tilde{h}\in H} A_{\tilde{h}}(\delta_y) I_{\tilde{h}}\Big\rangle = A_h(\delta_y) = 1/|h|$$

and the right-hand is

$$\Big\langle \sum\nolimits_{\tilde{f}\in F} A_{\tilde{f}}(\delta_x) I_{\tilde{f}}, \sum\nolimits_{\tilde{g}\in G} A_{\tilde{g}}(\delta_y) I_{\tilde{g}}\Big\rangle = A_f(\delta_x)A_g(\delta_y)\langle I_f, I_g\rangle = \frac{|f\cap g|}{|f|\,|g|},$$

where f is the block of F containing x and g is the block of G containing y; necessarily $f \subset h$ and $g \subset h$. ∎

3.60 Exercise [2]. Show of factors F and G that $G \leq F$ implies that F and G are orthogonal. Solve this exercise twice, once using (3.51) and once using (3.58). ◇

The following exercise gives a convenient reformulation of the orthogonality criterion (3.58).

3.61 Exercise [3]. Let F and G be factors. Show that F and G are orthogonal if and only if

$$\begin{array}{l}\text{for each level } h \text{ of } H \equiv F \wedge G \text{ and levels } g_1 \text{ and } g_2 \text{ of } G \text{ nested in } h,\\ \text{there exists a constant } c = c(h, g_1, g_2) \text{ such that } |f \cap g_1| = c|f \cap g_2| \\ \text{for all levels } f \text{ of } F \text{ nested in } h,\end{array} \qquad (3.62)$$

this being the so-called *condition of proportional cell counts*. ◇

3.63 Exercise [2]. Each of the four tables below gives the number of experimental units to be allocated to the cells of a 3×3 rectangular array. In which cases is the implied row factor R orthogonal to the implied column factor C?

Table 1

1	1	1
1	1	1
1	1	1

Table 2

3	3	3
2	2	2
1	1	1

Table 3

2	1	0
2	1	1
0	0	2

Table 4

2	1	0
2	1	0
0	0	2

◇

3.64 Exercise [2]. Consider a $2 \times 2 \times 2$ design having a row factor R with levels r_1 and r_2, a column factor C with levels c_1 and c_2, and a height factor H with levels h_1 and h_2. Suppose the 10 experimental units are allocated in such a way that

$$|r_i \cap c_j \cap h_k| = \begin{cases} 2, & \text{if } (i,j,k) = (1,1,2) \text{ or } (2,2,2), \\ 1, & \text{otherwise.} \end{cases}$$

Show that $R \perp H$ and $C \perp H$, but $(R \times C) \not\perp H$. Is $(R \wedge C) \perp H$? ◇

A *Tjur design* $\mathcal{D}$ is a collection of distinct factors such that

(D1) $F, G \in \mathcal{D}$ implies F and G are orthogonal,

(D2) $F, G \in \mathcal{D}$ implies $F \wedge G \in \mathcal{D}$, and (3.65)

(D3) the units factor U is in $\mathcal{D}$.

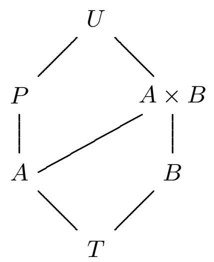
The subspaces L_F for F in a Tjur design $\mathcal{D}$ constitute a Tjur system $\mathcal{L}$; indeed, (D1) trivially implies (T1), (D2) implies (T2) by (3.52), and (D3) implies (T3) since $L_U = V$. Moreover by (3.51) the ordering of the factors $F \in \mathcal{D}$ corresponds to the ordering of the L_F's. The quantities $P_{L_F} \equiv P_F$, $d(L_F)$, and $SS_F = \|P_F v\|^2$ are easily obtained from (3.54), (3.49), and (3.56) respectively, and the analysis of variance table can be written down more or less at sight using a factor structure diagram and the method of $\text{su}^{\text{per}}_{\text{b}}$scripts.

3.66 Exercise *(A split-plot design)* [4]. Suppose that 5 treatments $a_1, \ldots, a_5$ are applied to 15 plots, each treatment being applied to 3 plots. Each plot is subdivided into 2 subplots to which two further treatments b_1, b_2 are applied, one treatment to each subplot. The relevant factors are: the units factor U (for subplots) with 30 levels; a "plot" factor P with 15 levels, an "a-treatment" factor A with 5 levels, a "b-treatment" factor B with 2 levels, a "combined treatment" factor $A \times B$ with 10 levels, and the trivial factor T with 1 level.

Using the following template of •'s to represent the 30 subplots, show one way of assigning the experimental units to these factors:

• • •   • • •   • • •   • • •   • • •
• • •   • • •   • • •   • • •   • • •

Verify that $\{U, P, A, B, A \times B, T\}$ is a Tjur design with factor structure diagram



and use the method of su$^{\text{per}}_{\text{b}}$scripts to compute the corresponding analysis of variance table. ◇

4. Self-adjoint transformations and the spectral theorem

Recall that a linear transformation $T: V \to V$ such that

$$\langle Tx, y \rangle = \langle x, Ty \rangle \quad \text{for all } x, y \in V \tag{4.1}$$

is said to be *self-adjoint.* In this section we are going to study the geometry of self-adjoint transformations.

