All Models Are Wrong

Concepts of Statistical Learning (CSL)

by Gaston Sanchez, and Ethan Marzban

2019-11-02

Contents

Part I

Welcome

Preface

This is a work in progress for an introductory text about concepts of Statistical Learning, covering common supervised as well as unsupervised methods.

How to cite this book:

Sanchez, G., Marzban E. (2019) **All Models Are Wrong: Concepts of Statistical Learning**. https://allmodelsarewrong.github.io

Chapter 1

About this book

Knowing that the field(s) of Machine Learning, Statisticial Learning, and any other name about learning from data, is a very broad subject, we should warn you that this book is not intended to be the ultimate compilation of every single SL technique ever devised.

Instead, we focus on the concepts that we consider the building blocks that any user or practitioner needs to make sense of most common SL techniques.

A big shortcoming of the book: we don't cover neural networks. At least not in this first round of iterations. Sorry!

On the plus side: We've tried hard to keep the notation as simple and consistent as possible. And we've also made a serious effort to make it very visual (lots of diagrams, pictures, plots, graphs, figures, ..., you name it).

1.1 Prerequisites

We are assuming that you already have some knowledge under your belt.

You will better understand (and hopefully enjoy) the book if you've taken one or more courses on the following subjects:

- linear or matrix algebra
- multivariable calculus
- statistics
- probability
- programming or scripting

1.2 Acknowledgements

Many thanks to the UC Berkeley students of Stat 154 Modern Statistical Prediction and Machine Learning (Fall 2017, Spring 2018, Fall 2019).

Part II

Intro

Chapter 2

Introduction

Picture a data set containing scores of several course for college students. For example, courses like matrix algebra, multivariable calculus, statistics, and probability. And say we also have historical data about a course in Statistical Learning. In particular we have final scores measured on a scale from 0 to 100, we also have final grades (letter grade scale), as well as a third interesting variable "Pass - Non-Pass" indicating whether the student passed statistical learning. Some data like that fits perfectly well in a tabular format. The rows contain the records for a bunch of students, and the columns refer to the variables.

Math 54	Math 55	Stat 135	Stat 134	Stat 154	Grade	P/NP
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	y_{15}	y_{16}
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	y_{15}	y_{16}
...	...	...	...	...	...	...
x_{n1}	x_{n2}	x_{n3}	x_{n4}	x_{n5}	y_{n5}	y_{n6}

Suppose that, based on this historical data, we wish to predict the score of a new student (whose Math 54, Math 55, and Stat 135 grades are known) in Stat 154.

To do so, we would fit some sort of model to our data; i.e. we would perform regression. This is a form of supervised learning, since our model is trained using known inputs (i.e. Math 54, Math 55, and Stat 135 grades) as well as known responses (i.e. the Stat 154 grades of the previous students).

Unsupervised Learning: where we have inputs, but not response variables.

2.1 Basic Notation

In this book we are going to use a fair amount of math notation. Becoming familiar with the meaning of all the different symbols as soon as possible, should allow you to keep the learning curve a little bit less steep.

The starting point is always the data, which we will assume to be in a tabular format, that can be translated into a mathematical matrix object. Here's an example of a data matrix $\mathbf{X}$ of size $n \times p$

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

By default, we will assume that the rows of a data matrix correspond to the individuals or objects. Likewise, we will also assume that the columns of a data matrix correspond to the variables or features observed on the individuals. In this sense, the symbol x_{ij} represents the value observed for the j-th variable on the i-th individual.

Throughout this book, every time you see the letter i, either alone or as an index associated with any other symbol (superscript or subscript), it means that such term corresponds to an individual or a row of some data matrix. For instance, symbols like x_i, $\mathbf{x_i}$, and α_i are all examples that refer to—or denote a connection with—individuals.

In turn, we will always use the letter j to conveyed association with variables. For instance, x_j, $\mathbf{x_j}$, and α_j are examples that refer to—or denote a connection with—variables.

For better or for worse, we've made the decision to represent row vectors and column vectors with the same notation: as bold lower case letters such as $\mathbf{x_i}$ and $\mathbf{x_j}$. Because of the risk of confusing a row vector with a column vector, sometimes we will use the arrow notation for row vectors: $\tilde{\mathbf{x}}_\mathbf{i}$.

So, going back to the above data matrix $\mathbf{X}$, we can represent the first variable as a vector $\mathbf{x_1} = (x_{11}, x_{21}, \ldots, x_{n1})$. Likewise, we can represent the first individual with the vector $\tilde{\mathbf{x}}_\mathbf{1} = (x_{11}, x_{12}, \ldots, x_{1n})$.

Here's a reference table with the notation and symbols used throughout the book

Symbol	Description
n	number of objects
p	number of variables
i	running index for rows

Symbol	Description
j	running index for columns
k	running index for sets of variables (i.e. blocks)
l, m, q	auxiliar index
$f()$, $g()$, $h()$	functions
$\lambda, \mu, \gamma, \alpha$	greek letters represent scalars
ε	decimal tolerance threshold (e.g. 0.00001)
$\mathbf{x}$, $\mathbf{y}$	variables, size determined by context
$\mathbf{w}$, $\mathbf{a}$, $\mathbf{b}$	vectors of weight coefficients
$\mathbf{z}$, $\mathbf{t}$, $\mathbf{u}$	components or latent variables
$\mathbf{X} : n \times p$	data matrix with n rows and p columns
x_{ij}	element of a matrix in i-th row and j-th column
$\mathbf{1}$	vector of ones, size determined by context
$\mathbf{I}$	identity matrix, size determined by context

By the way, there are many more symbols that will appear in later chapters. But for now these are the fundamental ones.

Common operators

Symbol	Description
$\mathbb{E}[X]$	expected value of a random variable X
$\|\mathbf{a}\|$	euclidean norm of a vector
$\mathbf{a}^\mathsf{T}$	transpose of a vector (or matrix)
$\mathbf{a}^\mathsf{T}\mathbf{b}$	inner product of two vectors
$\langle \mathbf{a}, \mathbf{b} \rangle$	inner product of two vectors
$det(\mathbf{A})$	determinant of a square matrix
$tr(\mathbf{A})$	trace of a square matrix
$\mathbf{A}^{-1}$	inverse of a square matrix
$diag(\mathbf{A})$	diagonal of a square matrix
$var()$	variance
$cov()$	covariance

2.2 About Statistical Learning

We will focus on supervised learning as well as unsupervised learning. We won't discuss more recent fields of deep learning and reinforcement learning.

To visualize the different types of learning, the different types of variables, and the methodology associated with each combination of learning/data types, we can use the following graphic:

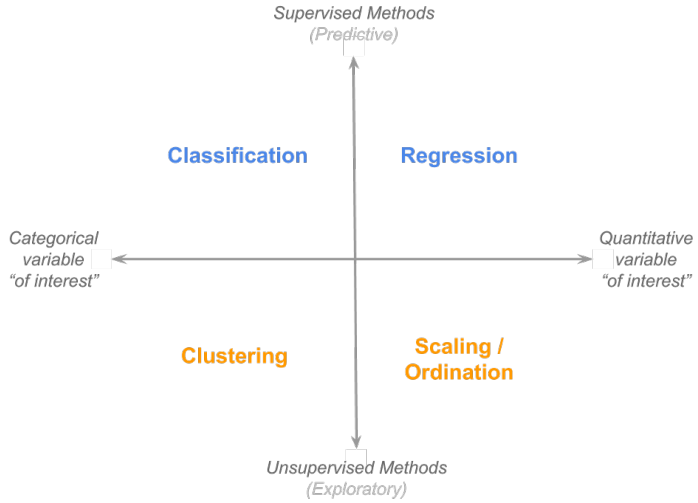

Figure 2.1: Supervised and Unsupervised

Chapter 3

Geometric Duality

The way we like to introduce you to the statistical learning world, is by talking and thinking about data in a geometric sense.

