Course Notes for
Statistics and Probability

Fei Li*

*Email: fei.li.best@gmail.com.

Contents

1 Probability **5**
- 1.1 Random Variables . . . 5
- 1.2 Inequalities . . . 5
- 1.3 Convergence of Random Variables . . . 6

2 Parametric Inference **9**
- 2.1 Sufficient Statistic . . . 10
- 2.2 Exponential Family . . . 13
- 2.3 Fisher Information . . . 14
- 2.4 Ancillarity and Completeness . . . 17
  - 2.4.1 Ancillarity . . . 17
  - 2.4.2 Completeness . . . 19
- 2.5 Estimation - Methods of Moments . . . 20
- 2.6 Estimation - Maximum Likelihood . . . 20
  - 2.6.1 Proof of Asymptotic Normality . . . 20
  - 2.6.2 Delta Method . . . 21
- 2.7 Methods for Evaluating Estimators . . . 22
  - 2.7.1 Mean Squared Error . . . 22
  - 2.7.2 Uniformly minimum variance unbiased estimators (UMVUE) . . . 23
- 2.8 Cramér-Rao Lower Bound . . . 29
- 2.9 Hypothesis Testing . . . 31
  - 2.9.1 Wald Test . . . 32
  - 2.9.2 Pearson's χ^2 test . . . 32
  - 2.9.3 Permutation Test . . . 33
  - 2.9.4 The Likelihood Ratio Test . . . 33

3 The Bootstrap **34**

4 EM Algorithm **38**
- 4.1 Foundations . . . 38
- 4.2 Separate Mixture of Normal Distributions . . . 38
- 4.3 Expectation and Maximization of Likelihoods . . . 40

5 Simulation Methods **45**
- 5.1 Monte Carlo Integration . . . 45
- 5.2 Importance Sampling . . . 47

5.3 Accept-Reject Algorithm 47
5.4 Markov Chain Monte Carlo (MCMC) 48
5.4.1 The Metropolis-Hastings Algorithm 49
5.4.2 Gibbs Sampling 50

6 Selected Topics **51**
6.1 Principal Component Analysis 51
6.2 Clustering 53

Part I

Probability

1 Probability

We record here only a small portion of probability theory. Readers should consult Wasserman 2004, Ash 1999; Chung and AitSahlia 1974; Durrett 2010; Resnick 2014 or Shreve 2004 for a complete study.

1.1 Random Variables

Given a probability space $(\Omega, \sigma(\Omega), \mathbb{P})$, a function $X : \Omega \to \mathbb{R}$ is called a *random variable* if

$$\forall A \in \mathcal{B}(\mathbb{R}), \quad X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\} \in \sigma(\Omega).$$

For a random variable X, its *cumulative distribution function (CDF)* is the function $F_X : \mathbb{R} \to [0, 1]$ such that

$$F_X(x) = \mathbb{P}\{X \leq x\}.$$

The following theorem tells us that the cumulative distribution function completely determines the distribution of a random variable.

Theorem 1.1. Let X have CDF F and let Y have CDF G. If $F(x) = G(x)$ for all x then $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$ for all (measurable) A.

1.2 Inequalities

Theorem 1.2 (Markov's Inequality)**.** Let X be a non-negative random variable and suppose $\mathbb{E}(X) < \infty$. For any $t > 0$,

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}(X)}{t}$$

Proof. $\mathbb{E}(X) = \int_0^\infty x f(x)dx = \int_0^t x f(x)dx + \int_t^\infty x f(x)dx \geq \int_t^\infty x f(x)dx \geq t \int_t^\infty f(x)dx = t\,\mathbb{P}(X > t)$. □

Theorem 1.3 (Chebyshev's Inequality)**.** Let $\mu = \mathbb{E}(X)$ and $\sigma^2 = \mathbb{V}(X)$. Then

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2} \quad \text{and} \quad \mathbb{P}(|Z| \geq k) \leq \frac{1}{k^2}$$

where $Z = (X - \mu)/\sigma$. In particular, $\mathbb{P}(|Z| > 2) \leq 1/4$ and $\mathbb{P}(|Z| > 3) \leq 1/9$.

Proof. Use Markov's inequality:

$$\mathbb{P}(|X - \mu| \geq t) = \mathbb{P}(|X - \mu|^2 \geq t^2) \leq \frac{\mathbb{E}(X - \mu)^2}{t^2} = \frac{\sigma^2}{t^2}.$$

The second part follows by setting $t = k\sigma$. □

Theorem 1.4 (Hoeffding's Inequality)**.** Let $Y_1, \ldots, Y_n$ be independent observations such that $\mathbb{E}(Y_i) = 0$ and $a_i \leq Y \leq b_i$. Let $\epsilon > 0$. Then for any $t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i \geq \epsilon\right) \leq e^{-t\epsilon} \prod_{i=1}^{n} e^{t^2(b_i-a_i)^2}/8.$$

Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Then, for any $\epsilon > 0$,

$$\mathbb{P}(|\bar{X}_n - p| > \epsilon) \leq 2e^{-2n\epsilon^2}$$

where $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_n$.

Theorem 1.5 (Cauchy-Schwarz inequality)**.** If X and Y have finite variances then

$$\mathbb{E}|XY| \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}.$$

Theorem 1.6 (Jensen's Inequality)**.** If g is convex then

$$\mathbb{E}g(X) \geq g(\mathbb{E}X).$$

If g is concave then

$$\mathbb{E}g(X) \leq g(\mathbb{E}X).$$

Proof. Let $L(x) = a + bx$ be a line, tangent to $g(x)$ at the point $(\mathbb{E}X, g(\mathbb{E}X))$. Since g is convex, g lies above the line $L(x)$, so $g(x) \geq L(x)$ for all x. Thus

$$\mathbb{E}g(X) \geq \mathbb{E}L(X) = \mathbb{E}(a + bX) = a + b\mathbb{E}(X) = L(\mathbb{E}(X)) = g(\mathbb{E}(X)).$$

□

1.3 Convergence of Random Variables

There are several types of convergence in probability theory. See Fig. 1 for their relationship.

- A sequence of random variables $X_1, X_2, \ldots$ converges *almost surely* to X if

  $$\mathbb{P}\{\omega : X_n(\omega) \to X(\omega)\} = 1,$$

  or using a short-hand notation

  $$\mathbb{P}\{\lim_{n\to\infty} X_n = X\} = 1.$$

- **converge in probability**: $X_n \xrightarrow{P} X$ if for every $\epsilon > 0$,

  $$\mathbb{P}(|X_n - X| > \epsilon) \to 0$$

  as $n \to \infty$.

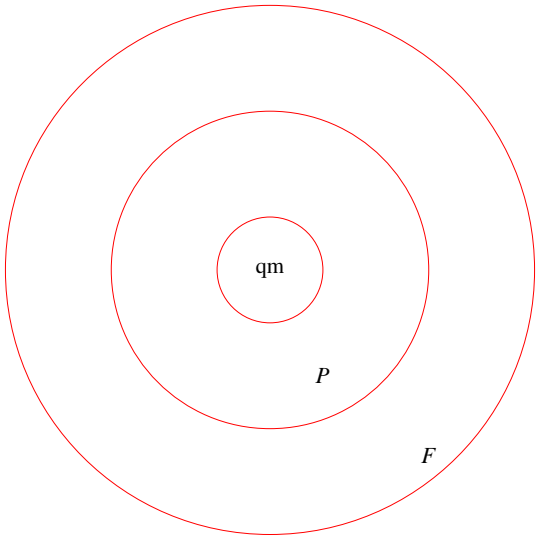

Figure 1: Relationship among different types of convergence.

- **converge in distribution**: $X_n \rightsquigarrow X$ if

$$\lim_{n\to\infty} F_n(t) = F(t)$$

  at all t for which F is continuous.

- **converge in quadratic mean (converge in L_2)**: $X_n \xrightarrow{qm} X$ if $\mathbb{E}(X_n - X)^2 \to 0$ as $n \to \infty$.

Theorem 1.7 (Weak Law of Large Numbers)**.** If $X_1, \ldots, X_n, \ldots$ are iid, then $\bar{X}_n \xrightarrow{P} \mu$.

Proof. Assume $\sigma < \infty$. Using Chebyshev's inequality,

$$\mathbb{P}(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\mathbb{V}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

which tends to 0 as $n \to \infty$. □

Theorem 1.8 (Strong Law of Large Numbers)**.** $\bar{X}_n \xrightarrow{\text{a.s.}} \mu$ as $n \to \infty$, i.e.

$$\mathbb{P}\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1$$

Proof. More complicated. □

Theorem 1.9 (Central Limit Theorem)**.** Let $X_1, X_2, \ldots$ be iid with mean μ and variance σ^2. Then

$$Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \rightsquigarrow Z$$

where $Z \sim N(0, 1)$.