A few words are in order about one reason why self-adjoint transformations are of special interest in the study of the GLM. Consider the canonical example: $(V, \langle \cdot, \cdot \rangle) = (\mathbb{R}^n, \text{dot product})$. Suppose $\boldsymbol{Y}^{n\times 1} = (Y_i)_{1 \le i \le n}$ is a random vector in $\mathbb{R}^n$ with dispersion matrix $\boldsymbol{\Sigma}^{n\times n} = (\sigma_{ij})_{1\le i \le n, 1 \le j \le n}$:

$$\sigma_{ij} = \text{Cov}(Y_i, Y_j).$$

The covariance between any two linear combinations $\sum_i c_i Y_i = \boldsymbol{c}^T\boldsymbol{Y}$ and $\sum_j d_j Y_j = \boldsymbol{d}^T\boldsymbol{Y}$ of the Y's is given by

$$\text{Cov}(\boldsymbol{c}^T\boldsymbol{Y}, \boldsymbol{d}^T\boldsymbol{Y}) = \sum\nolimits_{ij} c_i \sigma_{ij} d_j = \boldsymbol{c}^T\boldsymbol{\Sigma}\boldsymbol{d}.$$

The probabilistic identity

$$\text{Cov}(\boldsymbol{c}^T\boldsymbol{Y}, \boldsymbol{d}^T\boldsymbol{Y}) = \text{Cov}(\boldsymbol{d}^T\boldsymbol{Y}, \boldsymbol{c}^T\boldsymbol{Y})$$

implies the algebraic identity

$$\langle\, \boldsymbol{c}, \boldsymbol{\Sigma}\boldsymbol{d} \,\rangle = \boldsymbol{c}^T\boldsymbol{\Sigma}\boldsymbol{d} = \boldsymbol{d}^T\boldsymbol{\Sigma}\boldsymbol{c} = \langle\, \boldsymbol{\Sigma}\boldsymbol{c}, \boldsymbol{d} \,\rangle.$$

This says that the linear transformation

$$T_{\boldsymbol{\Sigma}}: \boldsymbol{x} \to \boldsymbol{\Sigma}\boldsymbol{x}$$

is self-adjoint. Notice also that

$$\langle T_{\boldsymbol{\Sigma}}\boldsymbol{x}, \boldsymbol{x}\rangle = \langle \boldsymbol{\Sigma x}, \boldsymbol{x}\rangle = \boldsymbol{x}^T\boldsymbol{\Sigma x} = \mathrm{Var}(\boldsymbol{x}^T\boldsymbol{Y}).$$

Thus variance/covariance structures are intimately related to self-adjoint transformations.

4.2 Exercise [2]. Let $T: V \to V$ be a linear transformation and let $\boldsymbol{A} = (a_{ik})$ be the matrix of T with respect to a given orthonormal basis $b_1, \ldots, b_n$ for V, so that

$$Tb_k = \sum\nolimits_i a_{ik} b_i \quad \text{for } 1 \le k \le n.$$

Show that

$$a_{jk} = \langle b_j, Tb_k\rangle$$

for all j, k and conclude that T is self-adjoint if and only if $\boldsymbol{A}$ is symmetric. ◇

We have seen that any orthogonal projection is self-adjoint. And, of course, the action of an orthogonal projection is geometrically obvious: P_M leaves each $x \in M$ fixed, kills each $y \in M^\perp$, and maps the general V vector $z = x + y$ with $x \in M$ and $y \in M^\perp$ into $P_M z = P_M x + P_M y$. More generally if

$$V = M_1 + \cdots + M_k$$

is a decomposition of V into mutually orthogonal subspaces and if $\lambda_1, \ldots, \lambda_k$ are any scalars, then

$$T = \sum\nolimits_{1\le i\le k} \lambda_i P_{M_i} \tag{4.3}$$

is self-adjoint, being a linear combination of self-adjoint transformations. Again the action of such a T is geometrically clear: since

$$T\big(\sum\nolimits_{1\le i\le k} x_i\big) = \sum\nolimits_{1\le i\le k} \lambda_i x_i \quad \text{for } x_i \in M_i,\ 1 \le i \le k,$$

T just dilates (or contracts) by a factor of λ_i within M_i, $1 \le i \le k$. The remarkable thing is that (4.3) describes the most general self-adjoint transformation on V:

4.4 Theorem (Spectral theorem). *Suppose $T: V \to V$ is linear and self-adjoint. Then there exists an orthogonal decomposition $V = \oplus_{i=1}^k M_i$ of V into non-trivial subspaces M_i and distinct scalars λ_i such that*

$$T = \sum\nolimits_{1\le i\le k} \lambda_i P_{M_i}. \tag{4.5}$$

Suppose T is self-adjoint and therefore of the form (4.5). One can characterize the λ_i's and M_i's in terms of the eigenvalues and eigenvectors of T. Recall that for any linear transformation $S: V \to V$, λ is called an *eigenvalue* of S if there is a non-zero *eigenvector* $x \in V$ such that

$$Sx = \lambda x; \tag{4.6}$$

the subspace of all $x \in V$ such that (4.6) holds is called the *eigenmanifold* of λ, denoted E_λ. From what was said earlier about the geometric action of T it is clear that the λ_i's are precisely the distinct eigenvalues of T and the M_i's are the corresponding eigenspaces. Also for any $x \in V$,

$$Q_T(x) \equiv \langle Tx, x \rangle = \left\langle \sum\nolimits_i \lambda_i x_i, \sum\nolimits_j x_j \right\rangle = \sum\nolimits_i \lambda_i \|x_i\|^2, \qquad (4.7)$$

where $x_i = P_{M_i} x$ for $1 \le i \le k$. (The Q in Q_T stands for "quadratic form".) For x of unit length, i.e., for

$$1 = \|x\|^2 = \sum\nolimits_i \|x_i\|^2,$$

$Q_T(x)$ is thus a weighted average of the λ_i's. This observation leads to the following extremal characterization of the λ_i's. Assuming the λ_i's to be indexed in decreasing order, so that $\lambda_1 > \lambda_2 > \cdots > \lambda_k$, one has

$$\begin{aligned} \lambda_1 &= \sup\{\, Q_T(x) : \|x\| = 1 \,\} \\ \lambda_2 &= \sup\{\, Q_T(x) : \|x\| = 1,\ x \perp E_{\lambda_1} \,\} \\ &\ \ \vdots \qquad\qquad\qquad \vdots \\ \lambda_k &= \sup\{\, Q_T(x) : \|x\| = 1,\ x \perp (E_{\lambda_1} + \cdots + E_{\lambda_{k-1}}) \,\}. \end{aligned} \qquad (4.8)$$