Let's suppose we have some data in the form of a data matrix. For convenience purposes, let's also suppose that all variables are measured in a real-value scale. Obviously not all data is expressed or even encoded numerically. You may have categorical or symbolic data. But for this illustration, let's assume that any categorical and symbolic data has already been transformed into a numeric scale.

It's very enlightening to think of a data matrix as viewed from the glass of Geometry. The key idea is to think of the data in a matrix as elements living in a multidimensional space. Actually, we can regard a data matrix from two apparently different perspectives that, in reality, are intimately connected: the **rows perspective** and the **columns perspective**. In order to explain these perspectives, let me use the following diagram of a data matrix $\mathbf{X}$ with n rows and p columns, with x_{ij} representing the element in the i-th row and j-th column.

When we look at a data matrix from the *columns* perpective what we are doing is focusing on the p variables. In a similar way, when looking at a data matrix from its *rows* perspective, we are focusing on the n individuals. Like a coin, though, this matrix has two sides: a rows side, and a columns side. That is, we could look at the data from the rows point of view, or the columns point of view. These two views are (of course) not completely independent. This double perspective or **duality** for short, is like the two sides of the same coin.

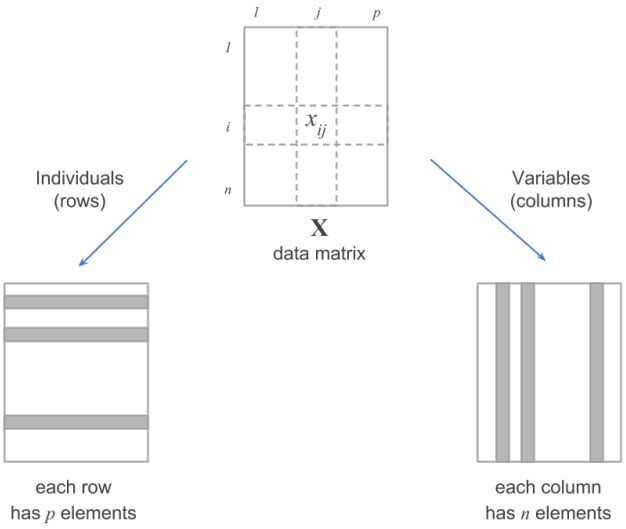

Figure 3.1: Duality of a data matrix

3.1 Rows Space

We know that human vision is limited to three-dimensions, but pretend that you had superpowers that let you visualize a space with any number of dimensions.

Because each row of the data matrix has p elements, we can regard individuals as objects that live in a p-dimensional space. For visualization purposes, think of each variable as playing the role of a dimension associated to a given axis in this space; likewise, consider each of the n individuals as being depicted as a point (or particle) in such space, like in the following diagram:

In the figure above, even though I'm showing only three axes, you should pretend that you are visualizing a p-dimensional space (imaging that there are p axes). Each point in this space corresponds to a single individual, and they all form what you can call a *cloud of points*.

3.2 Columns Space

We can do the same visual exercise with the columns of a data matrix. Since each variable has n elements, we can regard the set of p variables as objects that live in an n-dimensional space. However, instead of representing each variable with a dot, it's better to graphically represent them with an arrow (or vector). Why? Because of two reasons: one is to distinguish them from the individuals (dots). But more important, because the esential thing with a variable is not

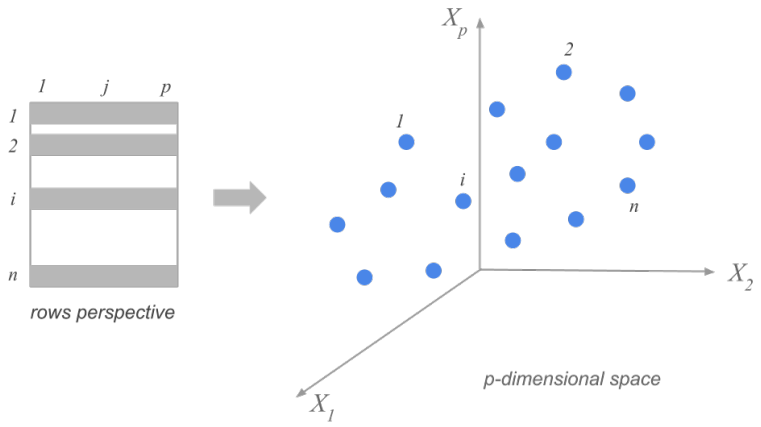

Figure 3.2: Rows space

really its magnitude (and therefore its position) but its direction. Often, as part of the preprocessing steps we apply transformations on variables that change their scales (e.g. shrinking them, or stretching them) without modifying their directions.

Analogously to the rows space and its cloud of individuals, you should also pretend that the image above is displaying an n-dimensional space with a bunch of blue arrows pointing in various directions.

What's next?

Now that we know how to think of data from a geometric perspective, the next step is to discuss a handful of common operations that can be performed with points and vectors that live in some geometric space.

3.3 Cloud of Individuals

In the previous chapter we introduce the powerful idea of looking at the rows and columns of a data matrix from the lens of geometry. We are assuming in general that the rows have to do with n individuals that lie in a p-dimensional space.

Let's start describing a set of common operations that we can apply on the individuals (living in a p-dimensional space).

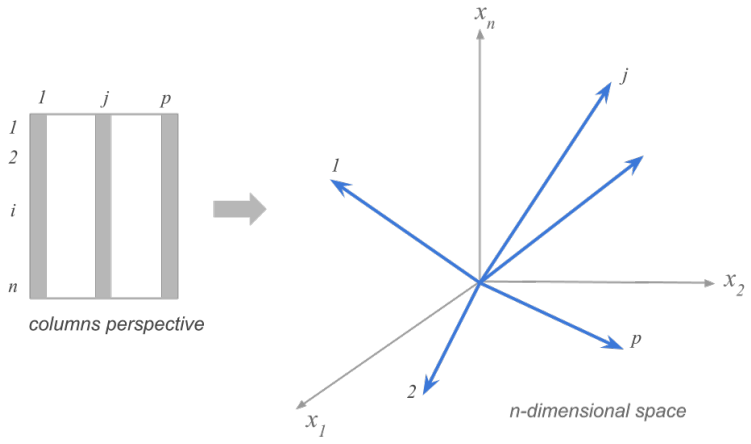

Figure 3.3: Columns space

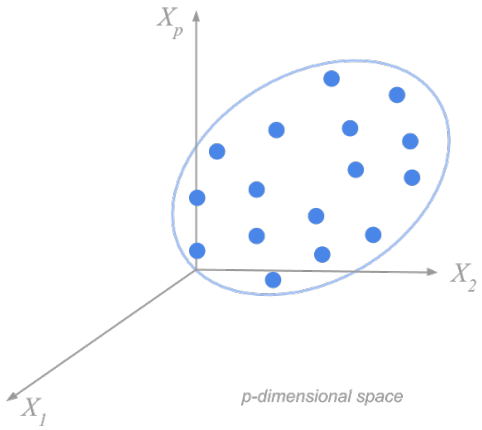

Figure 3.4: Cloud of points

Figure 3.5: Points in one dimension

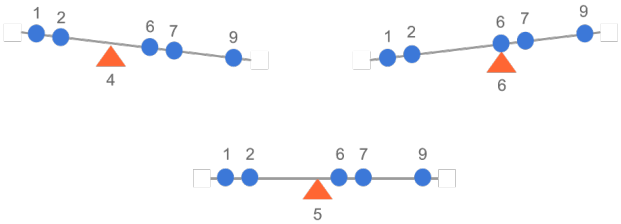

Figure 3.6: Average individual

3.3.1 Average Individual

We can ask about the typical or average individual.

