Multivariate Analysis

STATS 6M03

Fall 2022

Cameron Roopnarine[*] Paul McNicholas[†]

8th December 2022

Contents

LECTURE 1
7th September

1 Introduction to Multivariate Distributions

Notation & Terminology

- We observe n realizations $\underline{x}_1, \dots, \underline{x}_n$ of p-dimensional random variables $\underline{\mathcal{X}}_1, \dots, \underline{\mathcal{X}}_n$, where $\underline{\mathcal{X}}_i = (\mathcal{X}_{i1}, \dots, \mathcal{X}_{ip})'$ for $i = 1, \dots, n$.
- In matrix form,
$$\boldsymbol{\mathcal{X}} = (\underline{\mathcal{X}}_1, \dots, \underline{\mathcal{X}}_n)' = \begin{pmatrix} \underline{\mathcal{X}}_1' \\ \vdots \\ \underline{\mathcal{X}}_n' \end{pmatrix} = \begin{pmatrix} \mathcal{X}_{11} & \cdots & \mathcal{X}_{1p} \\ \vdots & \ddots & \vdots \\ \mathcal{X}_{n1} & \cdots & \mathcal{X}_{np} \end{pmatrix}.$$
- $\underline{\mathcal{X}}_i$ is called a **random vector**.
- $\boldsymbol{\mathcal{X}}$ is called an $n \times p$ **random matrix**.
- A matrix $\boldsymbol{A}$ with all entries constant is called a **constant matrix**.

Basic Definitions & Results

- If $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$ be $n \times p$ random matrices, then
$$\mathbb{E}[\boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{Y}}] = \mathbb{E}[\boldsymbol{\mathcal{X}}] + \mathbb{E}[\boldsymbol{\mathcal{Y}}].$$
Furthermore, if $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ are $m \times n$, $p \times q$, and $m \times q$ matrices of constants, respectively, then
$$\mathbb{E}[\boldsymbol{A}\boldsymbol{\mathcal{X}}\boldsymbol{B} + \boldsymbol{C}] = \boldsymbol{A}\,\mathbb{E}[\boldsymbol{\mathcal{X}}]\boldsymbol{B} + \boldsymbol{C}.$$

[*] LaTeXer

[†] Instructor

- If $\underline{\mathcal{X}}_i$ has mean $\underline{\mu}$, then the **covariance matrix** of $\underline{\mathcal{X}}_i$ is

$$\mathbf{\Sigma} = \mathrm{Var}(\underline{\mathcal{X}}_i) = \mathbb{E}\big[(\underline{\mathcal{X}}_i - \underline{\mu})(\underline{\mathcal{X}}_i - \underline{\mu})'\big].$$

- If p-dimensional $\underline{\mathcal{X}}_i$ has mean $\underline{\mu}$ and q-dimensional $\underline{\mathcal{Y}}_i$ has mean $\underline{\theta}$, then

$$\mathrm{Cov}(\underline{\mathcal{X}}_i, \underline{\mathcal{Y}}_i) = \mathbb{E}\big[(\underline{\mathcal{X}}_i - \underline{\mu})(\underline{\mathcal{Y}}_i - \underline{\theta})'\big].$$

Covariance Matrix I

- The covariance matrix $\mathbf{\Sigma}$ is

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{pmatrix}.$$

- The elements σ_{ii} are variances.
- The elements σ_{ij}, where $i \neq j$ are covariances.
- $\mathbf{\Sigma}$ is symmetric; that is, $\sigma_{ij} = \sigma_{ji}$.

Covariance Matrix II

- Let $\mathbf{\Sigma}$ be the covariance matrix of a $p \times 1$ random vector.

  (i) $\mathbf{\Sigma}$ is positive semi-definite; that is, for any $p \times 1$ constant vector $\underline{a} = (a_1, \ldots, a_p)'$,

$$\underline{a}'\mathbf{\Sigma}\underline{a} \geq \underline{0}.$$

  (ii) If $\boldsymbol{B}$ is a $q \times p$ constant matrix and $\underline{b}$ is a $q \times 1$ constant vector, then the covariance matrix of $\boldsymbol{\mathcal{Y}} = \boldsymbol{B}\boldsymbol{\mathcal{X}} + \underline{b}$ is

$$\mathrm{Var}(\boldsymbol{\mathcal{Y}}) = \boldsymbol{B}\mathbf{\Sigma}\boldsymbol{B}'.$$

- Note that the covariance matrix of $\mathbf{\Sigma}$ is **positive definite** if

$$\underline{a}'\mathbf{\Sigma}\underline{a} > \underline{0}$$

  for any constant vector $\underline{a} \neq \underline{0}$.