Proof of the Spectral theorem. Let T be self-adjoint and linear. We are going to produce the representation (4.5) using (4.8) to define the λ_i's. To begin with, consider

$$\lambda_1 \equiv \sup\{\, Q_T(x) : x \in B \,\}$$

where

$$B = \{\, x \in V : \|x\| = 1 \,\}.$$

Since with respect to the distance function

$$d(x, y) = \|y - x\|$$

the closed bounded set B is compact (see Exercise 1.23) and the mapping

$$x \to Q_T(x) = \langle Tx, x \rangle$$

is continuous, and since continuous functions on compact sets achieve their suprema, there exists an $x \in B$ such that

$$Q_T(x) = \lambda_1.$$

We claim that such an x is an eigenvector of T with eigenvalue λ_1, i.e.,

$$Tx = \lambda_1 x.$$

Notice that

$$\lambda_1 x = \langle Tx, x \rangle x = \frac{\langle Tx, x \rangle}{\langle x, x \rangle}\, x = P_{[x]}(Tx).$$

So if

$$Tx \neq \lambda_1 x$$

then the residual vector

$$r = Tx - P_{[x]}(Tx)$$

is non-zero, and the idea is to produce a contradiction by perturbing x in the direction r so as to produce a vector $y \in B$ such that $Q_T(y) > Q_T(x)$. Specifically for real δ set

$$y_\delta = \frac{x + \delta r}{\|x + \delta r\|} \in B$$

and note that

$$\begin{aligned}
Q_T(y_\delta) &= \langle Ty_\delta, y_\delta \rangle \\
&= \frac{\langle Tx + \delta Tr, x + \delta r \rangle}{\|x + \delta r\|^2} \\
&= \frac{\langle Tx, x \rangle + \delta\big(\langle Tr, x \rangle + \langle Tx, r \rangle\big) + \delta^2 \langle Tr, r \rangle}{\|x + \delta r\|^2} \\
&= \frac{\langle Tx, x \rangle + 2\delta \langle Tx, r \rangle + \delta^2 \langle Tr, r \rangle}{\|x + \delta r\|^2} && (T \text{ is self-adjoint}) \\
&= \frac{\langle Tx, x \rangle + 2\delta \|r\|^2 + \delta^2 \langle Tr, r \rangle}{\|x\|^2 + \delta^2 \|r\|^2} && (r = Q_{[x]}Tx \perp x).
\end{aligned}$$

Since $\|r\| \neq 0$ by supposition, $Q_T(y_\delta) > \langle Tx, x \rangle / \|x\|^2 = Q_T(x)$ for sufficiently small positive δ.

This shows that λ_1 is an eigenvalue of T. Let M_1 be the associated eigenspace. If $M_1 = V$, then $T = \lambda_1 P_{M_1}$ and we are done. Otherwise note that for $x \in M_1$ and $y \in M_1^\perp$,

$$\langle x, Ty \rangle = \langle Tx, y \rangle = \lambda_1 \langle x, y \rangle = 0,$$

so T maps $M_1^\perp$ into itself. Applying the preceding argument to the restriction T_1 of T to $M_1^\perp$ we find that

$$\lambda_2 = \sup\{\, Q_{T_1}(x) \equiv Q_T(x) : x \in B \text{ and } x \perp M_1 \,\}$$

is an eigenvalue of T_1 (and therefore of T) with some non-trivial eigenspace $M_2 \subset M_1^\perp$. If $M_2 = M_1^\perp$ we have $T = \lambda_1 P_{M_1} + \lambda_2 P_{M_2}$ and are done. Otherwise we recursively define λ_i and M_i for $i = 3, \dots$ by

$$\begin{aligned}
\lambda_i &= \sup\{\, Q_T(x) : x \in B \text{ and } x \perp (M_1 + \cdots + M_{i-1}) \,\} \\
M_i &= \text{eigenspace of } \lambda_i \text{ for } T \text{ restricted to } (M_1 + \cdots + M_{i-1})^\perp,
\end{aligned}$$

stopping the process with the first index k such that $\sum_{1 \le i \le k} M_i$ is V; such a k exists because each new eigenspace adds at least one dimension to the sum. We then have $T = \sum_{1 \le i \le k} \lambda_i P_{M_i}$. ∎

4.9 Exercise [3]. Find the spectral representation (4.5) of the transformation $T: \mathbb{R}^n \to \mathbb{R}^n$ corresponding to the matrix

$$\begin{pmatrix} a & b & b & \cdots & b \\ b & a & b & \cdots & b \\ b & b & a & \cdots & b \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b & b & b & \cdots & a \end{pmatrix}.$$

◇

4.10 Exercise [2]. Use the Spectral Theorem to solve Exercise 2.7. ◇

4.11 Exercise [3]. Use the Spectral Theorem to show that if $\boldsymbol{A}^{n\times n}$ is a symmetric matrix, then there exists an orthogonal matrix $\mathcal{O}^{n\times n}$ such that $\boldsymbol{D} \equiv \mathcal{O}^T \boldsymbol{A} \mathcal{O}$ is diagonal. ◇

4.12 Exercise [3]. Suppose $T: V \to V$ is linear self-adjoint and positive semi-definite ($\langle v, Tv \rangle \geq 0$ for all $v \in V$). Show that there exists a unique linear self-adjoint positive semi-definite transformation $S: V \to V$ such that $T = S^2$. (S is called the (positive) *square root* of T; one writes $S = \sqrt{T}$.) ◇

5. Representation of linear and bilinear functionals

For fixed $v \in V$, the mapping

$$x \to \langle x, v \rangle \quad (x \in V)$$

is a *linear functional* on V, i.e., a linear transformation from V to $\mathbb{R}$. The following result says every linear functional on V is of this form:

5.1 Proposition (Representation theorem for linear functionals). *If ψ is a linear functional on V, then there exists a unique $v \in V$ such that*

$$\psi(x) = \langle x, v \rangle \quad \textit{for all } x \in V. \tag{5.2}$$

The v of (5.2) is called the *coefficient vector* of ψ; one writes $v = c.v.(\psi)$, or just $v = cv(\psi)$.