Part II

Statistical Inference

2 Parametric Inference

A statistical model $\mathfrak{F}$ is a set of distributions. A parametric model is a set $\mathfrak{F}$ that can be parameterized by a finite number of parameters. In general a parametric model takes the form

$$\mathfrak{F} = \left\{ f(x, \theta) : \theta \in \Theta \right\}$$

where θ is an unknown parameter or a vector of parameters that can take value in the parameter space Θ. A statistical functional is any functional on $\mathfrak{F}$:

$$T : \mathfrak{F} \to \mathbb{R}.$$

Examples are mean $\mu(F) = \int x dF(x)$, variance $\sigma^2(F) = \int (x-\mu)^2 dF(x) = \int x^2 dF(x) - \left(\int x dF(x)\right)^2$ and median $m(F) = F^{-1}(1/2)$.

A point estimator of a parameter θ (or a statistics) is some function of the data $\{X_1, \ldots, X_n\}$, i.e. some $T_n = g(X_1, \ldots, X_n)$. The distribution of T_n is called the sampling distribution, and the standard deviation of T_n is called the standard error, denoted by se. The bias of an estimator is defined as $B_T(\theta) = \mathbb{E}_\theta[T] - \theta$. The estimator is called *unbiased* if $B_T(\theta) = 0$, and *consistent* if

$$T_n \xrightarrow{p} \theta \quad \text{as} \quad n \to \infty.$$

The mean squared error of an estimator is defined as $MSE_T(\theta) = \mathbb{E}_\theta\left[(T-\theta)^2\right]$. It is shown in Proposition 2.31 that the MSE can be written as

$$MSE_T(\theta) = \mathbb{V}_\theta(T) + B_T^2(\theta).$$

There is in general a bias-variance trade-off when we select estimators. If an estimator has low variance, then it typically has high bias (think about the mean). Conversely, if an estimator has low bias, then it may have high variance (think about perfect fitting). If $B_T(\theta) \to 0$ and $se = \sqrt{\mathbb{V}_\theta(T)} \to 0$ as $n \to \infty$ then $MSE_T(\theta) = \mathbb{E}_\theta\left[(T-\theta)^2\right] \to 0$, i.e. $T_n \xrightarrow{qm} \theta$, so that in particular $T_n \xrightarrow{p} \theta$, i.e. the estimator T_n is consistent. We see that consistency is generally a weaker condition than unbiasedness.

A $1-\alpha$ *confidence interval* for a parameter θ is an interval $C_n = (a, b)$ where $a = a(X_1, \ldots, X_n)$ and $b = b(X_1, \ldots, X_n)$ are functions of the data, and

$$\mathbb{P}(\theta \in C_n) \geq 1 - \alpha \quad \text{for all } \theta \in \Theta.$$

Definition 2.1. The empirical distribution function $\hat{F}_n$ is the CDF that puts mass $1/n$ at each data point X_i, i.e.

$$\hat{F}_n(x) = \frac{\sum_{i=1}^n I(X_i \leq x)}{n}.$$

Note that $\hat{F}_n$ is a function of $X_1, \ldots, X_n$.

We have the following theorems:

Theorem 2.2. At any fixed value for x:

- $\mathbb{E}(\hat{F}_n(x)) = F(x)$;
- $\mathbb{V}(\hat{F}_n(x)) = \dfrac{F(x)(1-F(x))}{n} \to 0$;
- $\hat{F}_n(x) \xrightarrow{p} F(x)$.

Theorem 2.3 (Glivenko-Cantelli Theorem)**.** Let $X_1, \ldots, X_n \sim F$. Then

$$\|\hat{F}_n(x) - F(x)\|_\infty = \sup_x |\hat{F}_n(x) - F(x)| \to 0$$

almost surely.

Definition 2.4. The plug-in estimator of $\theta = T(F)$ is defined by

$$\hat{\theta}_n = T(\hat{F}_n).$$

If T is linear in F, i.e. $T(F) = \int r(x) dF(x)$ for some function r, then the plug-in estimator is just $(1/n)\sum_{i=1}^n r(X_i)$. Sample mean is linear: $\hat{\theta}_n = T(\hat{F}_n) = (1/n)\sum_{i=1}^n X_i$. Variance is not:

$$\begin{aligned}
\hat{\sigma}^2 &= \int x^2 d\,\hat{F}_n(x) - \left(\int x d\,\hat{F}_n(x)\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2.
\end{aligned}$$

2.1 Sufficient Statistic

We use $\underline{X} = (X_1, \ldots, X_n)$ to denote our sample. A statistic $T : \mathcal{X} \to \mathcal{T}$ is a function mapping the set of all possible sample realizations to the set of possible values that T can take on.

Definition 2.5. T is *sufficient* for θ (or, more precisely, for the statistical model $\{f_\theta : \theta \in \Theta\}$) if and only if the conditional distribution of $\underline{X}$ given $T = t$ does not depend on θ for all $t \in \mathcal{T}$.

Example 2.6. Let $\underline{X} = (X_1, \ldots, X_n)$ with $X_i \overset{i.i.d}{\sim} Be(\theta)$ for $i = 1, \ldots, n$ and $\theta \in \Theta = (0, 1)$. We show that $T = \sum_{i=1}^n X_i$ is a sufficient statistic. We have

$$\mathbb{P}\{X_1 = x_1, \ldots, X_n = x_n | T = t\} = \frac{\mathbb{P}\left(\{X_1 = x_1, \ldots, X_n = x_n\} \cap \{T = t\}\right)}{\mathbb{P}\{T = t\}}. \tag{1}$$

Note that T has binomial distribution $bi(n,\theta)$. When $\sum_{i=1}^n x_i \neq t$, the two events $A = \{X_1 = x_1, \ldots, X_n = x_n\}$ and $B = \{T = t\}$ are disjoint, so that the probability $\mathbb{P}(A \cap B) = \mathbb{P}(\varnothing)$ is 0. If $\sum_{i=1}^n x_i = t$, then $A \subset B$, so that the conditional probability becomes

$$\begin{aligned}\frac{\mathbb{P}\{X_1 = x_1, \ldots, X_n = x_n\}}{\mathbb{P}\{T = t\}} &= \frac{\prod_{i=1}^n \mathbb{P}\{X_i = x_i\}}{\mathbb{P}\{T = t\}} \\ &= \frac{\prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\ &= \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\ &= \frac{1}{\binom{n}{t}}.\end{aligned}$$

Putting together we found

$$\mathbb{P}\{X_1 = x_1, \ldots, X_n = x_n | T = t\} = \begin{cases} 0 & \text{if } \sum_{i=1}^n x_i \neq t \\ \binom{n}{t}^{-1} & \text{if } \sum_{i=1}^n x_i = t \end{cases}$$

Since the conditional distribution does not depend on θ, we have established that $T = \sum_{i=1}^n X_i$ is a sufficient statistic for θ.

The definition of sufficient statistic provides a conceptual support for the notion of sufficiency. But

1. for more elaborated models it is not easy to verify;
2. it is not constructive in the sense that it only allows to verify whether a statistics is sufficient, but it does not provide a method to find a sufficient one.

To address the problems we have the following factorization theorem:

Theorem 2.7 (Neyman-Fisher Factorization Theorem)**.** $T(\underline{\mathrm{X}})$ is sufficient for θ if and only if there exist two non-negative functions $g(t,\theta)$ and $h(\underline{\mathrm{x}})$ s.t.

$$f_\theta^{\underline{\mathrm{X}}}(\underline{\mathrm{x}}) = g(t,\theta)h(\underline{\mathrm{x}}) \quad \forall \underline{\mathrm{x}} \in \mathcal{X}, \forall \theta \in \Theta$$

where $t = T(\underline{\mathrm{x}})$.

Proof. ("only if" part for the discrete case) If T is sufficient, then

$$\begin{aligned} f_\theta^{\underline{\mathrm{X}}}(\underline{\mathrm{x}}) = \mathbb{P}_\theta\{\underline{\mathrm{X}} = \underline{\mathrm{x}}\} &= \mathbb{P}_\theta\{\underline{\mathrm{X}} = \underline{\mathrm{x}}, T = t\} \\ &= \underbrace{\mathbb{P}\{\underline{\mathrm{X}} = \underline{\mathrm{x}} | T = t\}}_{=h(\underline{\mathrm{X}})} \underbrace{\mathbb{P}_\theta\{T = t\}}_{=g(t,\theta)}. \end{aligned}$$

□

We can use the factorization theorem to establish the following:

- $T = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is a sufficient statistic for $N(\mu, \sigma^2)$,
- $T = \sum_{i=1}^n X_i$ is a sufficient statistic for Poisson distribution $P(\theta)$, and
- $T = X_{(n)}$ is a sufficient statistic for uniform distribution $\mathcal{U}(0, \theta)$:

Example 2.8 (Normal Distribution)**.**

$$\begin{aligned} f_\theta^{\underline{\mathrm{X}}}(\underline{\mathrm{x}}) &= \prod_{i=1}^n \left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(x_i - \mu)^2 \right\} \right) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\} \\ &= \underbrace{\left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp\left\{ -\frac{1}{2\sigma^2} \left[\sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i + \mu^2 \right] \right\}}_{= g(t,\theta) = g((\sum_{i=1}^n x_i^2, \sum_{i=1}^n x_i), (\mu, \sigma^2)) \text{ with } h(\underline{\mathrm{X}}) = 1} \end{aligned}$$