Covariance Matrix III

- In general, the covariance matrix $\mathbf{\Sigma}$ is positive semi-definite.
- Therefore, the eigenvalues of $\mathbf{\Sigma}$ are non-negative, and are denoted

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0.$$

- Eigenvalues are also known as characteristic roots.
- Write $\underline{v}_i \neq \underline{0}$ to denote an eigenvector of $\mathbf{\Sigma}$ corresponding to the eigenvalue λ_i for $i = 1, \ldots, p$.
- Recall that eigenvalues and eigenvectors of $\mathbf{\Sigma}$ satisfy

$$\mathbf{\Sigma}\underline{v}_i = \lambda_i \underline{v}_i \tag{$\star$}$$

  for $i = 1, \ldots, p$.

Aside: Interesting Result

Let $\boldsymbol{A}$ be a symmetric matrix. Suppose $\underline{e}_1$ and $\underline{e}_2$ are eigenvectors of $\boldsymbol{A}$ with corresponding eigenvalues λ_1 and λ_2, respectively, such that $\lambda_1 \neq \lambda_2$. Then, $\underline{e}_1$ and $\underline{e}_2$ are orthogonal; that is,

$$\underline{e}_2'\underline{e}_1 = 0.$$

Since $\boldsymbol{A}\underline{e}_2 = \lambda_2\underline{e}_2$, we have

$$\underline{e}_1'\boldsymbol{A}\underline{e}_2 = \lambda_2\underline{e}_1'\underline{e}_2 \tag{1}$$

Now,

$$\begin{aligned}(\underline{e}_1'\boldsymbol{A}\underline{e}_2)' &= (\boldsymbol{A}\underline{e}_2)'(\underline{e}_1')' \\ &= \underline{e}_2'\boldsymbol{A}\underline{e}_1 \\ &= \underline{e}_2'(\lambda_1\underline{e}_1) \\ &= \lambda_1\underline{e}_2'\underline{e}_1.\end{aligned} \tag{2}$$

Using (2), we also have

$$\begin{aligned}(\underline{e}_1'\boldsymbol{A}\underline{e}_2)' &= \lambda_2(\underline{e}_1'\underline{e}_2)' \\ &= \lambda_2\underline{e}_2'\underline{e}_1.\end{aligned} \tag{3}$$

Equating (2) and (3) yields

$$\begin{aligned}\lambda_1\underline{e}_2'\underline{e}_1 &= \lambda_2\underline{e}_2'\underline{e}_1 \\ \Longrightarrow\ (\lambda_2 - \lambda_1)\underline{e}_2'\underline{e}_1 &= 0 \\ \Longrightarrow\ \underline{e}_2'\underline{e}_1 &= 0 \text{ since } \lambda_1 \neq \lambda_2.\end{aligned}$$

Covariance Matrix IV

Without loss of generality, it can be assumed that the eigenvectors are orthonormal; that is,

$$\underline{v}_i'\underline{v}_i = 1, \text{ and } \underline{v}_i'\underline{v}_j = 0,\ i \neq j.$$

Covariance Matrix V

- Write $\boldsymbol{V} = (\underline{v}_1, \ldots, \underline{v}_p)$ and

$$\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \ldots, \lambda_p\} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}.$$

- We can show that $(\star)$ can be written $\boldsymbol{\Sigma V} = \boldsymbol{V\Lambda}$.