Proof of the Representation theorem. *Uniqueness:* If v and w are both coefficient vectors for ψ, then their difference is orthogonal to every vector in V and so is 0.

Existence: If the desired representation holds, then the null space of ψ is the orthogonal complement of $[v]$ in V. This observation suggests how to proceed. To avoid trivialities, suppose ψ is not identically 0 and let N be its null space. Since

$$d(N) = d(V) - d\big(\mathcal{R}(\psi)\big) = d(V) - 1,$$

$N^\perp$ is 1-dimensional and so of the form $[w]$ for some $0 \neq w \in V$. The idea is to exhibit the desired coefficient vector as a scalar multiple of w, say dw.

Since the general vector

$$z = y + cw \qquad (y \in N,\ c \in \mathbb{R})$$

of V is mapped by ψ into

$$\psi(z) = \psi(y) + c\psi(w) = c\psi(w)$$

and by $\langle \cdot, dw \rangle$ into

$$\langle z, dw \rangle = \langle y + cw, dw \rangle = cd\langle w, w \rangle,$$

the appropriate choice of d is

$$d = \frac{\psi(w)}{\|w\|^2}\,. \qquad \blacksquare$$

There are two common ways to produce the coefficient vector v of a given linear functional. One way is the *direct-construction method* which constructs v as in the existence part of the proof. The other way is the *confirmation method* which checks that a conjectured v satisfies (5.2) and then appeals to uniqueness.

5.3 Exercise [3]. Suppose $x_1, x_2, \ldots, x_n$ is a basis for V, so that each $x \in V$ has a unique representation of the form

$$x = x_1\beta_1 + x_2\beta_2 + \cdots + x_n\beta_n.$$

Show that the coefficient vector of the so-called j^{th} *coordinate functional*

$$\psi_j : x \to \beta_j$$

is

$$v_j \equiv \left(\frac{Q_j x_j}{\|Q_j x_j\|} \right) \frac{1}{\|Q_j x_j\|} \tag{5.4}$$

where $Q_j = I - P_j$, P_j being projection onto the manifold spanned by the x_i's for $i \neq j$. Solve this exercise twice, once using the direct-construction method and once using the confirmation method. ◇

Suppose now $F: V \times V \to \mathbb{R}$ is a *bilinear functional,* i.e., linear in each component, the other component being held fixed. For each $z \in V$, the map

$$y \to F(y, z)$$

is linear and so of the form

$$F(y, z) = \langle y, Cz \rangle$$

for a unique vector $Cz \in V$. Since

$$\begin{aligned} \langle y, C(c_1 z_1 + c_2 z_2) \rangle &= F(y, c_1 z_1 + c_2 z_2) = c_1 F(y, z_1) + c_2 F(y, z_2) \\ &= \langle y, c_1 C z_1 + c_2 C z_2 \rangle \end{aligned}$$

for all $y \in V$, the mapping $C: V \to V$ is linear. We have established:

5.5 Proposition (Representation theorem for bilinear functionals). *If $F: V \times V \to \mathbb{R}$ is bilinear, then there exists a unique linear transformation $C: V \to V$ such that*

$$F(x, y) = \langle x, Cy \rangle \quad \textit{for all } x \textit{ and } y \in V. \tag{5.6}$$

This result has several important consequences. First, if $T: V \to V$ is linear, then

$$F(x, y) \equiv \langle Tx, y \rangle$$

defines a bilinear functional on V which is necessarily of the form

$$\langle Tx, y \rangle = \langle x, T'y \rangle$$

for some unique linear transformation $T': V \to V$. T' is called the *adjoint* of T. One has

$$(T')' = T, \quad (ST)' = T'S', \tag{5.7}$$

and $T' = T$ if and only if T is self-adjoint in the sense of (4.1). Moreover, T is orthogonal if and only if $T' = T^{-1}$, because the identity

$$\langle Tx, Ty \rangle - \langle x, y \rangle = \langle (T'T - I)x, y \rangle$$

for $x, y \in V$ implies that T preserves inner products if and only if $T'T = I$.

5.8 Exercise [1]. Let T be a linear transformation from V to V, and let $\boldsymbol{A} = (a_{ij})$ be the matrix of T with respect to a given orthonormal basis for V. Show that the matrix of T' is the transpose (a_{ji}) of $\boldsymbol{A}$. ◇

5.9 Exercise [3]. Let T be a linear transformation from V to V. Show that $\mathcal{N}(T') = \big(\mathcal{R}(T)\big)^{\perp}$ and deduce that T' is nonsingular if and only if T is. ◇

5.10 Exercise [3]. Let M be a subspace of V and let $\mathcal{O}: V \to V$ be an orthogonal linear transformation. Show that $\mathcal{O}P_M\mathcal{O}'$ is orthogonal projection onto $\mathcal{O}(M)$. ◇

5.11 Exercise [3]. Locate at least two places where use of the relation $(ST)' = T'S'$ would simplify the calculations in Section 2.2. ◇

Next suppose that $F: V \times V \to \mathbb{R}$ is both bilinear and symmetric. The representing transformation C is then self-adjoint and one can find an orthonormal basis for V consisting of eigenvectors $b_1, \ldots, b_n$ of C with associated eigenvalues $\lambda_1, \ldots, \lambda_n$. F then takes the form

$$F(x, y) = \langle x, Cy \rangle = \Big\langle \sum\nolimits_i c_i b_i, C \sum\nolimits_j d_j b_j \Big\rangle = \sum\nolimits_i \lambda_i c_i d_i, \tag{5.12}$$

where the $c_i = \langle x, b_i \rangle$'s and $d_j = \langle y, b_j \rangle$'s are the coefficients of x and y with respect to the b-basis. Thus when viewed from the right perspective, the general symmetric bilinear functional on V is just a weighted dot-product (the weights can be zero or negative).