If you only have one variable, then all the individual points lie in a one-dimensional space, which is basically a line:

In this case, the average individual is simply the average of the values, which geometrically corresponds to the balancing point:

Algebraically we have: individuals $x_1, x_2, \ldots, x_n$, and the average is:

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

In vector notation, the average can be calculated with an inner product between $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, and a constant vector of n-ones $\mathbf{1}$:

$$\bar{x} = \mathbf{x}^\mathsf{T}\mathbf{1}$$

What about the multivariate case? It turns out that we can also ask about the average individual of a cloud of points, like in the following figure:

The average individual, in a p-dimensional space is the point $\tilde{\mathbf{g}}$ containing as coordiantes the averages of all the variables:

$$\tilde{\mathbf{g}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_j)$$

where $\bar{x}_j$ is the average of the j-th variable.

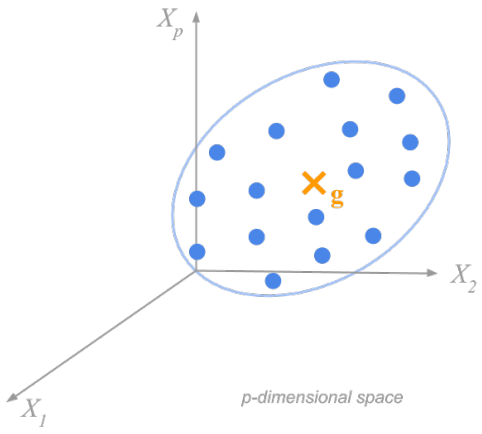

Figure 3.7: Cloud of points with centroid (i.e. average individual)

This average individual $\tilde{\mathbf{g}}$ is also known as the **centroid**, *barycenter*, or *center of gravity* of the cloud of points.

3.3.2 Centered Data

Often, it is convenient to transform the data in such a way that the centroid of a data set becomes the origin of the cloud of points. Geometrically, this type of transformation involves a shif of the axes in the p-dimensional space. Algebraically, this transformation corresponds to expresing the values of each variable in terms of deviations from their means.

3.3.3 Distance between individuals

Another common operation that we may be interested in is the distance between two individuals. Obviously the notion of distance is not unique, since you can choose different types of distance measures. Perhaps the most comon type of distance is the (squared) Euclidean distance. Unless otherwise mentioned, this will be the default distance used in this book.

If you have one variable X, then the squared distance $d^2(i,l)$ between two individuals x_i and x_l is:

$$d^2(i,l) = (x_i - x_l)^2$$

In general, with p variables, the squared distance between the i-th individual and the l-th individual is:

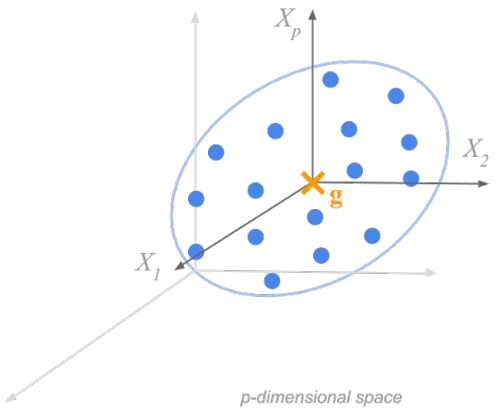

Figure 3.8: Cloud of points of mean-centered data

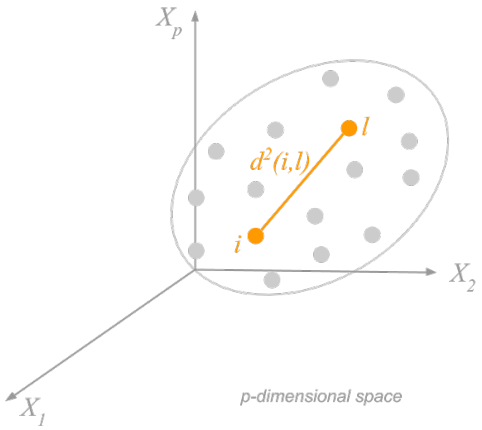

Figure 3.9: Distance between two individuals

$$\begin{aligned} d^2(i,l) &= (x_{i1} - x_{l1})^2 + (x_{i2} - x_{l2})^2 + \cdots + (x_{ip} - x_{lp})^2 \\ &= (\tilde{\mathbf{x}}_\mathbf{i} - \tilde{\mathbf{x}}_\mathbf{l})^\mathsf{T}(\tilde{\mathbf{x}}_\mathbf{i} - \tilde{\mathbf{x}}_\mathbf{l}) \end{aligned}$$

3.3.4 Distance to the centroid

A special case is the distance between any individual i and the average individual:

$$\begin{aligned} d^2(i,g) &= (x_{i1} - \bar{x}_1)^2 + (x_{i2} - \bar{x}_2)^2 + \cdots + (x_{ip} - \bar{x}_p)^2 \\ &= (\tilde{\mathbf{x}}_\mathbf{i} - \tilde{\mathbf{g}})^\mathsf{T}(\tilde{\mathbf{x}}_\mathbf{i} - \tilde{\mathbf{g}}) \end{aligned}$$

3.3.5 Measures of Dispersion

What else can we calculate with the individuals? Think about it. So far we've seen how to calculate the average individual, as well as distances between individuals. The average individual or centroid plays the role of a measure of center. And everytime you get a measure of center, it makes sense to get a measure of spread.

Overall Dispersion

One way to compute a measure of scatter among individuals is to consider all the squared distances between pairs of individuals. For instance, say you have three individuals a, b, and c. We can calculate all pairwise distances and add them up:

$$d^2(a,b) + d^2(b,a) + d^2(a,c) + d^2(c,a) + d^2(b,c) + d^2(c,b)$$

In general, when you have n individuals, you can obtain up to n^2 squared distances. We will give the generic name of **Overall Dispersion** to the sum of all squared pairwise distances:

$$\text{overall dispersion} = \sum_{i=1}^{n}\sum_{l=1}^{n} d^2(i,l)$$

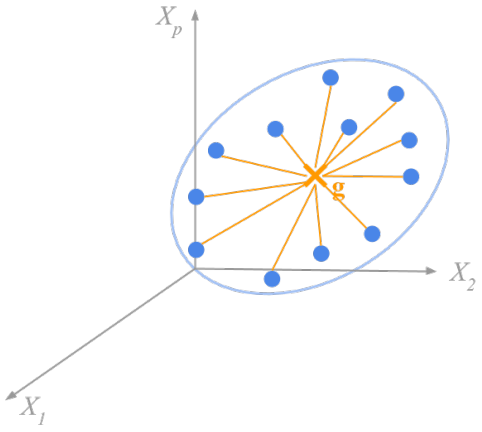

Figure 3.10: Inertia

Inertia

Another measure of scatter among individuals can be computed by adding the distances between all individuals and the centroid.