First,

$$\boldsymbol{\Sigma V} = \boldsymbol{\Sigma}(\underline{v}_1, \ldots, \underline{v}_p) = (\boldsymbol{\Sigma}\underline{v}_1, \ldots, \boldsymbol{\Sigma}\underline{v}_p). \tag{1}$$

Second,

$$\boldsymbol{V\Lambda} = (\lambda_1\underline{v}_1, \ldots, \lambda_p\underline{v}_p). \tag{2}$$

Now, (1) = (2) if and only if

$$\begin{aligned}\boldsymbol{\Sigma}\underline{v}_1 &= \lambda_1\underline{v}_1\\ &\vdots\\ \boldsymbol{\Sigma}\underline{v}_p &= \lambda_p\underline{v}_p.\end{aligned}$$

That is, $\boldsymbol{\Sigma}\underline{v}_i = \lambda_i\underline{v}_i$ for $i = 1, \ldots, p$.

- Note that $\boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}_p$ and
$$\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}' = \boldsymbol{V}\boldsymbol{\Lambda}^{1/2}\boldsymbol{V}'\boldsymbol{V}\boldsymbol{\Lambda}^{1/2}\boldsymbol{V}',$$
where $\boldsymbol{\Lambda}^{1/2} = \mathrm{diag}\{\sqrt{\lambda}_1, \ldots, \sqrt{\lambda}_p\}$.

Covariance Matrix VI

- We can also show that the determinant of the covariance matrix $\boldsymbol{\Sigma}$ can be written
$$|\boldsymbol{\Sigma}| = |\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}'| = |\boldsymbol{\Lambda}| = \prod_{i=1}^{p}\lambda_i.$$

$$|\boldsymbol{\Sigma}| = |\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}'| = |\boldsymbol{V}||\boldsymbol{\Lambda}||\boldsymbol{V}^{-1}| = \frac{|\boldsymbol{V}||\boldsymbol{\Lambda}|}{|\boldsymbol{V}|} = |\boldsymbol{\Lambda}| = \prod_{i=1}^{p}\lambda_i.$$

- The total variance of the covariance $\boldsymbol{\Sigma}$ can be written as
$$\mathrm{tr}\{\boldsymbol{\Sigma}\} = \mathrm{tr}\{\boldsymbol{\Lambda}\} = \sum_{i=1}^{p}\lambda_i.$$

$$\begin{aligned}\mathrm{tr}\{\boldsymbol{\Sigma}\} &= \mathrm{tr}\{\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}'\}\\ &= \mathrm{tr}\{\boldsymbol{\Lambda}\boldsymbol{V}\boldsymbol{V}'\}\\ &= \mathrm{tr}\{\boldsymbol{\Lambda}\}\\ &= \sum_{i=1}^{p}\lambda_i.\end{aligned}$$

- The determinant and total variance are sometimes used as summaries of the total scatter amongst the p variables.

Comments

- This lecture has presented some basics, much of which is a revision of ideas encountered in linear algebra.
- Please take some time to digest this material before the next class.
- In the next class, we will start to look at principal components analysis.

Lecture 2
8th September

2 Principal Component Analysis

What is a Principal Component?

- The first principal component is the direction of most variation (in the data).
- The second principal component is the direction of most variation (in the data) conditional on it being orthogonal to the first principal component.
- The third principal component is the direction of most variation (in the data) conditional on it being orthogonal to the first two principal components.
- For $r > 1$: the r^{th} principal component is the direction of most variation (in the data) conditional on it being orthogonal to the first $r-1$ principal components.

Definition

- Let $\underline{\mathcal{X}}$ be a p-dimensional random vector with mean $\underline{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
- Let $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ be the (ordered) eigenvalues of $\boldsymbol{\Sigma}$ and let $\underline{v}_1, \ldots, \underline{v}_p$ be the corresponding eigenvectors.
- The i^{th} principal component of $\underline{\mathcal{X}}$ is
$$\mathcal{W}_i = \underline{v}_i'(\underline{\mathcal{X}} - \underline{\mu}), \; i = 1, \ldots, p.$$
- That is,
$$\underline{\mathcal{W}} = \boldsymbol{V}'(\underline{\mathcal{X}} - \underline{\mu}),$$
where $\underline{\mathcal{W}} = (\mathcal{W}_1, \ldots, \mathcal{W}_p)'$ and $\boldsymbol{V} = (\underline{v}_1, \ldots, \underline{v}_p)$.