Example 2.9 (Poisson Distribution)**.**

$$f_\theta^{\underline{\mathrm{X}}}(\underline{\mathrm{x}}) = \prod_{i=1}^n \frac{\theta^{x_i} e^{-\theta}}{x_i!} = \frac{\theta^{\sum_{i=1}^n x_i} e^{-n\theta}}{\prod_{i=1}^n x_i!} = \underbrace{\left(\frac{1}{\prod_{i=1}^n x_i!} \right)}_{= h(\underline{\mathrm{X}})} \underbrace{\left(\theta^{\sum_{i=1}^n x_i} e^{-n\theta} \right)}_{g(t,\theta) = g(\sum_{i=1}^n x_i, \theta)}.$$

Example 2.10 (Uniform Distribution)**.**

$$f_\theta^{\underline{\mathrm{X}}}(\underline{\mathrm{x}}) = \prod_{i=1}^n \frac{1}{\theta} \mathbb{1}_{(0,\theta)}(x_i) = \frac{1}{\theta^n} \prod_{i=1}^n \mathbb{1}_{(0,\theta)}(x_i).$$

Now $\prod_{i=1}^n \mathbb{1}_{(0,\theta)}(x_i) = 1$ if and only if all x_i's are in $(0, \theta)$, or in other terms if and only if $0 < x_{(1)} < \cdots < x_{(n)} < \theta$, or $0 < x_{(1)} < x_{(n)} < \theta$. Thus we can write

$$f_\theta^{\underline{\mathrm{X}}}(\underline{\mathrm{x}}) = \underbrace{\mathbb{1}_{(0,x_{(n)})}(x_{(1)})}_{= h(\underline{\mathrm{X}})} \underbrace{\frac{1}{\theta^n} \mathbb{1}_{(0,\theta)}(x_{(n)})}_{= g(t,\theta) = g(x_{(n)}, \theta)}.$$

There can be many sufficient statistic for a parameter. The "minimal" one is the one that induces the coarsest partition of the set of sample realizations $\mathcal{X}$.

Proposition 2.11. Assume T is sufficient for θ and let l be an arbitrary function. If $T = l(T^*)$ then T^* is also sufficient.

Proof. By the factorization theorem

$$f_\theta^{\underline{X}}(\underline{x}) = g(T(\underline{x}), \theta)h(\underline{x}) = g(l(T^*(\underline{x})), \theta)h(\underline{x}) = g^*(T^*(\underline{x}), \theta)h(\underline{x})$$

where $g^* = g \circ l$. □

Definition 2.12. A sufficient statistic T is said to be *minimal* if T is a function of any other sufficient statistic T^*, i.e. for every sufficient statistic T^* there exists l^* s.t. $T = l^*(T^*)$. Equivalently, the partition induced by the minimal sufficient statistic is the one with the least number of elements.

The minimal sufficient statistic is unique up to one-to-one transformations. The following theorem can be used to test minimal sufficiency:

Theorem 2.13 (Lehmann-Scheffé)**.** If T is a statistic s.t.

$$\frac{f_\theta^{\underline{X}}(\underline{x})}{f_\theta^{\underline{X}}(\underline{y})} \text{ with } f_\theta^{\underline{X}}(\underline{y}) \neq 0$$

does not depend on θ if and only if $T(\underline{x}) = T(\underline{y})$, then T is minimal sufficient.

It can be shown that

- $T = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is minimal sufficient for $N(\mu, \sigma^2)$. The one-to-one transformations $T' = \left(\bar{X}, S^2 = \sum_{i=1}^n (X_i - \bar{X})^2\right)$ is also minimal sufficient.
- $T = (X_{(1)}, X_{(n)})$ is minimal sufficient for $\mathcal{U}(\theta, \theta + 1)$.

2.2 Exponential Family

Definition 2.14. $X \sim f_\theta$ with $\theta \in \Theta \subseteq \mathbb{R}$. We denote the support of f_θ by S_X. One-parameter exponential family:

$$f_\theta(x) = a(\theta)g(x)\exp\{b(\theta)R(x)\}$$

or equivalently

$$f_\theta(x) = g(x)\exp\{b(\theta)R(x) + c(\theta)\}$$

Note that since $e^x > 0$, the support of $f_\theta(x)$ does not depend on θ. Examples: Bernoulli, normal, beta, Poisson, exponential, Dirichlet, gamma, chi-squared, geometric.

Proposition 2.15. If X belongs to the one-parameter exponential family, then

$$\mathbb{E}_\theta[R(X)] = -\frac{c'(\theta)}{b'(\theta)}, \quad \mathbb{V}_\theta[R(X)] = -\frac{1}{b'(\theta)}\frac{d}{d\theta}\frac{c'(\theta)}{b'(\theta)}.$$

Proof. Differentiate both sides of $\int_{S_X} f_\theta(x)dx = 1$ and interchange derivative and integral:

$$\frac{\partial}{\partial\theta}\int_{S_X} f_\theta(x)dx = 0 \Longrightarrow \int_{S_X} \frac{\partial}{\partial\theta} f_\theta(x)dx = 0$$

and then use $f_\theta(x) = g(x)\exp\{b(\theta)R(x)+c(\theta)\}$ to expand the integral. One find $b'(\theta)\mathbb{E}_\theta[R(X)]+c'(\theta) = 0$. □

An important property of exponential family is closure under random sampling: if $X \sim f_\theta$ belongs to the exponential family, then the distribution of the samples $X_1, \ldots, X_n$ also belongs to the exponential family:

$$f_\theta^{\underline{X}}(\underline{x}) = \prod_{i=1}^{N} f_\theta(x_i) = \prod_{i=1}^{n} g(x_i)\exp\left\{b(\theta)\sum_{i=1}^{n} R(x_i) + nc(\theta)\right\}.$$

Apply the Neyman-Fisher factorization theorem we immediately get that $T = \sum_{i=1}^{n} R(x_i)$ is a sufficient statistic for θ.

Generalizations to vector parameter:

$$f_\theta(x) = g(x)\exp\left\{\sum_{j=1}^{k} b_j(\theta)R_j(x) + c(\theta)\right\}$$

where $\theta = (\theta_1, \ldots, \theta_k) \in \Theta \subseteq \mathbb{R}^k$. The family is said to be *minimal* if the functions $\{b_j(\theta)\}_{j=1,\ldots,k}$ are linearly independent and the functions $\{R_j(x)\}_{j=1,\ldots,k}$ are linearly independent. It is said to be *full-rank* if it is minimal and Θ includes a proper subset of $\mathbb{R}^k$.

$X \sim N(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R}\times\mathbb{R}^+$ belongs to the 2-parameter exponential family. $X \sim N(\theta, \theta^2)$ with $\theta \in \mathbb{R}$ belongs to the 2-parameter exponential family and is in minimal form, but is not full rank, since the parameter space degenerates to $\mathbb{R}$.

The multi-parameter exponential family is also closed under random sampling, and a general result is

Proposition 2.16. If X belongs to the k-parameter exponential family, then

$$T = \left(R_1^*(\underline{X}), \ldots, R_k^*(\underline{X})\right) = \left(\sum_{i=1}^{n} R_1(X_i), \sum_{i=1}^{n} R_2(X_i), \ldots, \sum_{i=1}^{n} R_k(X_i)\right)$$

is the minimal sufficient statistic for $\theta = (\theta_1, \ldots, \theta_k)$. Moreover, the distribution of T also belongs to the k-parameter exponential family.

2.3 Fisher Information

We first consider the one dimensional case $\Theta \subseteq \mathbb{R}$:

Definition 2.17 (Score function)**.** Let $\mathcal{P} = \{f_\theta : \theta \in \Theta\}$ be a C.R. regular statistical model, where $\Theta \subseteq \mathbb{R}$. The function

$$S_\theta(x) = \frac{\partial}{\partial\theta} \log f_\theta(x) = \frac{f'_\theta(x)}{f_\theta(x)}$$

is termed the *score function*.

The score function measures how sensitive the log likelihood function is to its parameter θ at a given x. Replace x by X we get the random variable

$$S_\theta(X) = \frac{\partial}{\partial\theta} \log f_\theta(X).$$

Since the data point X is generated by f_θ, we would expect that $f_\theta(X)$ attains its maximum for this particular θ. This implies that we should expect $\partial(\log f_\theta(X))/\partial\theta$ to be 0.

For example, we can calculate the score function for the exponential distribution $Exp(\lambda)$. Recall for the exponential distribution $f_\lambda(x) = \lambda e^{-\lambda x}$. Thus $\log f_\lambda(x) = \log\lambda - \lambda x$. The score function is thus $S_\lambda(X) = \frac{1}{\lambda} - X$. The expectation of $S_\lambda(X)$ is 0, and the variance of $S_\lambda(X)$ is equal to the variance of X, namely $1/\lambda^2$.