5.13 Exercise [3]. A *quadratic form* Q on V is a functional of the form

$$Q(v) = \langle Cv, v\rangle$$

for some linear transformation $C: V \to V$. Show that there is no loss of generality in supposing that C is self-adjoint, in which case C is uniquely determined by Q. [Hint: For uniqueness, generalize (1.6) to

$$\langle Cv, w\rangle = \frac{Q(v+w) - Q(v) - Q(w)}{2}\ .] \qquad (5.14)\ \diamond$$

Finally, suppose $F = [\cdot,\cdot]$ is another inner product on V, so that F is positive-definite in addition to being symmetric and bilinear. We then have

$$[x, y] = \langle x, Cy\rangle \quad \text{for all } x \text{ and } y \text{ in } V \qquad (5.15)$$

for a unique self-adjoint linear transformation C on V which is positive definite in the sense that

$$\langle x, Cx\rangle \geq 0 \text{ for all } x \in V \quad \text{and} \quad \langle x, Cx\rangle = 0 \text{ only for } x = 0.$$

In other words, with respect to an appropriate orthonormal basis $b_1, \ldots, b_n$, $[\cdot,\cdot]$ takes the form

$$[x, y] = \sum\nolimits_i \lambda_i c_i d_i, \quad \text{for } c_i = \langle x, b_i\rangle,\ d_i = \langle y, b_i\rangle \qquad (5.16)$$

with the weights $\lambda_1, \lambda_2, \ldots, \lambda_n$ being strictly positive.

5.17 Exercise [4]. Suppose V and W are both inner product spaces. Formulate and prove a representation theorem for bilinear functionals on $V \times W$. Show that if $T: V \to W$ is linear, then there exists a unique linear transformation $T': W \to V$ such that

$$\langle Tv, w\rangle_W = \langle v, T'w\rangle_V \quad \text{for all } v \in V \text{ and } w \in W; \qquad (5.18)$$

T' is called the *adjoint of* T. Formulate and prove extensions of (5.7) and the assertion of Exercise 5.8. $\diamond$

5.19 Exercise [2]. Suppose M is a subspace of V. Up to now we have considered P_M as a linear transformation from V to V, and as such it is self-adjoint. One can, however, think of P_M as being a linear transformation from V to M. If one takes this point of view, what then is the adjoint of P_M? (The inner product of two elements of M is defined to be their inner product in V.) $\diamond$

5.20 Exercise [3]. Suppose V and W are both inner product spaces and that $T: V \to W$ is linear. Show that

$$\mathcal{R}(T') = \mathcal{R}(T'T) \quad \text{and} \quad \mathcal{N}(T) = \mathcal{N}(T'T). \qquad (5.21)$$

Deduce that $T'T$ is nonsingular if and only if T is. $\diamond$

5.22 Exercise [3]. Suppose B and V are both inner product spaces and that $T\colon B \to V$ is linear. Put $M = T(B) = \{ Tb : b \in B \}$. Show that M is a subspace of V and that for each $v \in V$, one has $P_M v = Tb$ if and only if $b \in B$ satisfies the so-called *normal equations* $T'Tb = Tv$. Do not assume that T is nonsingular. ◇

5.23 Exercise [3]. Suppose $\langle \cdot, \cdot \rangle$ and $[\cdot, \cdot]$ are both inner products on V. Express the unique $[\cdot, \cdot]$-self-adjoint, positive definite linear transformation D such that $\langle x, y \rangle = [x, Dy]$ for all $x, y \in V$ in terms of the C of equation (5.15). ◇

6. Problem set: Cleveland's identity

Let $(V, \langle \cdot, \cdot \rangle)$ be an inner product space. Let $\langle \cdot, \cdot \rangle^* = \langle \cdot, \Delta \cdot \rangle$ be another inner product on V: Δ is a positive definite self-adjoint linear transformation from V to V. For any subspace M of V, let P_M denote projection onto M with respect to $\langle \cdot, \cdot \rangle$, and let P_M^* denote projection onto M with respect to $\langle \cdot, \cdot \rangle^*$. Denote the corresponding residual projections by Q_M and Q_M^* respectively. For any vector z in V, one has of course

$$\|Q_M^* z\|^2 \geq \|Q_M z\|^2.$$

The question of how much larger $\|Q_M^* z\|^2$ can be than $\|Q_M z\|^2$ is addressed by *Cleveland's identity,* which asserts

$$\sup_M \sup_{z \in V, z \notin M} \frac{\|Q_M^* z\|^2}{\|Q_M z\|^2} = \frac{(1+\tau)^2}{4\tau}\,, \tag{6.1}$$

where

$$\tau = \frac{\lambda_n}{\lambda_1} \tag{6.2}$$

is the ratio of the largest eigenvalue λ_n of Δ to the smallest eigenvalue λ_1 of Δ. By the extremal characterization of the eigenvalues of Δ, one has

$$\lambda_1 = \inf_{v \in V, v \neq 0} \left(\frac{\|v\|^*}{\|v\|} \right)^2 \quad \text{and} \quad \lambda_n = \sup_{v \in V, v \neq 0} \left(\frac{\|v\|^*}{\|v\|} \right)^2. \tag{6.3}$$

In this problem set you are asked to establish (6.1). Here are the steps in the argument. To begin with, suppose M and $z \notin M$ are given; one needs to show

$$\frac{\|Q_M^* z\|^2}{\|Q_M z\|^2} \leq \frac{(1+\tau)^2}{4\tau}\,. \tag{6.4}$$

A. Show that without loss of generality, one can consider just the case where $z \perp M$ (with respect to $\langle \cdot, \cdot \rangle$) with $\|z\| = 1$ and where M is one-dimensional, say $M = [y]$ with $\|y\| = 1$.
[Hint: Given z and M such that $P_M z \neq P_M^* z$, find a one-dimensional subspace L of M such that

$$\|Q_M^* z\| = \|Q_L^* u\| \quad \text{and} \quad \|Q_M z\| = \|Q_L u\|$$

for $u = Q_M z$.] ◦

B. Let y and z be as in (**A**). Put $W = [y, z]$. Show that there exists a basis for W, say $\{b_1, b_2\}$, orthonormal with respect to $\langle \cdot, \cdot \rangle$, such that for $v, w \in W$

$$\langle v, w \rangle^* = \beta_1 v_1 w_1 + \beta_2 v_2 w_2$$

where $v_i = \langle v, b_i \rangle$, $w_i = \langle w, b_i \rangle$, and

$$\beta_1 = \inf_{u \in W, u \neq 0} \left(\frac{\|u\|^*}{\|u\|} \right)^2 \quad \text{and} \quad \beta_2 = \sup_{u \in W, u \neq 0} \left(\frac{\|u\|^*}{\|u\|} \right)^2 .$$