Some Results

- $\mathbb{E}[\underline{\mathcal{W}}] = \underline{0}$.

$$\begin{aligned}
\mathbb{E}[\underline{\mathcal{W}}] &= \mathbb{E}[\boldsymbol{V}'(\underline{\mathcal{X}} - \underline{\mu})] \\
&= \boldsymbol{V}'\,\mathbb{E}[\underline{\mathcal{X}} - \underline{\mu}] \\
&= \boldsymbol{V}'\underline{0} \\
&= \underline{0}.
\end{aligned}$$

- $\operatorname{Var}(\underline{\mathcal{W}}) = \boldsymbol{\Lambda}$.

$$\begin{aligned}
\operatorname{Var}(\underline{\mathcal{W}}) &= \mathbb{E}[\underline{\mathcal{W}}\underline{\mathcal{W}}'] - \mathbb{E}[\underline{\mathcal{W}}]\,\mathbb{E}[\underline{\mathcal{W}}]' \\
&= \mathbb{E}[\underline{\mathcal{W}}\underline{\mathcal{W}}'] \\
&= \mathbb{E}\big[\boldsymbol{V}'(\underline{\mathcal{X}} - \underline{\mu})(\underline{\mathcal{X}} - \underline{\mu})'\boldsymbol{V}\big] \\
&= \boldsymbol{V}'\,\mathbb{E}\big[(\underline{\mathcal{X}} - \underline{\mu})(\underline{\mathcal{X}} - \underline{\mu})'\big]\boldsymbol{V} \\
&= \boldsymbol{V}'\operatorname{Var}(\underline{\mathcal{X}})\boldsymbol{V} \\
&= \boldsymbol{V}'\boldsymbol{\Sigma}\boldsymbol{V} \\
&= \boldsymbol{\Lambda}.
\end{aligned}$$

- $\underline{\mathcal{X}} = \underline{\mu} + \boldsymbol{V}\underline{\mathcal{W}} = \underline{\mu} + \sum_{i=1}^{p} \underline{v}_i \mathcal{W}_i$.

Using $\underline{\mathcal{W}} = \boldsymbol{V}'(\underline{\mathcal{X}} - \underline{\mu})$, we multiply $\boldsymbol{V}$ from the LHS to get

$$\begin{aligned} \boldsymbol{V}\underline{\mathcal{W}} &= \boldsymbol{V}\boldsymbol{V}'(\underline{\mathcal{X}} - \underline{\mu}) \\ &= \boldsymbol{I}_p(\underline{\mathcal{X}} - \underline{\mu}) \\ &= \underline{\mathcal{X}} - \underline{\mu}. \end{aligned}$$

Rearranging,

$$\underline{\mathcal{X}} = \underline{\mu} + \boldsymbol{V}\underline{\mathcal{W}} = \underline{\mu} + \sum_{i=1}^{p} \underline{v}_i \mathcal{W}_i.$$

- $\sum_{i=1}^{p} \mathrm{Var}(\mathcal{W}_i) = \sigma_{11} + \cdots + \sigma_{pp}$.

From the second result,

$$\begin{aligned} \mathrm{Var}(\mathcal{W}_i) &= \lambda_i \\ \implies \sum_{i=1}^{p} \mathrm{Var}(\mathcal{W}_i) &= \sum_{i=1}^{p} \lambda_i \\ &= \mathrm{tr}\{\boldsymbol{\Lambda}\} \\ &= \mathrm{tr}\{\boldsymbol{\Sigma}\} \\ &= \sigma_{11} + \cdots + \sigma_{pp}. \end{aligned}$$

Key Results

- Consider $\mathcal{W} = \underline{v}'(\underline{\mathcal{X}} - \underline{\mu})$ with $\underline{v}'\underline{v} = 1$.

  i. $\mathrm{Var}(\mathcal{W})$ is maximized when $\mathcal{W} = \mathcal{W}_1$, the first principal component.