Lemma 2.18. If $\mathcal{P}$ is C.R. regular, then we have

1. $\mathbb{E}_\theta[S_\theta(X)] = 0 \quad \forall\theta \in \Theta$
2. $\mathbb{V}_\theta[S_\theta(X)] = \mathbb{E}_\theta[S_\theta^2(X)] = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)^2\right]$

Proof. Again, we use $\int f_\theta(x) = 1 \Longrightarrow \frac{\partial}{\partial\theta}\int f_\theta(x)dx = 0$ to obtain

$$0 = \frac{\partial}{\partial\theta}\int f_\theta(x)dx = \int \frac{\partial}{\partial\theta} f_\theta(x)dx = \int \frac{f'_\theta(x)}{f_\theta(x)} f_\theta(x)dx = \int S_\theta(x) f_\theta(x)dx = \mathbb{E}_\theta[S_\theta(X)].$$

□

For a given θ, if $S_\theta(X)$ always concentrates near 0, then this implies that the distribution f_θ is very *flat*, namely it is nearly a constant, so that the derivative is close to 0. Different values of X appear with similar probabilities. If, on the other hand, $S_\theta(X)$ varies a lot with different X, then this means there are sharp increases or sharp decreases in f_θ, so that the distribution is uneven. Thus the variance of $S_\theta(X)$ can provide some information about the distribution f_θ.

Definition 2.19 (Fisher Information)**.** Let $\mathcal{P} = \{f_\theta : \theta \in \Theta\}$ be a C.R. regular statistical model. The variance of the score function is called the *Fisher information*:

$$I^X(\theta) = \mathbb{V}_\theta[S_\theta(X)].$$

Additivity of Fisher information:

Proposition 2.20. If X and Y are two independent random variables with C.R. regular distributions, then

$$I^{(X,Y)}(\theta) = I^X(\theta) + I^Y(\theta).$$

Proof. We use $f_\theta^{(X,Y)}(X,Y) = f_\theta^X(X) f_\theta^Y(Y)$ and $\log f_1 f_2 = \log f_1 + \log f_2$, and the fact that if X and Y are independent then $f_1(X)$ and $f_2(Y)$ are independent for any two measurable functions f_1 and f_2:

$$\begin{aligned}
\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_\theta^{(X,Y)}(X,Y)\right)^2\right] &= \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_\theta^X(X) f_\theta^Y(Y)\right)^2\right] \\
&= \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_\theta^X(X) + \frac{\partial}{\partial\theta}\log f_\theta^Y(Y)\right)^2\right] \\
&= \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_\theta^X(X)\right)^2\right] + \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_\theta^Y(Y)\right)^2\right] \\
&\quad + 2\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f_\theta^X(X)\frac{\partial}{\partial\theta}\log f_\theta^Y(Y)\right] \\
&= I^X(\theta) + I^Y(\theta) + 2\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f_\theta^X(X)\right]\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f_\theta^Y(Y)\right] \\
&= I^X(\theta) + I^Y(\theta).
\end{aligned}$$

□

Corollary 2.21. The Fisher information of a random sample $\underline{X} = (X_1, \ldots, X_n)$ with $X_i \overset{i.i.d.}{\sim} f_\theta$ (C.R. regular) is

$$I^{\underline{X}}(\theta) = n I^{X_1}(\theta).$$

Further simplifications in the computation of the Fisher information:

Lemma 2.22. If the model is C.R. regular then

$$I^X(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f_\theta(X)\right]$$

Proof. First note

$$\begin{aligned}
\frac{\partial^2}{\partial\theta^2}\log f_\theta(x) = \frac{\partial}{\partial\theta}\left[\frac{f_\theta'(x)}{f_\theta(x)}\right] &= \frac{f_\theta''(x) f_\theta(x) - (f_\theta'(x))^2}{f_\theta(x)^2} \\
&= -\left(\frac{f_\theta'(x)}{f_\theta(x)}\right)^2 + \frac{f_\theta''(x)}{f_\theta(x)} = -\left(\frac{\partial}{\partial\theta}\log f_\theta(x)\right)^2 + \frac{f_\theta''(x)}{f_\theta(x)}.
\end{aligned}$$

Take the expectation we get

$$\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f_\theta(X)\right] = -\underbrace{\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)^2\right]}_{=I^X(\theta)} + \underbrace{\mathbb{E}_\theta\left[\frac{f_\theta''(X)}{f_\theta(X)}\right]}_{=0}.$$

The last term is zero due to

$$\mathbb{E}_\theta\left[\frac{f_\theta''(X)}{f_\theta(X)}\right] = \int \frac{f_\theta''(x)}{f_\theta(x)} f_\theta(x)dx = \int f_\theta''(x)dx = \frac{\partial^2}{\partial\theta^2}\underbrace{\int f_\theta(x)dx}_{=1} = 0.$$

□

Example 2.23. As an example, we calculate the Fisher information for a sample $\underline{\mathrm{X}} = (X_1, \ldots, X_n)$ where $X_i \overset{i.i.d.}{\sim} f_\lambda(x) = \lambda e^{-\lambda x}$. The first derivative of $\log f_\lambda(x)$ is $1/\lambda - x$, so the second derivative of $\log f_\lambda(x)$ is

$$\frac{\partial^2}{\partial\lambda^2}\log f_\lambda(x) = \frac{\partial}{\partial\lambda}\left(\frac{1}{\lambda} - x\right) = -\frac{1}{\lambda^2}.$$

Thus we have

$$I^{X_1}(\lambda) = -\mathbb{E}_\lambda\left[\frac{1}{\lambda^2}\right] = \frac{1}{\lambda^2}.$$

So the Fisher information of the sample is $I^{\underline{\mathrm{X}}} = n/\lambda^2$.

The following result is very useful:

Theorem 2.24. (Assume every distribution involved is C.R. regular) Suppose $T = T(\underline{\mathrm{X}})$ is a statistic of a sample $\underline{\mathrm{X}} = (X_1, \ldots, X_n)$. Then

$$I^{\underline{\mathrm{X}}}(\theta) \geq I^T(\theta) \quad \forall\theta \in \Theta$$

with equality holds if and only if T is sufficient.

The theorem provides great convenience, since if T is sufficient, then to calculate the Fisher information of T we don't have to work out the distribution of T. We can just compute $I^{X_1}(\theta)$ instead. We then have $I^T(\theta) = I^{\underline{\mathrm{X}}}(\theta) = nI^{X_1}(\theta)$.

2.4 Ancillarity and Completeness

2.4.1 Ancillarity

Definition 2.25 (Ancillary statistic)**.** A statistic $V = V(X_1, \ldots, X_n)$ is said to be *ancillary* for θ if the distribution of V does not depend on θ.

For example, suppose our sample is $X_1, \ldots, X_n$ with $X_i \overset{i.i.d.}{\sim} N(\theta, 1)$. Then $T = X_1 - X_2 \sim N(0, 2)$ is ancillary.

- We say X belongs to a *location family* if $X \sim F(x-\theta)$ where F is some known distribution. If X is continuous then this is also equivalent to $f_\theta^X(x) = f(x-\theta)$ for some known f. Think about the normal distributions with different means. To construct a location family, let $Z \sim F$ and let $X = Z + \theta$. Then

$$F_\theta^X(x) = \mathbb{P}\{X \leq x\} = \mathbb{P}\{Z + \theta \leq x\} = \mathbb{P}\{Z \leq x - \theta\} = F(z - \theta).$$

- Ancillary statistics for samples of X belonging to a location family could be
  - $T = X_i - X_j$ for $i \neq j$.
  - $T = X_{(i)} - X_{(j)}$ for $i \neq j$.
- We say X belongs to a *scale family* if $X \sim F(\frac{x}{\theta})$ for some known F. Equivalently, $f_\theta^X(x) = \frac{1}{\theta} f(\frac{x}{\theta})$ for some known f. To construct one, we can do $X = \theta Z$ where $Z \sim F$ is known. Then

$$F_\theta^X(x) = \mathbb{P}\{X \leq x\} = \mathbb{P}\{\theta Z \leq x\} = \mathbb{P}\{Z \leq \frac{x}{\theta}\} = F(\frac{x}{\theta}).$$

  Examples: $X \sim N(0, \sigma^2)$, which has pdf

$$f_{\sigma^2}^X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2\right\}.$$

- Ancillary statistics for samples of X belonging to a scale family could be
  - $T = \left(\sum_{i=1}^n X_i\right) / X_n$.
  - $T = \left(\frac{X_1}{X_n}, \ldots, \frac{X_{n-1}}{X_n}\right)$.
- We say that X belongs to a *location-scale* family if

$$X \sim F\left(\frac{x-\mu}{\sigma}\right)$$

  for F known and $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$. An example would be the normal distribution $X \sim N(\mu, \sigma^2)$. To construct a location-scale family, we can do $X = \sigma Z + \mu$ with $Z \sim F$ known.
- Ancillary statistics for samples of X belonging to a location-scale family could be
  - $T = \left(\frac{X_1 - X_n}{S}, \ldots, \frac{X_{n-1} - X_n}{S}\right)$, where $S^2 = (1/n)\sum_{i=1}^n (X_i - \bar{X})^2$ is the sample variance.

2.4.2 Completeness

Completeness of T corresponds to requiring that there does not exist any function of T that can lead to 1st order ancillarity, except the constant function. If T is rendered to a constant, it is of course irrelevant of θ and hence ancillary.