[Hint: Restrict $\langle \cdot, \cdot \rangle$ and $\langle \cdot, \cdot \rangle^*$ to $W \times W$.] ◦

C. Let y and z be as in (**A**), b_1, b_2, β_1, β_2 as in (**B**). Show that

$$\frac{\|Q_M^* z\|^2}{\|Q_M z\|^2} = \frac{\beta_1^2 y_1^2 + \beta_2^2 y_2^2}{(\beta_1 y_1^2 + \beta_2 y_2^2)^2} ,$$

where $y_i = \langle y, b_i \rangle$.

[Hint: Compute! Note that the matrix $\begin{pmatrix} y_1 & z_1 \\ y_2 & z_2 \end{pmatrix}$ is orthogonal.] ◦

D. Show that

$$\sup_{p_1, p_2 \geq 0;\, p_1 + p_2 = 1} \frac{\beta_1^2 p_1 + \beta_2^2 p_2}{(\beta_1 p_1 + \beta_2 p_2)^2} = \frac{(\beta_1 + \beta_2)^2}{4 \beta_1 \beta_2}$$

and deduce (6.4). ◦

E. Show that equality holds in (6.1) by exhibiting M and z such that

$$\frac{\|Q_M^* z\|^2}{\|Q_M z\|^2} = \frac{(1 + \tau)^2}{4\tau} .$$ ◦

7. Appendix: Rudiments

This appendix reviews basic definitions and facts about linear algebra with which we presume the reader is acquainted.

7A. Vector spaces

A (real) *vector space* is a triple $(V, +, \cdot)$ consisting of a set V of objects called vectors and two operations, vector addition, "+", which associates to each pair v_1, v_2 of vectors in V a vector $v_1 + v_2$ in V, and scalar multiplication, "·", which associates to each vector $v \in V$ and each scalar $c \in \mathbb{R}$ a vector $c \cdot v \equiv cv$ in V. The operations are assumed to have the following properties:

(1) For all $v_1, v_2, v_3 \in V$, $(v_1 + v_2) + v_3 = v_1 + (v_2 + v_3)$.
(2) For all v_1, $v_2 \in V$, $v_1 + v_2 = v_2 + v_1$.
(3) There exists a vector $0 \in V$ such that $v + 0 = v$ for all $v \in V$.

(4) For each $v \in V$, there exists a vector $-v \in V$ such that $-v + v = 0$. For $v_1 + (-v_2)$ we write also $v_1 - v_2$.
(5) For all $c \in \mathbb{R}$ and $v_1, v_2 \in V$, $c(v_1 + v_2) = cv_1 + cv_2$.
(6) For all $c_1, c_2 \in \mathbb{R}$ and $v \in V$, $(c_1 + c_2)v = c_1 v + c_2 v$.
(7) For all $c_1, c_2 \in \mathbb{R}$ and $v \in V$, $(c_1 c_2)v = c_1(c_2 v)$.
(8) For all $v \in V$, $1 \cdot v = v$.

An important example of a vector space is $\mathbb{R}^n$, the set of n-tuples (or $n \times 1$ column vectors) of real numbers, for which vector addition and multiplication by scalars are defined componentwise as addition and multiplication of real numbers. The *transpose* of a column vector $\boldsymbol{x} \in \mathbb{R}^n$ with components $x_1, \ldots, x_n$ is the row vector $\boldsymbol{x}^T = (x_1, \ldots, x_n)$.

Vectors $v_1, \ldots, v_n$ in a vector space V are said to be *(linearly) independent* if

$$\sum_{1 \le i \le n} c_i v_i = 0$$

implies that the c_i's are all zero; otherwise the v_i's are said to be *(linearly) dependent.* The subset of M of vectors in V of the form $\sum_{1 \le i \le n} c_i v_i$, where each c_i varies over $\mathbb{R}$, is called the *span* of $v_1, \ldots, v_n$, denoted $[v_1, \ldots, v_n]$. Vectors $v_1, \ldots, v_n$ are said to form a *basis* for V if and only if they are linearly independent and V is their span, i.e., if and only if each vector in V has a unique representation of the form $\sum_i c_i v_i$. When $v_1, \ldots, v_n$ is a basis for V, c_i is called the *coordinate of* $\sum_{1 \le j \le n} c_j v_j$ *with respect to* v_i.

V is said to be *finite dimensional* if it has a basis consisting of finitely many elements. Every basis for a finite dimensional vector space has the same number of elements; this number is called the *dimension* of V, denoted $\dim(V)$, or just $d(V)$. If V is an n-dimensional vector space, then $v_1, \ldots, v_n$ is a basis for V if and only if the v_i's are linearly independent, or, equivalently, if and only if they span V. The *usual coordinate basis* for $\mathbb{R}^n$ is comprised of the vectors $\boldsymbol{e}^{(1)}, \ldots, \boldsymbol{e}^{(n)}$ defined by $\boldsymbol{e}_j^{(i)} = \delta_{ij}$.

We assume henceforth that all vector spaces we deal with are finite dimensional.