Note that $\mathrm{Var}(\mathcal{W}) = \underline{v}'\boldsymbol{\Sigma}\underline{v}$. We can write

$$\underline{v} = c_1\underline{v}_1 + \cdots + c_p\underline{v}_p = \boldsymbol{V}\underline{c},$$

where $\underline{c}$ satisfies $\underline{c}'\underline{c} = 1$.
Note: $\underline{v}'\underline{v} = \underline{c}'\boldsymbol{V}'\boldsymbol{V}\underline{c} = \underline{c}'\underline{c} = 1$.
Therefore,

$$\mathrm{Var}(\mathcal{W}) = \underline{c}'\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}'\underline{c} = c_1^2\lambda_1 + \cdots + c_p^2\lambda_p,$$

which is maximized when $c_1 = 1$ and $c_2 = \cdots = c_p = 0$; that is, when $\underline{v} = \underline{v}_1$.

  ii. If $\mathcal{W}$ is uncorrelated with the first $k < p$ principal components, $\mathcal{W}_1, \ldots, \mathcal{W}_k$, the $\mathrm{Var}(\mathcal{W})$ is maximized when $\mathcal{W} = \mathcal{W}_{k+1}$, the $(k+1)^{\text{th}}$ principal component.

$$\begin{aligned}
\operatorname{Cov}(\mathcal{W}, \mathcal{W}_i) &= \mathbb{E}\big[\underline{v}'(\underline{\mathcal{X}} - \underline{\mu})(\underline{\mathcal{X}} - \underline{\mu})'\underline{v}_i\big] \\
&= \underline{v}'\, \mathbb{E}\big[(\underline{\mathcal{X}} - \underline{\mu})(\underline{\mathcal{X}} - \underline{\mu})'\big]\underline{v}_i \\
&= \underline{v}' \operatorname{Var}(\underline{\mathcal{X}})\underline{v}_i \\
&= \underline{v}'\boldsymbol{\Sigma}\underline{v}_i \\
&= \underline{c}'\boldsymbol{V}\boldsymbol{\Sigma}\underline{v}_i \\
&= \lambda_i\underline{c}'\boldsymbol{V}\underline{v}_i && \text{since } \boldsymbol{\Sigma}\underline{v}_i = \lambda_i\underline{v}_i \\
&= \lambda_i(\boldsymbol{V}\underline{c})'\underline{v}_i \\
&= \lambda_i c_i.
\end{aligned}$$

The maximum of the previous result under $\lambda_i c_i = 0$ for $i = 1, \ldots, p$ is when $\underline{v} = \underline{v}_{k+1}$.

- The proportion of the total variation explained by the i^{th} principal component is $\lambda_i / \operatorname{tr}\{\boldsymbol{\Sigma}\}$.
- The proportion of the total variation explained by the first k principal components is

$$\frac{\sum_{i=1}^{k}\lambda_i}{\operatorname{tr}\{\boldsymbol{\Sigma}\}}.$$

The last two key results follow directly from the fourth result in the “some results” subsection.

Next Step

- The next step is to do some principal components analyses.
- We will do this in R.
- R code from class is posted on the course materials page.
- So don’t waste your time writing down the code; rather, try to focus on understanding the analyses (while taking any necessary notes).

Lecture 22
14th October

3 Introduction to Bayesian Inference

Introduction

- I would like to discuss some Bayesian topics in multivariate statistics.
- But I am aware that many of you will not yet have encountered the Bayesian paradigm.
- Therefore, I am going to devote a lecture or two to introducing the Bayesian paradigm.
- This introduction will assume n realizations of a univariate random variable X, which we denote $x_1, \ldots, x_n$ or $\underline{x}$.

Probability & the Bayesian Approach

- First, we are going to look a little more closely at what a probability means.
- We will also look at some aspects of the Bayesian approach.
- I am going to present this material from what may be considered a reasonably pro-Bayesian viewpoint.
- This is only fair really…

What is Probability?

- You will have seen this standard definition before.
- We talk about the probability of an event. If E is an event, then we denote the probability of E occurring by $\mathbb{P}(E)$.
- For an experiment with possible outcomes $E_1, \dots, E_n$, the probability $\mathbb{P}(E_i)$ must obey the following rules:
  - $0 \leq \mathbb{P}(E_i) \leq 1$ for all E_i.
  - $\mathbb{P}(E_1) + \cdots + \mathbb{P}(E_n) = 1$.
  - $\mathbb{P}(\emptyset) = 0$.
  - The OR law for mutually exclusive events.