Definition 2.26. T is complete if

$$\mathbb{E}_\theta[g(T)] \text{ does not depend on } \theta \implies \text{ it must be that } g \text{ is a constant function (a.s.)}$$

Or equivalently

$$\mathbb{E}_\theta[g(T)] = 0 \quad \forall \theta \in \Theta \implies \mathbb{P}\{g(T) = 0\} = 1.$$

We can show that $T = \sum_{i=1}^n X_i$ is complete for Bernoulli sample $X_i \sim Be(\theta)$, $T = X_{(n)}$ is complete for uniform sample $X_i \sim \mathcal{U}(0, \theta)$, and $T = \bar{X}$ is complete for the normal sample $X_i \sim N(\mu, \sigma^2)$. For the k-parameter exponential family,

$$T = \left(\sum_{i=1}^n R_1(X_i), \sum_{i=1}^n R_2(X_i), \ldots, \sum_{i=1}^n R_k(X_i) \right)$$

is complete.

Why we want completeness? We can view ancillarity as the opposite of sufficiency. However, even if a statistic T is minimal sufficient, there may exist one-to-one transformation of T such that ancillarity appears. An example is $T = (X_{(1)}, X_{(n)})$ for the distribution $X_i \sim \mathcal{U}(\theta, \theta + 1)$. The transformation $T^* = (X_{(n)} - X_{(1)}, (X_{(1)} + X_{(n)})/2)$ is also minimal sufficient, but it contains ancillary statistic $X_{(n)} - X_{(1)}$.

Thus, we can view completeness as "removing ancillarity from sufficient statistic". If T is sufficient and complete, we should expect that it is independent from ancillary statistic.

Theorem 2.27 (Basu's Theorem)**.** If T is a sufficient and complete statistic for $\{f_\theta : \theta \in \Theta\}$ and V is any ancillary statistic, then T and V are independent.

Proof. We prove for the discrete case. Since $\mathbb{P}\{\underline{X} = \underline{x} | T = t\}$ does not depend on θ, and V is a function of $\underline{X}$, we have $\mathbb{P}\{V \in B | T = t\} := h_B(t) = \mathbb{E}[\mathbb{1}_B(V) | T = t]$ does not depend on θ. By law of iterated expectation, the expected value of $h_B(T)$ is $\mathbb{E}[\mathbb{1}_B(V)] = \mathbb{P}\{V \in B\}$. Since T is complete, the function h_B is constant (a.s.) and is equal to $\mathbb{P}\{V \in B\}$. In other words,

$$\mathbb{P}\{V \in B | T = t\} = \mathbb{P}\{V \in B\} \quad \text{a.s.}$$

□

An application of Basu's theorem would be to prove $\bar{X}$ and S^2 are independent for the normal sample $X_i \sim N(\mu, \sigma^2)$. For every fixed σ^2, $\bar{X}$ is (minimal) sufficient and complete, while S^2 is ancillary ($S^2 = (1/n)\sum_{i=1}^n (X_i - \bar{X})^2 = (1/n)\sum_{i=1}^n (Z_i - \bar{Z})^2$ with $Z_i = X_i - \mu \sim N(0, \sigma^2)$), so that the two are independent (for σ^2 known or unknown).

2.5 Estimation - Methods of Moments

Suppose the model is $X \sim f_\theta$ where $\theta = (\theta_1, \ldots, \theta_k) \in \Theta \subseteq \mathbb{R}^k$, so we have k parameters to estimate. The analytical formulas of the moments $\mathbb{E}_\theta[X^j] = \int x^j f_\theta$, $j = 1, 2, \ldots$ are known and depend on $\theta = (\theta_1, \ldots, \theta_k)$. The empirical moments are

$$m_j = \frac{1}{n}(X_1^j + X_2^j + \ldots + X_n^j) \quad j = 1, 2, \ldots.$$

The methods of moments equates the first k analytical moments with the corresponding empirical ones, and then solve for $\theta_1, \ldots, \theta_k$.

2.6 Estimation - Maximum Likelihood

Properties of maximum likelihood estimation:

1. The MLE is *consistent*: $\widehat{\theta}_n \xrightarrow{p} \theta$
2. The MLE is *equivariant*: if $\widehat{\theta}_n$ is the MLE of θ then $g(\widehat{\theta}_n)$ is the MLE of $g(\theta)$.
3. The MLE is *asymptotically normal*:
   (a) $se = \sqrt{\mathbb{V}(\widehat{\theta}_n)} \approx \sqrt{1/I_n(\theta)}$ and $(\widehat{\theta}_n - \theta)/se \rightsquigarrow N(0, 1)$.
   (b) Let $\widehat{se} = \sqrt{1/I_n(\widehat{\theta}_n)}$, then $(\widehat{\theta}_n - \theta)/\widehat{se} \rightsquigarrow N(0, 1)$ (Note that $I_n(\theta) = nI(\theta)$)
4. The MLE is *asymptotically optimal* or *efficient*: among all well-behaved estimators, the MLE has the smallest variance, at least for large samples.

2.6.1 Proof of Asymptotic Normality

We prove the asymptotic normality for the MLE. Let $\ell(\theta)$ denote the log-likelihood. For MLE $\widehat{\theta}$ we have $\ell'(\widehat{\theta}) = 0$. Use the first order Taylor polynomial to approximate $\ell'(\widehat{\theta})$ at the point $\widehat{\theta}$:

$$0 = \ell'(\widehat{\theta}) \approx \ell'(\theta) + (\widehat{\theta} - \theta)\ell''(\theta)$$

so we have

$$\sqrt{n}(\widehat{\theta} - \theta) \approx \frac{\ell'(\theta)/\sqrt{n}}{-\ell''(\theta)/n}.$$

$\ell'(\theta)$ is n times the score function $S_\theta(X)$ for the distribution $X \sim f_\theta$. Recall $\mathbb{E}[S_\theta(X)] = 0$ and $\mathbb{V}[S_\theta(X)] = I(\theta)$. Hence the numerator

$$n^{-1/2} \sum_i S_\theta(X_i) = \sqrt{n} \cdot \overline{S_\theta(X_i)} = \sqrt{n}(\overline{S_\theta(X_i)} - 0) \rightsquigarrow W \sim N(0, I(\theta))$$

by the central limit theorem. For the denominator, let $A_i = -\partial^2 \log f_\theta(X_i)/\partial\theta^2$. Then the denominator is $\overline{A}$ and $\mathbb{E}(A_i) = I(\theta)$ for each i. Thus by the law of large numbers the denominator converges to $I(\theta)$, the theoretical mean. We conclude by Slutsky's theorem that

$$\sqrt{n}(\widehat{\theta} - \theta) \rightsquigarrow \frac{W}{I(\theta)} \sim N\left(0, \frac{1}{I(\theta)}\right).$$

Recall $\widehat{se} = \sqrt{1/I_n(\widehat{\theta}_n)} = \sqrt{1/nI(\widehat{\theta}_n)}$, so we can rewrite the above as

$$\begin{aligned}\frac{\widehat{\theta}_n - \theta}{\widehat{se}} &= \sqrt{n} I^{1/2}(\widehat{\theta}_n)(\widehat{\theta}_n - \theta) \\ &= \left[\sqrt{n} I^{1/2}(\theta)(\widehat{\theta}_n - \theta)\right] \sqrt{\frac{I(\widehat{\theta}_n)}{I(\theta)}}\end{aligned}$$

The first term tends to $N(0,1)$. Assuming that I is continuous, we have $I(\widehat{\theta}_n) \xrightarrow{p} I(\theta)$, so that the second term tends to 1.

2.6.2 Delta Method

We can get the distribution of functions of parameters $g(\theta)$ in a similar fashion. We first show the univariate case.

Theorem 2.28. If the sequence of variables $\{X_n\}$ satisfies

$$\sqrt{n}[X_n - \theta] \rightsquigarrow N(0, \sigma^2)$$

then

$$\sqrt{n}[g(X_n) - g(\theta)] \rightsquigarrow N(0, \sigma^2 \cdot [g'(\theta)]^2)$$

Proof. First order Taylor approximation at the point X_n:

$$g(X_n) = g(\theta) + g'(\tilde{\theta})(X_n - \theta)$$

where $\tilde{\theta}$ lies between X_n and θ. Note that since $X_n \xrightarrow{P} \theta$ and $X_n < \tilde{\theta} < \theta$, it must be that $\tilde{\theta} \xrightarrow{P} \theta$ and since $g'(\theta)$ is continuous, applying the continuous mapping theorem yields $g'(\tilde{\theta}) \xrightarrow{P} g'(\theta)$.