The sum of squared distances from each point to the centroid then becomes

$$\frac{1}{n}\sum_{i=1}^{n} d^2(i,g) = \frac{1}{n}\sum_{i=1}^{n}(\tilde{\mathbf{x}}_\mathbf{i} - \tilde{\mathbf{g}})^\mathsf{T}(\tilde{\mathbf{x}}_\mathbf{i} - \tilde{\mathbf{g}})$$

We will name this measure **Inertia**, borrowing this term from the concept of inertia used in mechanics (in physics).

$$\text{Inertia} = \frac{1}{n}\sum_{i=1}^{n} d^2(i,g)$$

What is the motivation behind this measure? Consider the $p = 1$ case; i.e. when $\mathbf{X}$ is simply a column vector

$$\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

The centroid will simply be the mean of these points: i.e. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

The sum of squared-distances from each point to the centroid then becomes:

$$(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Does the above formula look familiar? What if we take the average of the squared distances:

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$$

Same question: Do you recognize this formula? You better do… This is nothing else than the formula of the variance of X. And yes, we are dividing by n (not by $n - 1$). Hence, you can think of inertia as a multidimensional extension of variance, which gives the typical squared distance around the centroid.

Overall Dispersion and Inertia

Interestingly, the *overall dispersion* and the *inertia* are connected through the following relation:

$$\begin{aligned} \text{overall dispersion} &= \sum_{i=1}^{n} \sum_{l=1}^{n} d^2(i, l) \\ &= 2n \sum_{i=1}^{n} d^2(i, g) \\ &= (2n^2)\text{Inertia} \end{aligned}$$

The proof of this relation is left as a homework exercise.

3.4 Cloud of Variables

The starting point when analyzing variables involves computing various summary measures—such as means, and variances—to get an idea of the common or central values, and the amount of variability of each variable. In this chapter we will review how concepts like the mean of a variable, the variance, covariance, and correlation, can be interpreted in a geometric sense, as well as their expressions in terms of vector-matrix operations.

3.4.1 Mean of a Variable

To measure variation, we usually begin by calculating a "typical" value. The idea is to summarize the values of a variable with one or two representative values. You will find this notion under several terms like measures of center, location, central tendency, or centrality.

The prototypical summary value of center is the **mean**, sometimes referred to as average. The mean of an $n-$element variable $X = (x_1, x_2, \ldots, x_n)$, represented by $\bar{x}$, is obtained by adding all the x_i values and then dividing by their total number n:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Using summation notation we can express $\bar{x}$ in a very compact way as:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

If you associate a constant weight of $1/n$ to each observation x_i, you can look at the formula of the mean as a weighted sum:

$$\bar{x} = \frac{1}{n} x_1 + \frac{1}{n} x_2 + \cdots + \frac{1}{n} x_n$$

This is a slightly different way of looking at the mean that will allow you to generalize the concept of an "average" as a *weighted aggregation of information*. For example, if we denote the weight of the i-th individual as w_i, then the average can be expressed as:

$$\bar{x} = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n = \sum_{i=1}^{n} w_i x_i$$

3.4.2 Variance of a Variable

A measure of center such as the mean is not enoguh to summarize the information of a variable. We also need a measure of the amount of variability. Synonym terms are variation, spread, scatter, and dispersion.

Because of its relevance and importance for statistical learning methods, we will focus on one particular measure of spread: the **variance** (and its square root the standard deviation).

Simply put, the variance is a measure of spread around the mean. The main idea behind the calculation of the variance is to quantify the typical concentration

of values around the mean. The way this is done is by averaging the squared deviations from the mean.

$$var(X) = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Let's disect the terms and operations involved in the formula of the variance.

- the main terms are the *deviations from the mean* $(x_i - \bar{x})$, that is, the difference between each observation x_i and the mean $\bar{x}$.
- conceptually speaking, we want to know what is the average size of the deviations around the mean.
- simply averaging the deviations won't work because their sum is zero (i.e. the sum of deviations around the mean will cancel out because the mean is the balancing point).
- this is why we square each deviation: $(x_i - \bar{x})^2$, which literally means getting the squared distance from x_i to $\bar{x}$.
- having squared all the deviations, then we average them to get the variance.

Because the variance has squared units, we need to take the square root to "recover" the original units in which X is expressed. This gives us the **standard deviation**

$$sd(X) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

In this sense, you can say that the standard deviation is roughly the average distance that the data points vary from the mean.

Sample Variance

In practice, you will often find two versions of the formula for the variance: one in which the sum of squared deviations is divided by n, and another one in which the division is done by $n-1$. Each version is associated to the statistical inference view of variance in terms of whether the data comes from the *population* or from a *sample* of the population.

The *population variance* is obtained dividing by n:

$$\text{population variance:} \quad \frac{1}{(n)} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

The *sample variance* is obtained dividing by $n-1$ instead of dividing by n. The reason for doing this is to get an unbiased estimor of the population variance:

$$\text{sample variance:} \quad \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

It is important to note that most statistical software compute the variance with the unbiased version. If you implement your own functions and are planning to compare them against other software, then it is crucial to known what other programmers are using for computing the variance. Otherwise, your results might be a bit different from the ones with other people's code.

In this book, unless indicated otherwise, we will use the factor $\frac{1}{n}$ when introducing concepts of variance, and related measures. If needed, we will let you know when a formula needs to use the factor $\frac{1}{n-1}$.

3.4.3 Variance with Vector Notation

In a similar way to expressing the mean with vector notation, you can also formulate the variance in terms of vector-matrix notation. First, notice that the formula of the variance consists of the addition of squared terms. Second, recall that a sum of numbers can be expressed with an inner product by using the unit vector (or summation operator). If we denote a vector of ones of size n as $\mathbf{1}_n$, then the variance of a vector $\mathbf{x}$ can be obtained with the following inner product:

$$var(\mathbf{x}) = \frac{1}{n} (\mathbf{x} - \bar{\mathbf{x}})^\mathsf{T} (\mathbf{x} - \bar{\mathbf{x}})$$

where $\bar{\mathbf{x}}$ is an n-element vector of mean values $\bar{x}$.

Assuming that $\mathbf{x}$ is already mean-centered, then the variance is proportional to the squared norm of $\mathbf{x}$

$$var(\mathbf{x}) = \frac{1}{n} \mathbf{x}^\mathsf{T} \mathbf{x} = \frac{1}{n} \|\mathbf{x}\|^2$$

This means that we can formulate the variance with the general notion of an inner product:

$$var(\mathbf{x}) = \frac{1}{n} \langle \mathbf{x}, \mathbf{x} \rangle$$

3.4.4 Standard Deviation as a Norm

If we use a metric matrix $\mathbf{D} = diag(1/n)$ then we have that the variance is given by a special type of inner product:

$$var(\mathbf{x}) = \langle \mathbf{x}, \mathbf{x} \rangle_D = \mathbf{x}^\mathsf{T} \mathbf{D} \mathbf{x}$$

From this point of view, we can say that the variance of $\mathbf{x}$ is equivalent to its squared norm when the vector space is endowed with a metric $\mathbf{D}$. Consequently, the standard deviation is simply the length of $\mathbf{x}$ in this particular geometric space.

$$sd(\mathbf{x}) = \|\mathbf{x}\|_D$$

When looking at the standard deviation from this perspective, you can actually say that the amount of spread of a vector $\mathbf{x}$ is actually its length (under the metric $\mathbf{D}$).