The Rules of Probability

- A probability is a number that measures our uncertainty in a random variable X.
- A popular way to view a probability is to consider that it must obey the following three laws, known as the **calculus of probability**:
  - Convexity: $0 \leq \mathbb{P}(X) \leq 1$.
  - Additivity: If both X_1 and X_2 are mutually exclusive then
    $$\mathbb{P}(X_1 \cup X_2) = \mathbb{P}(X_1) + \mathbb{P}(X_2).$$
  - Multiplicativity:
    $$\mathbb{P}(X_1 \cap X_2) = \mathbb{P}(X_1)\, \mathbb{P}(X_2 \mid X_1).$$
- We must also consider that a certain event that is assigned a probability of 1 and an impossible event is assigned a probability of 0.

Alternatives to Probability

- There are several competing theories to probability for quantifying uncertainty.
- The most well known is fuzzy logic and its counterpart fuzzy set theory.
- Another is possibility theory.
- These competitors have been used successfully in some applications.
- However, they all lack a justification from more fundamental ideas about uncertainty that probability possesses.

The Meaning of a Probability

- What precisely is the meaning of a probability?
- We can easily define a probability very precisely, but that will not tell us it means.
- The actually meaning of a probability is a topic that continues to be the subject of some debate.
- We are going to look at two popular concepts.
- Namely, **physical probability** and **psychological probability**.

Physical Probability

- Also known as material probability, intrinsic probability, objective probability or propensity.
- Probability is viewed as a property of the material world, like mass or volume, which exists irrespective of minds and logic.
- A physical probability is typically defined in one of two ways: first principles and relative frequency.
- **First Principles**: There are n possible outcomes to some random quantity. For an event E to occur, one of r specific outcomes must occur. By making an assumption that each outcome is equally likely, $\mathbb{P}(E) = r/n$.
- **Relative Frequency**: The proportion of times that X has been observed to occur in a long sequence of essentially identical experiments.
- The relative frequency notion is usually adopted and people who adopt this interpretation of probability are called frequentists.

Psychological Probability

- A degree of belief, or intensity of conviction, used for betting or making decisions, not necessarily after mature consideration.
- It does not have to be consistent with one's other opinions.
- When a psychological probability is coherent (see De Finetti) and obeys the laws of probability then it is called **subjective probability**.
- The subjective probability notion is usually adopted and people who adopt this notion are called **subjectivists** or, some would say, **Bayesians**.

Frequentist vs Subjectivist

- Frequentist
  1. Probably still the most widely used.
  2. Consistent (everyone should get the same probability for a given event).
  3. Assumes that an experiment can be repeated indefinitely under almost identical conditions. This excludes one-off events.
  4. What are 'almost identical conditions'?
- Subjectivist
  1. Concerned with individual behaviour.
  2. Varies from individual to individual — no "correct" probability for a given event.
  3. Applies to a wider range of situations, including one-off situations.

The Bayesian Paradigm

- For our purposes today, we are going to assume that all Bayesians are subjectivists.
- Bayesians obey the laws of probability strictly and conduct statistical inference accordingly.
- One can think of Bayesians as using observed data to update their existing, or prior, knowledge.
- In a moment, an example, first some revision.

Conditional Probability

- Recall the **AND** Rule: if E and F are two events then

$$\mathbb{P}(E \cap F) = \mathbb{P}(E)\,\mathbb{P}(F \mid E),$$

where $F \mid E$ means the occurrence of an event F given that an event E has already occurred.

- Now, dividing both sides of the above equation by $\mathbb{P}(E)$ gives us the definition of a conditional probability.
- The probability that an event F occurs given that an event E has already occurred is given by

$$\mathbb{P}(F \mid E) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(E)}.$$

- Using the fact that $\mathbb{P}(E \cap F) = \mathbb{P}(F \cap E)$ and noting that $\mathbb{P}(F \cap E) = \mathbb{P}(E \mid F)\,\mathbb{P}(F)$, we can write the expression for $\mathbb{P}(F \mid E)$ as follows.

$$\mathbb{P}(F \mid E) = \frac{\mathbb{P}(E \mid F)\,\mathbb{P}(F)}{\mathbb{P}(E)}. \tag{1}$$

- Now, we can also rewrite the term $\mathbb{P}(E)$ in this equation.