Rearranging the terms and multiplying by $\sqrt{n}$ gives

$$\sqrt{n}[g(X_n) - g(\theta)] = g'(\tilde{\theta})\sqrt{n}[X_n - \theta]$$

and the delta method follows. □

In the multivariate case, suppose $\sqrt{n}(B-\beta) \rightsquigarrow N(0, \Sigma)$, and we want to derive the distribution of $h(\beta)$ for some function h. We again use the Taylor approximation at point B:

$$h(B) \approx h(\beta) + \nabla h(\beta)^T \cdot (B - \beta)$$

$$\Downarrow$$

$$\sqrt{n}\,(h(B) - h(\beta)) \approx \sqrt{n} \cdot \nabla h(\beta)^T \cdot (B - \beta)$$

The variance of the right hand side is

$$\begin{aligned}
\mathbb{V}(\sqrt{n} \cdot \nabla h(\beta)^T \cdot (B - \beta)) &= n \cdot \mathbb{V}(\nabla h(\beta)^T \cdot B) \\
&= n \cdot [\nabla h(\beta)^T \cdot Cov(B) \cdot \nabla h(\beta)] \\
&= n \cdot [\nabla h(\beta)^T \cdot \Sigma/n \cdot \nabla h(\beta)] \\
&= \nabla h(\beta)^T \cdot \Sigma \cdot \nabla h(\beta)
\end{aligned}$$

where we have used the fact that $Cov(B) \approx \Sigma/n$ from $\sqrt{n}(B - \beta) \rightsquigarrow N(0, \Sigma)$. Thus the conclusion is that

$$\sqrt{n}(h(B) - h(\beta)) \rightsquigarrow N(0, \nabla h(\beta)^T \cdot \Sigma \cdot \nabla h(\beta)).$$

2.7 Methods for Evaluating Estimators

2.7.1 Mean Squared Error

Definition 2.29. Given an estimator T for a parameter $\theta \in \Theta$, the MSE is defined as

$$MSE_T(\theta) = \mathbb{E}_\theta\left[(T - \theta)^2\right]$$

Definition 2.30. T is an *unbiased* estimator for θ ($g(\theta)$) if

$$\mathbb{E}_\theta[T] = \theta \quad \forall \theta \in \Theta \qquad \Big(\mathbb{E}_\theta[T] = g(\theta) \quad \forall \theta \in \Theta\Big).$$

The quantity

$$B_T(\theta) = \mathbb{E}_\theta(T) - \theta \qquad \Big(B_T(\theta) = \mathbb{E}_\theta(T) - g(\theta)\Big)$$

is termed *bias*.

Proposition 2.31. $MSE_T(\theta) = \mathbb{E}_\theta\left[(T - \theta)^2\right] = \mathbb{V}_\theta(T) + B_T^2(\theta)$.

Proof.

$$\begin{aligned}
\mathbb{E}_\theta\left[(T-\theta)^2\right] &= \mathbb{E}_\theta\left[(T-\mathbb{E}_\theta(T)+\mathbb{E}_\theta(T)-\theta)^2\right] \\
&= \underbrace{\mathbb{E}_\theta\left[(T-\mathbb{E}_\theta(T))^2\right]}_{=\mathbb{V}_\theta(T)} + \mathbb{E}_\theta\left[\underbrace{(\mathbb{E}_\theta(T)-\theta)}_{=B_T(\theta)}^2\right] \\
&\quad + 2\mathbb{E}_\theta[(T-\mathbb{E}_\theta(T))(\mathbb{E}_\theta(T)-\theta)] \\
&= \mathbb{V}_\theta(T) + B_T(\theta)^2 + 2(\mathbb{E}_\theta(T)-\theta)\underbrace{\mathbb{E}_\theta[T-\mathbb{E}_\theta(T)]}_{=0} \\
&= \mathbb{V}_\theta(T) + B_T(\theta)^2
\end{aligned}$$

□

If T is unbiased, then $MSE_T(\theta) = \mathbb{V}_T(\theta)$ for any $\theta \in \Theta$. This implies that finding an estimator that minimizes the MSE within the class of unbiased estimators, is equivalent to finding T for which $\mathbb{V}_\theta(T)$ is minimal.

2.7.2 Uniformly minimum variance unbiased estimators (UMVUE)

Definition 2.32. T^* is the *uniformly minimum variance unbiased estimator* (UMVUE) for $g(\theta)$ if

- $\mathbb{E}_\theta(T^*) = g(\theta) \quad \forall\theta \in \Theta$
- $\mathbb{V}_\theta(T^*) \leq \mathbb{V}_\theta(T) \quad \forall\theta \in \Theta$ and T unbiased.

Example 2.33. Let $\underline{X} = (X_1, \ldots, X_n)$ be i.i.d with $X_i \sim f$ s.t. $\mathbb{E}(X_i) = \mu$ and $\mathbb{V}(X_i) = \sigma^2$. Clearly $\bar{X}$ is an unbaised estimator for μ. We want an unbiased estimator for σ^2. For the sample variance

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

we have

$$\begin{aligned}
\mathbb{E}(S^2) &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2\right] - 2\frac{1}{n}\mathbb{E}\sum_{i=1}^{n} X_i\bar{X} + \mathbb{E}[\bar{X}^2] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - 2\mathbb{E}\frac{X_1+\cdots+X_n}{n}\cdot\bar{X} + \mathbb{E}[\bar{X}^2] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - 2\mathbb{E}[\bar{X}^2] + \mathbb{E}[\bar{X}^2] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \mathbb{E}[\bar{X}^2] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \left(\mathbb{V}[\bar{X}] + \mathbb{E}[\bar{X}]^2\right) \\
&= \frac{1}{n}\cdot n\cdot(\sigma^2+\mu^2) - \frac{\sigma^2}{n} - \mu^2 \\
&= \left(1-\frac{1}{n}\right)\sigma^2 = \frac{n-1}{n}\sigma^2.
\end{aligned}$$

Thus S^2 is a biased estimator. The estimator

$$S_c^2 = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

would be unbiased since $\mathbb{E}[S_c^2] = \frac{n}{n-1}\mathbb{E}[S^2] = \sigma^2$. To compare the MSE of the two estimators, we need to calculate the second moments according to Proposition 2.31. At this point we add the assumption that $X_i \sim N(\mu,\sigma^2)$. Then the two would follow gamma distributions. It can be shown that $S_c^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$, and so $\mathbb{V}(S_c^2) = \frac{\sigma^4}{(n-1)^2}2(n-1) = \frac{2\sigma^4}{(n-1)}$. The MSE for S_c^2 is then $\frac{2\sigma^4}{(n-1)}$. Similarly, $S^2 \sim \frac{\sigma^2}{n}\chi^2_{n-1}$ and $\mathbb{V}(S^2) = \frac{\sigma^4}{n^2}2(n-1)$. Thus the MSE for S^2 is $\mathbb{V}_{\sigma^2}(S^2) + B^2_{S^2}(\sigma^2) = \frac{2\sigma^4(n-1)}{n^2} + \left(-\frac{\sigma^2}{n}\right)^2 = \sigma^4\frac{2n-1}{n^2}$. Since

$$\frac{2}{n-1} > \frac{2n-1}{n^2}$$

the MSE for S^2 is smaller than the MSE for S_c^2, even though S^2 is biased.

Theorem 2.34 (Rao-Blackwell Theorem)**.** Let U be an unbiased estimator for $g(\theta)$ and T a sufficient statistic for θ. Then

$$T^* = \mathbb{E}[U|T]$$

is an unbiased estimator for $g(\theta)$ and

$$\mathbb{V}_\theta(T^*) \le \mathbb{V}_\theta(U) \quad \forall\theta\in\Theta$$

with the equality holding if and only if $T^* = U$ with probability 1.

Proof. We need to show four things:

1. $T^* = \mathbb{E}[U|T]$ is a statistic, i.e. it does not depend on θ.
2. T^* is unbiased.
3. $\mathbb{V}_\theta(T^*) \leq \mathbb{V}_\theta(U)$.
4. $\mathbb{V}_\theta(T^*) = \mathbb{V}_\theta(U)$ if and only if $\mathbb{P}\{T^* = U\} = 1$.

As for Item 1, since T is sufficient, $\underline{X}|T$ does not depend on θ. Moreover U is a function of $\underline{X}$, so T^* does not depend on θ.

As for Item 2, by law of iterated expectation $\mathbb{E}[T^*] = \mathbb{E}[\mathbb{E}[U|T]] = \mathbb{E}[U] = g(\theta)$.

As for Item 3, we compute $\mathbb{V}_\theta(U)$:

$$\begin{aligned}
\mathbb{V}_\theta(U) &= \mathbb{E}_\theta\left[(U - g(\theta))^2\right] \\
&= \mathbb{E}_\theta[(U - T^* + T^* - g(\theta))^2] \\
&= \mathbb{E}_\theta[(U - T^*)^2] + \mathbb{E}_\theta[(T^* - g(\theta))^2] \\
&\quad + 2\mathbb{E}_\theta\left[(U - T^*)(T^* - g(\theta))\right] \\
&= \mathbb{V}_\theta(T^*) + \underbrace{\mathbb{E}_\theta[(U - T^*)^2]}_{\geq 0} + 0
\end{aligned}$$

from which we conclude $\mathbb{V}_\theta(U) \geq \mathbb{V}_\theta(T^*)$.

As for Item 4, we showed previously that

$$\mathbb{V}_\theta(U) = \mathbb{V}_\theta(T^*) + \mathbb{E}_\theta\left[(U - T^*)^2\right] \quad \forall \theta \in \Theta$$

and so $\mathbb{E}_\theta\left[(U - T^*)^2\right] = 0$ if and only if $\mathbb{P}\{U = T^*\} = 1$. □

Theorem 2.35. T^* is UMVUE for $g(\theta)$ if and only if $Cov_\theta(T^*, U) = 0 \forall \theta$ where U is any unbiased estimator of 0 (i.e.$\mathbb{E}_\theta[U] = 0 \forall \theta$).