3.4.5 Covariance

The covariance generalizes the concept of variance for two variables. Recall that the formula for the covariance between $\mathbf{x}$ and $\mathbf{y}$ is:

$$cov(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

where $\bar{x}$ is the mean value of $\mathbf{x}$ obtained as:

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

and $\bar{y}$ is the mean value of $\mathbf{y}$:

$$\bar{y} = \frac{1}{n}(y_1 + y_2 + \cdots + y_n) = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Basically, the covariance is a statistical summary that is used to assess the **linear association between pairs of variables**.

Assuming that the variables are mean-centered, we can get a more compact expression of the covariance in vector notation:

$$cov(\mathbf{x}, \mathbf{y}) = \frac{1}{n}(\mathbf{x}^\mathsf{T}\mathbf{y})$$

Properties of covariance:

- the covariance is a symmetric index: $cov(X, Y) = cov(Y, X)$
- the covariance can take any real value (negative, null, positive)
- the covariance is linked to variances under the name of the Cauchy-Schwarz inequality:

$$cov(X, Y)^2 \leq var(X)var(Y)$$

3.4.6 Correlation

Although the covariance indicates the direction—positive or negative—of a possible linear relation, it does not tell us how big or small the relation might be. To have a more interpretable index, we must transform the convariance into a unit-free measure. To do this we must consider the standard deviations of the variables so we can normalize the covariance. The result of this normalization is the coefficient of linear correlation defined as:

$$cor(X, Y) = \frac{cov(X, Y)}{\sqrt{var(X)}\sqrt{var(Y)}}$$

Representing X and Y as vectors $\mathbf{x}$ and $\mathbf{y}$, we can express the correlation as:

$$cor(\mathbf{x}, \mathbf{y}) = \frac{cov(\mathbf{x}, \mathbf{y})}{\sqrt{var(\mathbf{x})}\sqrt{var(\mathbf{y})}}$$

Assuming that $\mathbf{x}$ and $\mathbf{y}$ are mean-centered, we can express the correlation as:

$$cor(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^\mathsf{T}\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$$

As it turns out, the norm of a mean-centered variable $\mathbf{x}$ is proportional to the square root of its variance (or standard deviation):

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}^\mathsf{T}\mathbf{x}} = \sqrt{n}\sqrt{var(\mathbf{x})}$$

Consequently, we can also express the correlation with inner products as:

$$cor(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^\mathsf{T}\mathbf{y}}{\sqrt{(\mathbf{x}^\mathsf{T}\mathbf{x})}\sqrt{(\mathbf{y}^\mathsf{T}\mathbf{y})}}$$

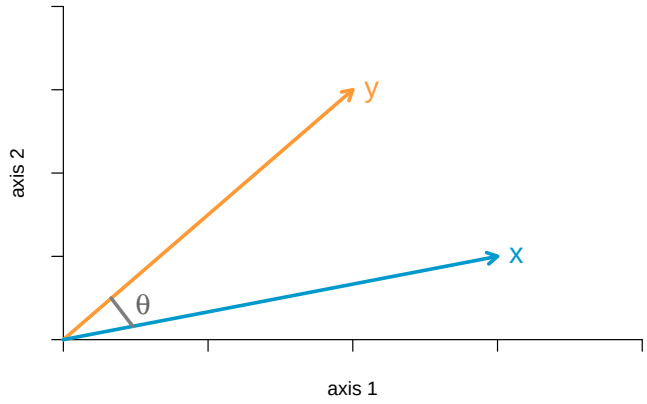

Figure 3.11: Two vectors in a 2-dimensional space

or equivalently:

$$cor(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^\mathsf{T}\mathbf{y}}{\|\mathbf{x}\| \, \|\mathbf{y}\|}$$

In the case that both $\mathbf{x}$ and $\mathbf{y}$ are standardized (mean zero and unit variance), that is:

$$\mathbf{x} = \begin{bmatrix} \frac{x_1 - \bar{x}}{\sigma_x} \\ \frac{x_2 - \bar{x}}{\sigma_x} \\ \vdots \\ \frac{x_n - \bar{x}}{\sigma_x} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} \frac{y_1 - \bar{y}}{\sigma_y} \\ \frac{y_2 - \bar{y}}{\sigma_y} \\ \vdots \\ \frac{y_n - \bar{y}}{\sigma_y} \end{bmatrix}$$

the correlation is simply the inner product:

$$cor(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\mathsf{T}\mathbf{y} \quad \text{(standardized variables)}$$

3.4.7 Geometry of Correlation

Let's look at two variables (i.e. vectors) from a geometric perspective.

The inner product ot two mean-centered vectors $\langle \mathbf{x}, \mathbf{y} \rangle$ is obtained with the following equation:

$$\mathbf{x}^\mathsf{T}\mathbf{y} = \|\mathbf{x}\| \, \|\mathbf{y}\| \, cos(\theta_{x,y})$$

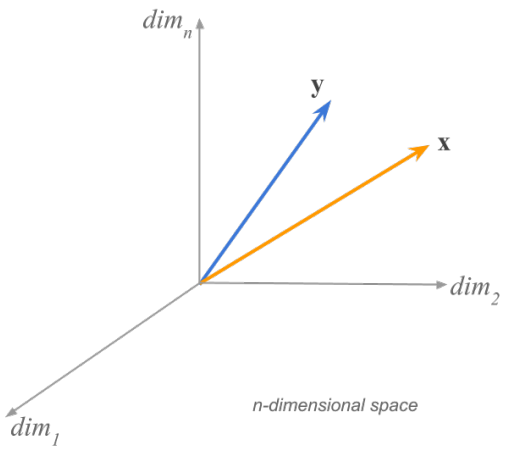

Figure 3.12: Two vectors in n-dimensional space

where $cos(\theta_{x,y})$ is the angle between $\mathbf{x}$ and $\mathbf{y}$. Rearranging the terms in the previous equation we get that:

$$cos(\theta_{x,y}) = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} = cor(\mathbf{x}, \mathbf{y})$$

which means that the correlation between mean-centered vectors $\mathbf{x}$ and $\mathbf{y}$ turns out to be the cosine of the angle between $\mathbf{x}$ and $\mathbf{y}$.

3.4.8 Orthogonal Projections

Last but not least, we finish this chapter with a discussion of projections. To be more specific, the statistical interpretation of orthogonal projections.

Let's motivate this discussion with the following question: Consider two variables $\mathbf{x}$ and $\mathbf{y}$. Can we approximate one of the variables in terms of the other? This is an asymmetric type of association since we seek to say something about the variability of one variable, say $\mathbf{y}$, in terms of the variability of $\mathbf{x}$.

We can think of several ways to approximate $\mathbf{y}$ in terms of $\mathbf{x}$. The approximation of $\mathbf{y}$, denoted by $\hat{\mathbf{y}}$, means finding a scalar b such that:

$$\hat{\mathbf{y}} = b\mathbf{x}$$

The common approach to get $\hat{\mathbf{y}}$ in some optimal way is by minimizing the square difference between $\mathbf{y}$ and $\hat{\mathbf{y}}$.