The Partition Law

- This approach can be generalized to get a general formula for $\mathbb{P}(E)$ in terms of conditional probabilities.
- **The Partition Theorem**: Suppose the outcome of an event E depends on an event F which has possible outcomes $F_1, \ldots, F_n$, then

$$\mathbb{P}(E) = \sum_{i=1}^{n} \mathbb{P}(E \mid F_i)\,\mathbb{P}(F_i).$$

Bayes' Theorem

- The Partition Theorem replaces $\mathbb{P}(E)$ in the denominator of Equation (1) to give Bayes' Theorem.
- **Bayes' Theorem**: Suppose the outcome of an event E depends on an event F which has possible outcomes $F_1, \ldots, F_n$, then

$$\mathbb{P}(F_j \mid E) = \frac{\mathbb{P}(E \mid F_j)\,\mathbb{P}(F_j)}{\sum_{i=1}^{n} \mathbb{P}(E \mid F_i)\,\mathbb{P}(F_i)},$$

for $j = 1, 2, \ldots, n$.

- However, the result is not actually due to Bayes.

Thomas Bayes

- Fisher wrote that:

“For the first serious attempt known to us to give a rational account of the process of scientific inference as a means of understanding the real world, in the sense in which this term is understood by experimental investigators, we must look back over two hundred years to an English clergyman, the Reverend Thomas Bayes, whose life spanned the first half of the eighteenth century.”

- What Bayes actually showed was for one case of continuous X, and it is not clear that he would agree entirely with all aspects of what has become known as **Bayesian inference**.

A Bayesian Example

- Suppose I am Bayesian.
- I want to estimate the height of men in Ireland.
- A reasonable description of my (prior) belief is that the height of men in Ireland is $\mathcal{N}(1.70, 0.15^2)$.
- Data are then collected, and my views are 'updated'.
- But how?

The Bayesian Approach I

- Denote the probability density that describes my prior belief by $h(\theta)$.
- Then, $h(\theta)$ is called the **prior distribution**.
- Once we observe our data, we write down the likelihood $\mathcal{L}(\theta \mid \underline{x}) = p(\underline{x} \mid \theta)$.
- We then compute the **posterior distribution** $h(\theta \mid \underline{x})$.
- We can find $h(\theta \mid \underline{x})$ using Bayes theorem.

The Bayesian Approach II

- From Bayes theorem,
$$h(\theta \mid \underline{x}) = \frac{p(\underline{x} \mid \theta)h(\theta)}{\int_\Theta p(\underline{x} \mid \theta)h(\theta)\,\mathrm{d}\theta}.$$
- Noting that the denominator is a constant with respect to θ, we can write
$$h(\theta \mid \underline{x}) \propto p(\underline{x} \mid \theta)h(\theta).$$
- We also know that
$$\int_\Theta h(\theta \mid \underline{x})\,\mathrm{d}\theta = 1.$$

Example 1

- Suppose we have $x_1, x_2, \ldots, x_n$ each from a BERN(θ).
- Suppose that our prior is a Beta(α, β), so that
$$h(\theta) = \frac{\Gamma(\alpha, \beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},\ 0 < \theta < 1.$$
- What is the posterior distribution?

First,
$$p(\underline{x} \mid \theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}.$$

Therefore,

$$\begin{aligned} h(\theta \mid \underline{x}) &\propto p(\underline{x} \mid \theta) h(\theta) \\ &= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \frac{\Gamma(\alpha, \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{\sum_{i=1}^n x_i+\alpha-1}(1-\theta)^{n-\sum_{i=1}^n x_i+\beta-1}, \end{aligned}$$

which is the functional form of Beta$(\sum_{i=1}^n x_i + \alpha, n - \sum_{i=1}^n x_i + \beta)$.

- What is the posterior mean?