7B. Subspaces

A nonempty subset M of a vector space V which is closed under vector addition and multiplication by scalars is called a *subspace,* or *(linear) manifold,* of V. For example, the span $[v_1, \ldots, v_m]$ of given vectors $v_1, \ldots, v_m$ is a subspace. We write the trivial subspace $[0]$ simply as 0. A subset F of V of the form $F = v_0 + M = \{ v_0 + m : m \in M \}$, where M is a subspace and $v_0 \in V$, is called a *flat,* or *shifted manifold;* F is a flat if and only if $c_1 f_1 + c_2 f_2 \in F$ whenever $f_1, f_2 \in F$ and $c_1 + c_2 = 1$. The dimension of F is defined to be the dimension of M.

Two subspaces of V are said to be *disjoint* if they have only the zero vector in common (this notion should not be confused with the set-theoretic one). The *intersection,* $M_1 \cap M_2$, of two subspaces is defined to be the subspace of V consisting of all vectors common to M_1 and to M_2. To say $M_1 \cap M_2 = 0$ is to say M_1 and M_2 are disjoint. The *sum* $M_1 + M_2$ of two subspaces of V is defined to be the subspace of V consisting of all vectors of the form $m_1 + m_2$ where $m_1 \in M_1$ and $m_2 \in M_2$. A sum $M_1 + M_2$ is said to be a *direct sum,* denoted $M_1 \oplus M_2$, if M_1 and M_2 are disjoint, i.e., if the representation of $m \in M_1 + M_2$ as $m = m_1 + m_2$ with $m_i \in M_i$ is unique. As regards to dimension, one has

$$d(M_1 + M_2) + d(M_1 \cap M_2) = d(M_1) + d(M_2); \tag{7.1}$$

in particular, $d(M_1 \oplus M_2) = d(M_1) + d(M_2)$.

7C. Linear functionals

A mapping ψ of V into $\mathbb{R}$ is called a *linear functional* on V provided it preserves addition and scalar multiplication: $\psi(c_1v_1 + c_2v_2) = c_1\psi(v_1) + c_2\psi(v_2)$ for all $c_1, c_2 \in \mathbb{R}$ and $v_1, v_2 \in V$. Operations of addition and scalar multiplication on the set V^* of all linear functionals on V are defined pointwise, so that $(\psi_1+\psi_2)(v) = \psi_1(v)+\psi_2(v)$ and $(c\psi)(v) = c\psi(v)$. Under these operations, V^* is itself a vector space, called the *dual* of V. V^* has the same dimension as V. Examples of linear functionals on V are the *coordinate functionals* ψ_i defined relative to a given basis $v_1, \ldots, v_n$ for V by $\psi_i(\sum_{1\le j\le n} c_jv_j) = c_i$ for $i = 1, \ldots, n$.

7D. Linear transformations

A mapping T of a vector space V into a vector space W is called a *linear transformation* if it preserves linear structure: $T(c_1v_1 + c_2v_2) = c_1T(v_1) + c_2T(v_2)$ for all $c_1, c_2 \in \mathbb{R}$ and $v_1, v_2 \in V$. Addition and scalar multiplication of linear transformations from V to W are defined pointwise. A linear functional is simply a linear transformation from V into $\mathbb{R}^1$. A linear transformation $T: V \to W$ is called a *isomorphism* if it is one-to-one and onto. V and W are said to be *isomorphic* if there is an isomorphism mapping one onto the other; isomorphic vector spaces are identical so far as their linear structure is concerned. The *composition* of the linear transformations $S: U \to V$ and $T: V \to W$ is defined to be the linear transformation $TS: U \to W$ which sends $u \in U$ to $T(S(u)) \in W$. A linear transformation $T: V \to V$ such that $T = T^2$ ($\equiv TT$) is said to be *idempotent.* The linear transformation which leaves each $v \in V$ fixed is called the *identity* transformation and is commonly denoted by I, or I_V.

The *range* of a linear transformation T mapping V to W is the subspace of W, denoted $\mathcal{R}(T)$ or $\mathcal{R}(T)$, consisting of all vectors of the form Tv for $v \in V$. The *rank*, $\rho(T)$, of T, is the dimension of $\mathcal{R}(T)$. The *null space* of T is the subspace of V, denoted $\mathcal{N}(T)$ or $\mathcal{N}(T)$, consisting of all vectors in V which are mapped by T into zero in W.

The formula

$$d\big(\mathcal{R}(T)\big) + d\big(\mathcal{N}(T)\big) = d(V), \tag{7.2}$$

holding for any linear transformation T from V to W, has many applications. In particular, if V and W have the same dimension and $\mathcal{N}(T) = 0$, then T is an isomorphism of V and W. A transformation T such that $\mathcal{N}(T) = 0$ is said to be *nonsingular.* The *inverse*, T^{-1}, of a nonsingular transformation exists on $\mathcal{R}(T)$ and is linear.

The *matrix* of a linear transformation $T\colon V \to W$ with respect to bases $v_1, \ldots, v_m$ for V and $w_1, \ldots, w_n$ for W is the $n \times m$ array $[T] \equiv (t_{ij})$ defined by $T(v_j) = \sum_{1\le i\le n} t_{ij} w_i$ for $1 \le j \le m$, or, equivalently, by the condition $T(\sum_j c_j v_j) = \sum_i (\sum_j t_{ij} c_j) w_i$ for $c_1, \ldots, c_m \in V$. If $S\colon U \to V$ and $T\colon V \to W$ are both linear, then the identity $[TS] = [T][S]$ holds relative to fixed bases for U, V, and W.