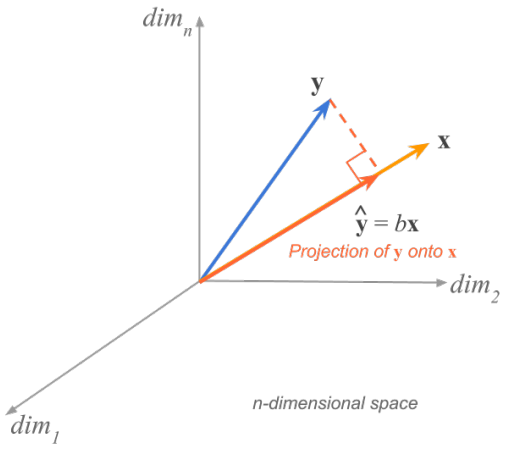

Figure 3.13: Orthogonal projection of y onto x

The answer to this question comes in the form of a projection. More precisely, we orthogonally project $\mathbf{y}$ onto $\mathbf{x}$:

$$\hat{\mathbf{y}} = \mathbf{x}\left(\frac{\mathbf{y}^\mathsf{T}\mathbf{x}}{\mathbf{x}^\mathsf{T}\mathbf{x}}\right)$$

or equivalently:

$$\hat{\mathbf{y}} = \mathbf{x}\left(\frac{\mathbf{y}^\mathsf{T}\mathbf{x}}{\|\mathbf{x}\|^2}\right)$$

For convenience purposes, we can rewrite the above equation in a slightly different format:

$$\hat{\mathbf{y}} = \mathbf{x}(\mathbf{x}^\mathsf{T}\mathbf{x})^{-1}\mathbf{x}^\mathsf{T}\mathbf{y}$$

If you are familiar with linear regression, you should be able to recognize this equation. We'll come back to this when we get to the chapter about Linear regression.

3.4.9 The mean as an orthogonal projection

Let's go back to the concept of mean of a variable. As we previously mention, a variable $X = (x_1, \ldots, x_n)$, can be thought of a vector $\mathbf{x}$ in an n-dimensional

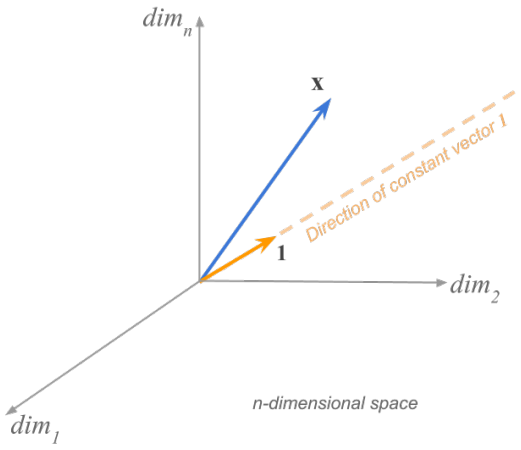

Figure 3.14: Two vectors in n-dimensional space

space. Furthermore, let's also consider the constant vector $\mathbf{1}$ of size n. Here's a conceptual diagram for this situation:

Out of curiosity, what happens when we ask about the orthogonal projection of $\mathbf{x}$ onto $\mathbf{1}$? Something like in the following picture:

This projection is expressed in vector notation as:

$$\hat{\mathbf{x}} = \mathbf{1}\left(\frac{\mathbf{x}^\mathsf{T}\mathbf{1}}{\mathbf{1}^\mathsf{T}\mathbf{1}}\right)$$

or equivalently:

$$\hat{\mathbf{x}} = \mathbf{1}\left(\frac{\mathbf{x}^\mathsf{T}\mathbf{1}}{\|\mathbf{1}\|^2}\right)$$

Note that the term in parenthesis is just a scalar, so we can actually express $\hat{\mathbf{x}}$ as $b\mathbf{1}$. This means that a projection implies multiplying $\mathbf{1}$ by some number b, such that $\hat{\mathbf{x}} = b\mathbf{1}$ is a stretched or shrinked version of $\mathbf{1}$. So, what is the scalar b? It is simply the mean of $\mathbf{x}$:

$$\hat{\mathbf{x}} = \mathbf{1}\left(\frac{\mathbf{x}^\mathsf{T}\mathbf{1}}{\|\mathbf{1}\|^2}\right) = \bar{x}\mathbf{1}$$

This is better appreciated in the following figure.

What this tells us is that the mean of the variable X, denoted by $\bar{x}$, has a very interesting geometric interpretation. As you can tell, $\bar{x}$ is the length of the

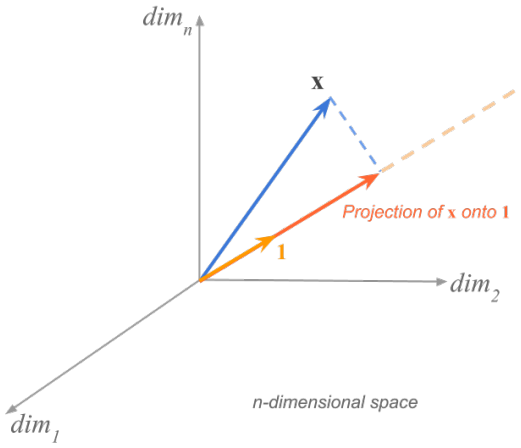

Figure 3.15: Orthogonal projection of vector x onto constant vector 1

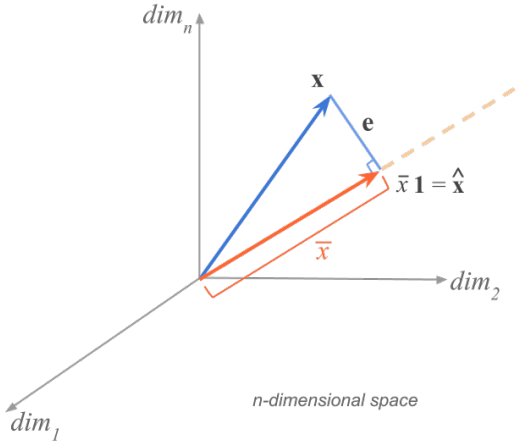

Figure 3.16: Mean of x as length of its projection onto constant vector 1

projected vector $\hat{\mathbf{x}}$. Or in more formal terms, $\bar{x}$ is the scalar projection of $\mathbf{x}$ onto $\mathbf{1}$.

Part III

Unsupervised I: PCA

Chapter 4

Principal Components Analysis

Our first unsupervised method of the book is Principal Components Analysis, commonly referred to as PCA.

Principal Components Analysis (PCA) is the workhorse method of multivariate data analysis. Simply put, PCA helps us study and explore a data set of quantitative variables measured on a set of objects. One way to look at the purpose of principal components analysis is to get the *best* low-dimensional representation of the variation in data. Among the various appealing features of PCA is that it allows us to obtain a visualization of the objects in order to see their proximities. Likewise, it also provides us results to get a graphic representation of the variables in terms of their correlations. Overall, PCA is a multivariate technique that allows us to summarize the systematic patterns of variations in a data set.