From our distribution above,

$$\begin{aligned} \mathbb{E}[\theta \mid \underline{x}] = \frac{\sum_{i=1}^n x_i + \alpha}{\alpha + n + \beta} &= \frac{\sum_{i=1}^n x_i + \alpha}{\alpha + n + \beta} \\ &= \left(\frac{n}{\alpha+\beta+n}\right)\frac{1}{n}\sum_{i=1}^n x_i + \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\frac{\alpha}{\alpha+\beta} \\ &= \left(1 - \frac{\alpha+\beta}{\alpha+\beta+n}\right)\frac{1}{n}\sum_{i=1}^n x_i + \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\frac{\alpha}{\alpha+\beta} \\ &= (1-\gamma_n)\hat{\theta} + \gamma_n \, \mathbb{E}[\theta], \end{aligned}$$

where $\mathbb{E}[\theta] = \frac{\alpha}{\alpha+\beta}$ is the prior mean, $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n x_i$ is the MLE of the data, and $\gamma_n = \frac{\alpha+\beta}{\alpha+\beta+n}$. We observe that as $n \to \infty$, $\gamma_n \to 0$, so that $\mathbb{E}[\theta \mid \underline{x}] \to \hat{\theta}$, which is obvious. If we collect an infinite amount of data, we do not care about our prior knowledge.

Notes

- Note that in using this beta prior, we have distilled our prior beliefs to two numbers: α and β.
- Note also that some people use a Beta$(1,1)$, which is a uniform, to indicate 'no' prior knowledge.
- There are problems using uniform priors in this way, which we shall see later on.
- When the prior and the posterior for θ are from the same distribution, this is called **conjugacy**.
- The definition of a conjugate prior is important, so let's be careful.
- A **conjugate prior** $h(\theta)$ is a distribution such that the posterior $h(\theta \mid \underline{x})$ is of the same distribution type.
- If $X \sim$ Beta(α, β), then

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}, \quad \mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Example 2

- Show that the conjugate prior distribution for the POI(λ) is the Gamma distribution.

First,

$$p(\underline{x} \mid \lambda) = \prod_{i=1}^{n} p(x_i \mid \lambda) = \lambda^{x_i} \frac{e^{-\lambda}}{x_i!} = \lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda} \prod_{i=1}^{n} \frac{1}{x_i} \propto \lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}.$$

Assuming that $\lambda \sim \text{GAM}(\alpha, \beta)$, we have

$$h(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \propto \lambda^{\alpha-1} e^{-\beta\lambda}.$$

Therefore,

$$\begin{aligned} h(\lambda \mid \underline{x}) &\propto p(\underline{x} \mid \lambda) h(\lambda) \\ &\propto \lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda} \lambda^{\alpha-1} e^{-\beta\lambda} \\ &= \lambda^{\alpha + \sum_{i=1}^{n} x_i - 1} e^{-(\beta+n)\lambda}, \end{aligned}$$

which is the functional form of $\text{GAM}(\alpha + \sum_{i=1}^{n} x_i, \beta + n)$.

Example 3

- We want to assess the probability θ that a quarter, when tossed, will land heads-up.
- Your prior belief is that θ is probably 0.5 but may be anywhere from 0.3 to 0.7.
- Construct a beta distribution for θ that reflects your prior beliefs.
- The coin is tossed 10 times and lands heads-up 7 times.
- What is the posterior mean and variance?

Since prior belief is that θ is probably 0.5, but may be anywhere from 0.3 to 0.7, we have

$$\mathbb{E}[\theta] \pm 2\sqrt{\text{Var}(\theta)} = (0.3, 0.7),$$

where $\mathbb{E}[\theta] = 0.5$. Hence,

$$0.5 \pm 2\sqrt{\text{Var}(\theta)} = (0.3, 0.7) \implies \text{Var}(\theta) = 0.01.$$

Hence,

$$\frac{\alpha}{\alpha + \beta} = 0.5 \implies \alpha = \beta.$$

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = 0.01 \implies \alpha = 12.$$

Therefore, $h(\theta)$, our prior distribution, is

$$h(\theta) \propto \theta^{11}(1-\theta)^{11},$$

which is $\text{Beta}(12, 12)$.

From earlier, we see that $h(\underline{x} \mid \theta) \propto \text{Beta}(7 + 12, 10 - 7 + 12) = \text{Beta}(19, 15)$.