Theorem 2.36 (Lehmann-Scheffé Theorem)**.** Let T be a sufficient and complete statistic for θ and $T^* = h(T)$ be any unbiased estimator for $g(\theta)$. Then T^* is UMVUE for $g(\theta)$ and essentially unique.

Proof. 1. First we show that T^* is unique. Let $\tilde{T}$ be another estimator which is

   (a) a function of T: $\tilde{T} = \tilde{h}(T)$
   (b) unbiased: $\mathbb{E}_\theta(\tilde{T}) = g(\theta)$.

   Then we have $\mathbb{E}_\theta(\tilde{T} - T^*) = 0$. Moreover, since both T^* and $\tilde{T}$ is a function of T, their difference is also a function of T. By completeness of T we have $\mathbb{P}\{\tilde{T} = T^*\} = 1$.

2. Let U be any unbiased estimator of $g(\theta)$. Then by the Rao-Balckwell Theorem $\mathbb{E}[U|T]$ is also unbiased and $\mathbb{V}_\theta(\mathbb{E}[U|T]) \leq \mathbb{V}_\theta(U) \forall \theta$. Since the function of T that is unbiased is unique, we have $T^* = \mathbb{E}[U|T]$ a.s. and $\mathbb{V}_\theta(T^*) \leq \mathbb{V}_\theta(U)$. Since U is arbitrary $\mathbb{V}_\theta(T^*)$ is minimal and T^* is UMVUE.

□

We next show some applications of the Rao-Blackwell theorem and the Lehmann-Scheffé Theorem.

Example 2.37. Let $\underline{X} = (X_1, \ldots, X_n)$ be i.i.d with $X_i \sim N(\mu, \sigma^2)$. We distinguish 3 cases:

1. σ^2 known
   $T = \sum_{i=1}^n X_i$ is sufficient and complete, and $\bar{X} = h(T) = (1/n)T = (1/n)\sum_{i=1}^n X_i$ is unbiased for μ. Hence by the Lehmann-Scheffé Theorem $\bar{X}$ is UMVUE for μ.

2. μ known
   $T = \sum_{i=1}^n (X_i - \mu)^2$ is sufficient and complete. We have

   $$\mathbb{E}[T] = \sum_{i=1}^n \mathbb{E}[X_i - \mu]^2 = n\sigma^2$$

   and thus $T^* = (1/n)T$ is a function of T that is unbiased, and so is UMVUE by the Lehmann-Scheffé Theorem.

3. Both μ and σ^2 unknown
   We know that $T = (\sum_{i=1}^n X_i, \sum_{i=1}^n (X_i - \mu)^2)$ is sufficient and complete for (μ, σ^2). The same is true for the one-to-one transformation $T^* = (\bar{X}, S_c^2)$. It is unbiased, and so by the Lehmann-Scheffé Theorem it is UMVUE for (μ, σ^2).

Example 2.38. Let $\underline{X} = (X_1, \ldots, X_n)$ be i.i.d with $X_i \sim Be(\theta)$.

1. *UMVUE for* θ: We know that $T = \sum X_i$ is sufficient and complete. Then

   $$T^* = h(T) = \frac{1}{n}T = \bar{X}$$

   is a function of T and is unbiased, and so T^* is UMVUE for θ.

2. *UMVUE for* $\theta(1-\theta)$: Suppose we wan want to know the UMVUE for $g(\theta) = \theta(1-\theta) = \mathbb{V}(X_i)$. We might first want to try the MLE estimate $T = \widehat{\theta(1-\theta)} = \hat{\theta}(1-\hat{\theta}) = \bar{X}(1-\bar{X})$.

Let's check whether the MLE is unbiased:

$$\begin{aligned}\mathbb{E}_\theta[\bar{X}(1-\bar{X})] &= \mathbb{E}_\theta(\bar{X}-\bar{X}^2)\\ &= \mathbb{E}_\theta(\bar{X}) - \mathbb{E}_\theta(\bar{X}^2)\\ &= \theta - [\mathbb{V}_\theta(\bar{X}) + \mathbb{E}(\bar{X})^2]\\ &= \theta - \frac{\theta(1-\theta)}{n} - \theta^2\\ &= \frac{n-1}{n}\theta(1-\theta).\end{aligned}$$

Hence the MLE is biased but it is easy to correct the bias: for $T^* = h(T) = \frac{n}{n-1}T = \frac{n}{n-1}\bar{X}(1-\bar{X})$ we have $\mathbb{E}_\theta(T^*) = \theta(1-\theta)$. T^* is a function of the complete and sufficient statistic T that is unbiased, so we conclude that T^* is UMVUE for $\theta(1-\theta)$.

Example 2.39. Let $\underline{X} = (X_1, \ldots, X_n)$ be i.i.d. with $X_i \sim Po(\lambda)$. Recall that $T = \sum_{i=1}^n X_i$ is a sufficient and complete statistic for λ.

1. The UMVUE for λ is clearly $T^* = \bar{X}$.
2. Find the UMVUE for
$$g_k(\lambda) = \mathbb{P}\{X_i = k\} = \frac{e^{-\lambda}\lambda^k}{k!}.$$
To use the Rao-Blackwell theorem we need to first come up with an unbiased estimator U for $g_k(\lambda)$ and then conditioning on T. $\mathbb{1}_k(X_1)$ would be such an estimator:
$$\mathbb{E}[\mathbb{1}_k(X_1)] = \mathbb{P}\{X_1 = k\} = \frac{e^{-\lambda}\lambda^k}{k!}.$$
Then $T^* = \mathbb{E}[\mathbb{1}_k(X_1)|T]$ would be the UMVUE for $g_k(\lambda)$. Let's compute it:
$$\begin{aligned}T^* &= \mathbb{E}[\mathbb{1}_k(X_1)|T]\\ &= \mathbb{P}\{X_1 = k|T\}\\ &= \frac{\mathbb{P}\{X_1 = k, T = t\}}{\mathbb{P}\{T = t\}}\\ &= \begin{cases}0 & t < k\\ (*) & t \geq k\end{cases}\end{aligned}$$

where

$$\begin{aligned}(*) &= \frac{\mathbb{P}\{X_1 = k, \sum_{i=2}^n X_i = t-k\}}{\mathbb{P}\{\sum_{i=1}^n X_i = t\}} \\ &= \frac{\mathbb{P}\{X_1 = k\}\mathbb{P}\{\sum_{i=2}^n X_i = t-k\}}{\mathbb{P}\{\sum_{i=1}^n X_i = t\}} \\ &= \frac{e^{-\lambda}\lambda^k}{k!} \cdot \frac{e^{-(n-1)\lambda}((n-1)\lambda)^{t-k}}{(t-k)!} \Bigg/ \frac{e^{-n\lambda}(n\lambda)^t}{t!} \\ &= \frac{t!}{k!(t-k)!} \cdot \frac{(n-1)^{t-k}}{n^t} \\ &= \binom{t}{k}\left(\frac{1}{n}\right)^k \left(\frac{n-1}{n}\right)^{t-k}.\end{aligned}$$

Hence the UMVUE for $g_k(\lambda)$ is

$$T^* = \begin{cases} \binom{T}{k}\left(\frac{1}{n}\right)^k \left(\frac{n-1}{n}\right)^{t-k} & k = 0, \dots, T \\ 0 & \text{otherwise.} \end{cases}$$

In particular, if $k = 0$, then the UMVUE for $g_0(\lambda) = e^{-\lambda} = \mathbb{P}\{X_i = 0\}$ would be

$$T^* = \left(1 - \frac{1}{n}\right)^T.$$

Example 2.40. Let $\underline{X} = (X_1, \dots, X_n)$ be i.i.d. with $X_i \sim \mathcal{U}(0, \theta)$. We know that $T = X_{(n)}$ is sufficient and complete for θ. The CDF of T is given by

$$F_\theta^T(t) = [F^X(t)]^n = \left(\frac{t}{\theta}\right)^n,$$

so the PDF of T is

$$f_\theta^T(t) = n\left(\frac{t}{\theta}\right)^{n-1} \frac{1}{\theta} \mathbb{1}_{(0,\theta)}(t).$$

The expectation of T is thus

$$\begin{aligned}\mathbb{E}(T) &= \int_0^\theta t f_\theta^T(t) dt = \frac{n}{\theta^n} \int_0^\theta t^n dt \\ &= \frac{n}{\theta^n} \frac{t^{n+1}}{n+1}\Bigg|_0^\theta = \frac{n}{\theta^n} \frac{\theta^{n+1}}{n+1} = \frac{n}{n+1}\theta.\end{aligned}$$

Hence $T = X_{(n)}$ is biased but for $T^* = ((n+1)/n)T$ we have

$$\mathbb{E}_\theta(T^*) = \theta$$

and so T^* is UMVUE.