The classic reference for PCA is the work by the eminent British biostatistician Karl Pearson "On Lines and Planes of Closest Fit to Systems of Points in Space," from 1901. This publication presents the PCA problem under a purely geometric standpoint, describing how to find low-dimensional subspaces that best fit—in the least squares sense—a cloud of points. The other seminal work of PCA is the one by the American mathematician and economic theorist Harold Hotelling with "Analysis of a Complex of Statistical Variables into Principal Components," from 1933. Unlike Pearson, Hotelling finds the principal components as orthogonal linear combinations of the variables of maximum variance.

PCA is one of those methods that can be approached from multiple, seemingly unrelated, perspectives. The way we are going to introduce PCA is not the typical way in which PCA is discussed in most books published in English.

Figure 4.1: Cloud of points in the form of a mug

However, our introduction is actually based on the ideas and concepts originally published in Karl Pearson's 1901 paper *On lines and planes of closest fit to systems of points in space*. This is what can be considered to be the first paper on PCA, although keep in mind that Karl Pearson never used the term *principal components analysis*. That term was coined by Harold Hotelling, who formalized the method by giving it a more mature statistical perspective.

4.1 Low-dimensional Representations

Let's play the following game. Imagine for a minute that you have the super-power to see any type of multidimensional space (not just three-dimensions). As we mentioned before, we think of the individuals as forming a cloud of points in a p-dim space, and the variables forming a cloud of arrows in an n-dim space.

Pretend that you have some data in which its cloud of points has the shape of a mug, like in the following diagram:

This mug is supposed be high-dimensional, and something that you are not supposed to ever see in real life. So the question is: Is there a way in which we can get a low-dimensional representation of this data?

Luckily, the answer is: YES, we can!

How? Well, the name of the game is **projections**: we can look for projections of the data into sub-spaces of lower dimension, like in the diagram below.

Think of *projections* as taking photographs or x-rays of the mug. You can take a photo of the mug from different angles. For instance, a picture in which the lens of the camera lies on the top of the mug, or another picture in which the lens is below the mug (from the bottom), and so on.

As you can tell from the above figure, we have three candidate subspaces: $\mathbb{H}_A$, $\mathbb{H}_B$, and $\mathbb{H}_C$. Among the three possible projections, subspace $\mathbb{H}_C$ is the one

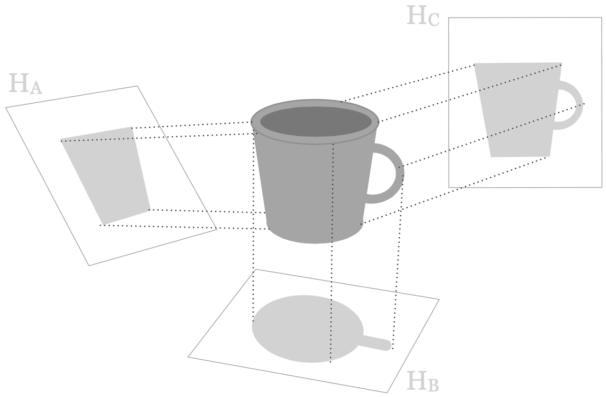

Figure 4.2: Various projections onto subspaces

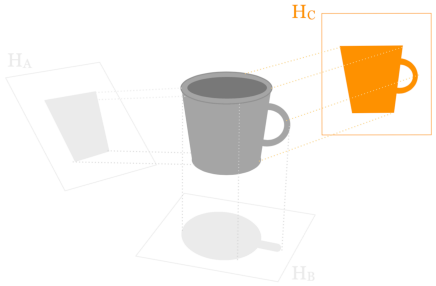

Figure 4.3: The shape of the projection is similar to the original mug shape.

that provides the best low dimensional representation, in the sense that the projected silhouette is the most similar to the original mug shape. We can say that this "photo" is the one that most resembles the original object. Now, keep in mind that the resulting image in the low-dimensional space is not capturing the whole pattern. In other words, there is some loss of information. However, by chosing the right project, we hope to minimize such loss.

4.2 PCA Idea

The overall idea behind PCA is the following. Given a set of p variables $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}$, we want to obtain new k variables $\mathbf{z_1}, \mathbf{z_2}, \ldots, \mathbf{z_k}$, called the **Principal Components** (PCs).

A principal component is a linear combination: $\mathbf{z} = \mathbf{Xv}$. The first PC is a linear mix:

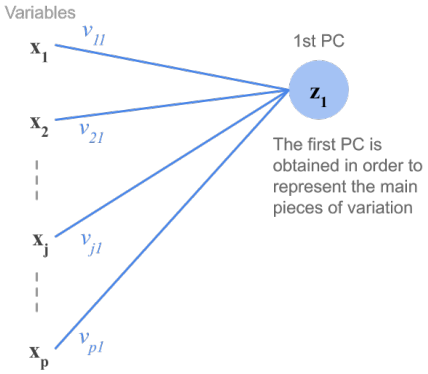

Figure 4.4: PCs as linear combinations of X-variables

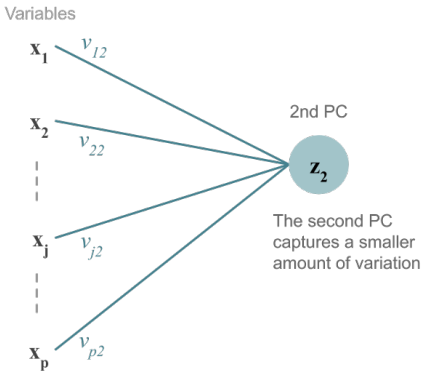

Figure 4.5: PCs as linear combinations of X-variables

The second PC is another linear mix:

We want to compute the **PCs as linear combinations** of the original variables.

$$\begin{aligned} \mathbf{z_1} &= v_{11}\mathbf{x_1} + \cdots + v_{1p}\mathbf{x_p} \\ \mathbf{z_2} &= v_{21}\mathbf{x_1} + \cdots + v_{2p}\mathbf{x_p} \\ &\vdots \\ \mathbf{z_k} &= v_{k1}\mathbf{x_1} + \cdots + v_{kp}\mathbf{x_p} \end{aligned}$$

Or in matrix notation:

$$\mathbf{Z} = \mathbf{XV}$$

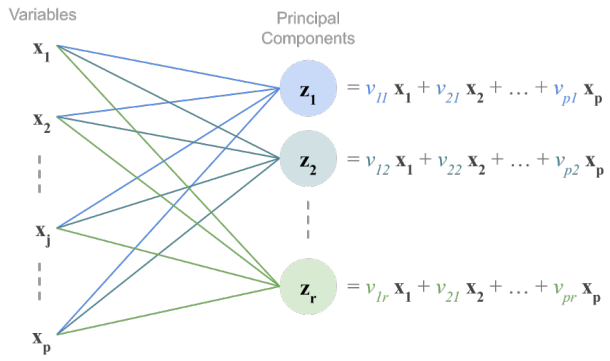

Figure 4.6: PCs as linear combinations of X-variables

where $\mathbf{Z}$ is an $n \times k$ matrix of principal components, and $\mathbf{V}$ is a $p \times k$ matrix of weights, also known as directional vectors of the principal axes. The following figure shows a graphical representation of a PCA problem in diagram notation:

We look to transform the original variables into a smaller set of new variables, the Principal Components (PCs), that summarize the variation in data. The PCs are obtained as linear combinations (i.e. weighted sums) of the original variables. We look for PCs is in such a way that they have maximum variance, and being mutually uncorrelated.