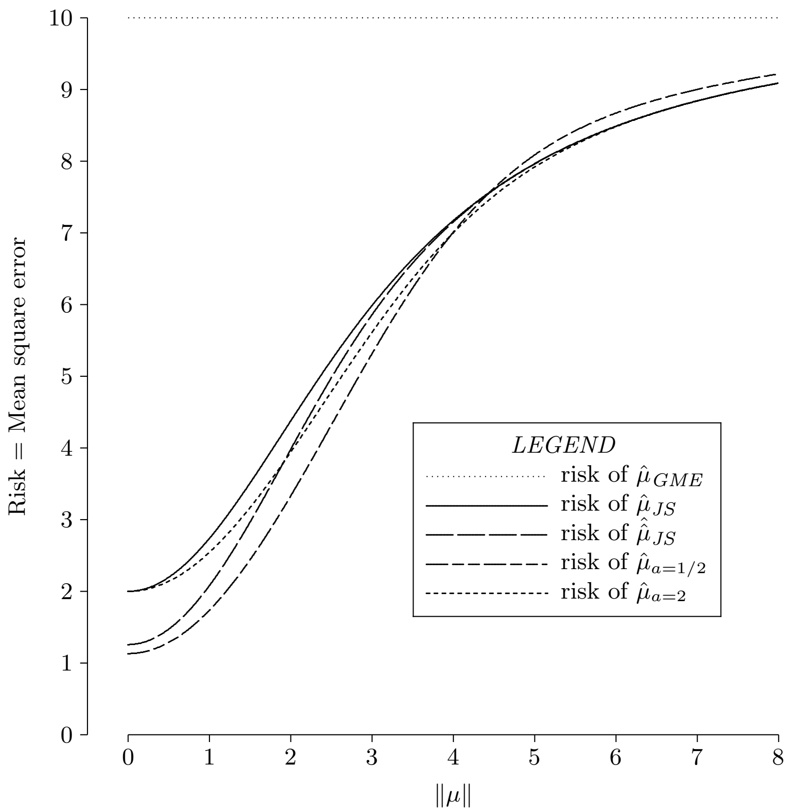
The *transpose* of an $m \times n$ matrix $\boldsymbol{A} = (a_{ij})_{i=1,\ldots,m,\, j=1,\ldots n}$ is the $n \times m$ matrix $\boldsymbol{A}^T = (b_{ji})_{j=1,\ldots,n,\, i=1,\ldots,m}$ where $b_{ji} = a_{ij}$.

(3) $\hat{\mu}_{JS}$ is minimax, since it dominates the minimax estimator $X = P_M Y$. $\hat{\mu}_{JS}$ is not, however, admissible, since by Exercise 4.20 it is itself dominated by the so called *positive-part estimator*

$$\hat{\hat{\mu}}_{JS} \equiv \Big(1 - \frac{p-2}{S}\Big)^{+} X = \max\Big(0, 1 - \frac{p-2}{S}\Big) X, \tag{4.25}$$

which unlike $\hat{\mu}_{JS}$ never reverses the sign of X in estimating μ. It is known that every admissible rule in the problem at hand must be a smooth function of the data, so $\hat{\hat{\mu}}_{JS}$ is also inadmissible; a dominating rule has yet to be found.

4.26 Figure. *For the case $p = 10$ the following diagram displays the risk functions of the the GME $\hat{\mu}_{GME} = X = P_M Y$, the James-Stein estimator $\hat{\mu}_{JS}$ (4.24), the positive part estimator $\hat{\hat{\mu}}_{JS}$ (4.25), and the generalized Bayes estimator $\hat{\mu}_a$ of the next section for the cases $a = 1/2$ and $a = 2$ (see (5.13)). The risk functions depend on μ only through $\|\mu\|$, which is given on the horiontal axis. Evidently all four of the alternatives to $\hat{\mu}_{GME}$ have substantially smaller risk than $\hat{\mu}_{GME}$ for μ's near 0, but there is not much difference in the risks of the alternative estimators. The Bayes estimator $\hat{\mu}_{a=1/2}$ is somewhat preferable to $\hat{\mu}_{JS}$ and $\hat{\hat{\mu}}_{JS}$ in the region where those estimators provide a considerable reduction in risk. $\hat{\mu}_{a=2}$ dominates $\hat{\mu}_{JS}$.*



CHAPTER 9

REFERENCES

The references given here comprise only those works cited in the notes, and papers from which the problem sets were derived.

[1] Baranchik, A. J. (1970). A family of minimax estimator of the mean of a multivariate normal distribution. *Ann. Math. Statist.,* **41,** 642–645.

[2] Berger, J. O. (1976). Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss. *Ann. Statist.,* **4,** 223–226.

[3] Bickel, P. J., and Doksum, K. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics.* Holden Day, San Francisco.

[4] Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Statist.,* **42,** 855–903.

[5] Cleveland, W. S. (1971). Projecting with the wrong inner product and its application to regression with correlated errors and linear filtering of time series. *Ann. Math. Statist.,* **42,** 616–624.

[6] Efron, B., and Morris, C. (1973). Stein's estimation rule and its competitors — An empirical Bayes approach. *J. Amer. Statist. Assoc.,* **68,** 117–130.

[7] Efron, B., and Morris, C. (1976). Families of minimax estimators of the mean of a multivariate normal distribution. *Ann. Statist.,* **4,** 11–21.

[8] Haberman, S. J. (1975). How much do Gauss-Markov and least squares estimates differ? — A coordinate-free approach. *Ann. Statist.,* **3,** 982–990.

[9] Kruskal, W. K. (1960). Untitled. Unpublished monograph on the coordinate-free appraoch to linear models.

[10] Kruskal, W. K. (1961). The coordinate-free approach to Gauss-Markov estimation, and its application to missing and extra observations. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1,** 435–451.

[11] Lehmann, E. L. (1959). *Testing Statistical Hypotheses.* Wiley, New York.

[12] Rao, C. R. (1965). *Linear Statistical Inference and Its Applications.* Wiley, New York.

[13] Stein, C. (1966). An approach to the recovery of inter-block information in balanced incomplete designs, in *Festschrift for J. Neyman,* 351–366, Wiley, New York.

[14] Stein, C. (1973). Estimation of the mean of a multivariate normal distribution. *Proc. Prague. Symp. Asymptotic Statist.,* 345–381.

[15] Strawderman, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Statist.,* **42,** 385–388.

[16] Wald, A. (1942). On the power function of the analysis of variance test. *Ann. Math. Statist.,* **13,** 434–439.

APPENDIX 1

INDEX