4.2.1 Finding Principal Components

The way to find principal components is to construct them as weighted sums of the original variables, looking to optimize some criterion and following some constraints. One way in which we can express the criterion is to require components $\mathbf{z_1}, \mathbf{z_2}, \ldots, \mathbf{z_k}$ that capture most of the variation in the data $\mathbf{X}$. "Capturing most of the variation," implies looking for a vector $\mathbf{v_j}$ such that a component $\mathbf{z_h} = \mathbf{X}\mathbf{v_h}$ has maximum variance:

$$\max_{\mathbf{v_h}} \; var(\mathbf{z_h}) \quad \Rightarrow \quad \max_{\mathbf{v_h}} \; var(\mathbf{X}\mathbf{v_h})$$

that is

$$\max_{\mathbf{v_h}} \; \frac{1}{n}\mathbf{v_h}^{\top}\mathbf{X}^{\top}\mathbf{X}\mathbf{v_h}$$

As you can tell, this is a maximization problem. Without any constraints, this problem is unbounded, not to mention useless. We could take $\mathbf{v_h}$ as bigger as

we want without being able to reach any maximum. To get a feasible solution we need to impose some kind of restriction. The standard adopted constraint is to require $\mathbf{v_h}$ to be of unit norm:

$$\|\mathbf{v_h}\| = 1 \;\; \Rightarrow \;\; \mathbf{v_h^T v_h} = 1$$

Note that $(1/n)\mathbf{X^T X}$ is the variance-covariance matrix. If we denote $\mathbf{S} = (1/n)\mathbf{X^T X}$ then the criterion to be maximized is:

$$\max_{\mathbf{v_h}} \; \mathbf{v_h^T S v_h}$$

subject to $\mathbf{v_h^T v_h} = 1$

To avoid a PC $\mathbf{z_h}$ from capturing the same variation as other PCs $\mathbf{z_l}$ (i.e. avoiding redundant information), we also require them to be **mutually orthogonal** so they are uncorrelated with each other. Formally, we impose the restriction $\mathbf{z_h}$ to be perpendicular to other components: $\mathbf{z_h^T z_l} = 0; (h \neq l)$.

4.2.2 Finding the first PC

In order to get the first principal component $\mathbf{z_1} = \mathbf{X v_1}$, we need to find $\mathbf{v_1}$ such that:

$$\max_{\mathbf{v_1}} \; \mathbf{v_1^T S v_1}$$

subject to $\mathbf{v_1^T v_1} = 1$

Being a maximization problem, the typical procedure to find the solution is by using the **Lagrangian multiplier** method. Using Lagrange multipliers we get:

$$\mathbf{v_1^T S v_1} - \lambda(\mathbf{v_1^T v_1} - 1) = 0$$

Differentiation with respect to $\mathbf{v_1}$, and equating to zero gives:

$$\mathbf{S v_1} - \lambda_1 \mathbf{v_1} = \mathbf{0}$$

Rearranging some terms we get:

$$\mathbf{S v_1} = \lambda_1 \mathbf{v_1}$$

What does this mean? It means that λ_1 is an eigenvalue of $\mathbf{S}$, and $\mathbf{v_1}$ is the corresponding eigenvector.

4.2.3 Finding the second PC

In order to find the second principal component $\mathbf{z_2} = \mathbf{X}\mathbf{v_2}$, we need to find $\mathbf{v_2}$ such that

$$\max_{\mathbf{v_2}} \mathbf{v_2}^\mathsf{T}\mathbf{S}\mathbf{v_2}$$

subject to $\|\mathbf{v_2}\| = 1$ and $\mathbf{z_1}^\mathsf{T}\mathbf{z_2} = 0$. Remember that $\mathbf{z_2}$ must be uncorrelated to $\mathbf{z_1}$. Applying the Lagrange multipliers, it can be shown that the desired $\mathbf{v_2}$ is such that

$$\mathbf{S}\mathbf{v_2} = \lambda_2\mathbf{v_2}$$

In other words. λ_2 is an eigenvalue of $\mathbf{S}$ and $\mathbf{v_2}$ is the corresponding eigenvector.

4.2.4 Finding all PCs

All PCs can be found simultaneously by **diagonalizing S**. Diagonalizing $\mathbf{S}$ involves expressing it as the product:

$$\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\mathsf{T}$$

where:

- $\mathbf{D}$ is a diagonal matrix
- the elements in the diagonal of $\mathbf{D}$ are the eigenvalues of $\mathbf{S}$
- the columns of $\mathbf{V}$ are orthonormal: $\mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}$
- the columns of $\mathbf{V}$ are the eigenvectors of $\mathbf{S}$
- $\mathbf{V}^\mathsf{T} = \mathbf{V}^{-\mathbf{1}}$

Diagonalizing a symmetric matrix is nothing more than obtaining its **eigenvalue decomposition** (a.k.a. spectral decomposition). A $p \times p$ symmetric matrix $\mathbf{S}$ has the following properties:

- $\mathbf{S}$ has p real eigenvalues (counting multiplicites)
- the eigenvectors corresponding to different eigenvalues are orthogonal
- $\mathbf{S}$ is orthogonally diagonalizable ($\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\mathsf{T}$)
- the set of eigenvalues of $\mathbf{S}$ is called the **spectrum** of $\mathbf{S}$

In summary: The PCA solution can be obtained with an Eigenvalue Decomposition of the matrix $\mathbf{S} = (1/n)\mathbf{X}^\mathsf{T}\mathbf{X}$

4.3 PCA Model

Formally, PCA involves finding scores and loadings such that the data can be expressed as a product of two matrices:

$$\underset{n\times p}{\mathbf{X}} = \underset{n\times k}{\mathbf{Z}}\,\underset{k\times p}{\mathbf{V}^{\mathsf{T}}}$$

where $\mathbf{Z}$ is the matrix of PCs or *scores*, and $\mathbf{V}$ is the matrix of *loadings*. We can obtain as many different eigenvalues as the rank of $\mathbf{S}$ denoted by k. Ideally, we expect k to be smaller than p so we get a convenient data reduction. But usually we will only retain just a few PCs (i.e. $k \ll p$) expecting not to lose too much information:

$$\underset{n\times p}{\mathbf{X}} \approx \underset{n\times k}{\mathbf{Z}}\,\underset{k\times p}{\mathbf{V}^{\mathsf{T}}} + \text{Residual}$$

The previous expression means that just a few PCs will *optimally* summarize the main structure of the data

4.4 Another Perspective

Finding $\mathbf{z_h} = \mathbf{X}\mathbf{v_h}$ with maximum variance has another important property that it is not always mentioned in multivariate textbooks but that we find worth mentioning. $\mathbf{z_h}$ is such that

$$\max \sum_{j=1}^{p} cor^2(\mathbf{z_h}, \mathbf{x_j})$$

What this expression implies is that principal components $\mathbf{z_h}$ are computed to be the *best* representants in terms of maximizing the sum of squared correlations with the variables $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_j}$. Interestingly, you can think of PCs as predictors of the variables in $\mathbf{X}$. Under this perspective, we can reverse the relations and see PCA from a regression-like model perspective:

$$\mathbf{x_j} = v_{jh}\mathbf{z_h} + \mathbf{e_h}$$

Notice that the regression coefficient is the j-th element of the h-th eigenvector.