2.8 Cramér-Rao Lower Bound

Theorem 2.41 (Cramér-Rao Lower Bound)**.** $\underline{\mathrm{X}} \sim f_\theta$ where $\mathcal{P} = \{f_\theta : \theta \in \Theta\}$ be a C.R. regular model. Let $T = T(\underline{\mathrm{x}})$ be an estimator for θ. Assume $\mathbb{V}_\theta(T) < \infty$ and

$$\frac{\partial}{\partial\theta}\int T(\underline{\mathrm{x}})\, f_\theta(\underline{\mathrm{x}})\, d\underline{\mathrm{x}} = \int T(\underline{\mathrm{x}})\frac{\partial}{\partial\theta} f_\theta(\underline{\mathrm{x}})\, d\underline{\mathrm{x}} \quad \forall\theta \in \Theta.$$

Then one has

$$\mathbb{V}_\theta(T) \geq \left(\frac{\partial}{\partial\theta}\mathbb{E}_\theta[T]\right)^2 \Big/ I^{\underline{\mathrm{X}}}(\theta).$$

Proof. We compute $\left(\frac{\partial}{\partial\theta}\mathbb{E}_\theta[T]\right)^2$. By definition we have

$$\mathbb{E}_\theta[T] = \int T(\underline{\mathrm{x}})\, f_\theta(\underline{\mathrm{x}})\, d\underline{\mathrm{x}}.$$

Take derivatives on both sides w.r.t. θ:

$$\begin{aligned}
\frac{\partial}{\partial\theta}\mathbb{E}_\theta[T] &= \int T(\underline{\mathrm{x}})\frac{\partial}{\partial\theta} f_\theta(\underline{\mathrm{x}})\, d\underline{\mathrm{x}} \quad \text{(by assumption)} \\
&= \int T(\underline{\mathrm{x}})\, \underbrace{\frac{\frac{\partial}{\partial\theta} f_\theta(\underline{\mathrm{x}})}{f_\theta(\underline{\mathrm{x}})}}_{=S_\theta(\underline{\mathrm{x}})}\, f_\theta(\underline{\mathrm{x}})\, d\underline{\mathrm{x}} \\
&= \int T(\underline{\mathrm{x}}) S_\theta(\underline{\mathrm{x}})\, f_\theta(\underline{\mathrm{x}})\, d\underline{\mathrm{x}} \\
&= \mathbb{E}_\theta\left[T(\underline{\mathrm{X}}) S_\theta(\underline{\mathrm{X}})\right]. \\
&= Cov\left[T(\underline{\mathrm{X}}) S_\theta(\underline{\mathrm{X}})\right].
\end{aligned}$$

Recall the Cauchy-Schwarz inequality

$$Cov^2(X, Y) \leq \mathbb{V}(X)\mathbb{V}(Y),$$

so that

$$\left(\frac{\partial}{\partial\theta}\mathbb{E}_\theta(T)\right)^2 = Cov^2(T, S_\theta) \leq \mathbb{V}_\theta(T)\mathbb{V}_\theta(S_\theta).$$

Rearranging we get the desired result. □

Corollary 2.42. If T is also an unbiased estimator of $g(\theta)$, then the inequality becomes

$$\mathbb{V}_\theta(T) \geq \left(\frac{\partial}{\partial\theta} g(\theta)\right)^2 \Big/ I^{\underline{\mathrm{X}}}(\theta).$$

In the special case $g(\theta) = \theta$, the bound is

$$\mathbb{V}_\theta(T) \geq 1 \Big/ I^{\underline{\mathrm{X}}}(\theta).$$

Example 2.43. The theorem does not hold for non C.R. regular models. For example, consider $\underline{X} = (X_1, \dots, X_n)$ with $X_i \overset{i.i.d.}{\sim} \mathcal{U}(0, \theta)$. We know from Example 2.40 that $T^* = \frac{n+1}{n} X_{(n)}$ is UMVUE for θ. From the PDF of $X_{(n)}$

$$f_\theta^{X_{(n)}} = \frac{n y^{n-1}}{\theta^n} \mathbb{1}_{(0,\theta)}(y)$$

one can calculate the variance of T^*, which turns out to be $\theta^2 \frac{1}{(n+2)n}$. The Fisher information for the sample $\underline{X}$ is $I^{\underline{X}}(\theta) = n/\theta^2$, so that $\mathbb{V}_\theta(T^*) < 1/I^{\underline{X}}(\theta)$ in this case.

A natural question is when the inequality becomes equality. The following result provides a characterization in terms of the score function.

Corollary 2.44. The C.R. inequality becomes an equality if and only if the score function is linear in T:

$$S_\theta(\underline{x}) = r(\theta)[T(\underline{x}) - g(\theta)]$$

for some function $r(\theta) \neq 0$ and $g(\theta) = \mathbb{E}_\theta[T(\underline{X})]$.

Proof. This stems from the fact that the Cauchy-Schwarz inequality

$$Cov^2(X, Y) \leq \mathbb{V}(X)\mathbb{V}(Y)$$

becomes an equality if and only if there is a linear relationship between X and Y. □

If the variance an estimator attains the C.R. lower bound l_{CR}, then we say that the estimator is *efficient*.

Definition 2.45. Given two unbiased estimator for $g(\theta)$, say T_1 and T_2 with $\mathbb{V}_\theta(T_1) < \infty$ and $\mathbb{V}_\theta(T_2) < \infty$, T_1 is *more efficient* that T_2 if

$$\mathbb{V}_\theta(T_1) < \mathbb{V}_\theta(T_2) \quad \forall \theta \in \Theta.$$

An estimator is termed *efficient* if

$$\mathbb{V}_\theta(T) = l_{CR} \quad \forall \theta \in \Theta.$$

It is said to be *asymptotically efficient* if

$$\lim_{n\to\infty} \frac{l_{CR}}{\mathbb{V}_\theta(T)} = 1 \quad \forall \theta \in \Theta.$$

If an estimator T attains the C.R. lower bound (i.e. is efficient), then it is UMVUE. The converse is not true: the variance of an UMVUE estimator may not be able to reach the C.R. lower bound. One can prove that an efficient estimator exists essentially only if the model belongs to the exponential family and the goal is to estimate $g(\theta) = -\frac{c'(\theta)}{b'(\theta)}$ (or a linear transformation), the expected value of $T = \sum R(X_i)$.

Example 2.46. Let $\underline{X} = (X_1, \ldots, X_n)$ be an i.i.d. sample with $X_i \sim f_\theta$ belonging to the exponential family:

$$f_\theta(x_i) = g(x_i) \exp\{b(\theta) R(x_i) + c(\theta)\}.$$

Let's compute the score function of $\underline{X}$:

$$\begin{aligned}
\frac{\partial}{\partial \theta} \log f_\theta(\underline{x}) &= \frac{\partial}{\partial \theta} \left[b(\theta) \sum_{i=1}^n R(x_i) + n c(\theta) + \sum_{i=1}^n \log g(x_i) \right] \\
&= b'(\theta) \sum_{i=1}^n R(x_i) + n c'(\theta) \\
&= n b'(\theta) \left[\frac{1}{n} \sum_{i=1}^n R(x_i) - \left(-\frac{c'(\theta)}{b'(\theta)} \right) \right].
\end{aligned}$$

Now recall $\mathbb{E}[R(X_i)] = -c'(\theta)/b'(\theta)$, so that

$$\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n R(x_i) \right] = -\frac{c'(\theta)}{b'(\theta)}.$$

We apply Corollary 2.44 to conclude that $T = \frac{1}{n} \sum_{i=1}^n R(x_i)$ is the efficient estimator for $-c'(\theta)/b'(\theta)$.

Example 2.47. By computing the score function of the samples, we can show that

- For $X_i \sim Po(\lambda)$, $T = \bar{X}$ is an efficient estimator for λ.
- For $X_i \sim B(\theta)$, $T = \bar{X}$ is an efficient estimator for θ.
- For $X_i \sim N(\mu, \sigma^2)$, S_c is the UMVUE for σ^2, but it is not efficient: $\text{eff}(S_c^2) = \frac{n-1}{n} < 1$. However, it is asymptotically efficient.

2.9 Hypothesis Testing

- In parametric tests, the hypotheses split the parameter space Θ into two non-overlapping sets:

$$\Theta = \Theta_0 \cup \Theta_1 \quad \Theta_0 \cap \Theta_1 = \varnothing.$$

- The test splits the set $\mathcal{X}$ of all observations into two subsets $\mathcal{R}$ (*critical region*) and $\mathcal{A}$ (*acceptance region*), such that $\mathcal{X} = \mathcal{R} \cup \mathcal{A}$ and $\mathcal{R} \cap \mathcal{A} = \varnothing$. The rule is

$$\begin{aligned}
\underline{X} \in \mathcal{R} &\Rightarrow \text{reject } H_0 \\
\underline{X} \in \mathcal{A} &\Rightarrow \text{accept } H_0
\end{aligned}$$

Typically, the rejection region is defined as

$$\mathcal{R} = \{\underline{x} \in \mathcal{X} : T(\underline{x}) > c\}$$

where T is a *test statistic* and c is a *critical value*.

- Type I error: $\underline{X} \in \mathcal{R}$ when H_0 is true.
- Type II error: $\underline{X} \in \mathcal{A}$ when H_0 is false.

Definition 2.48. The *power function* of a test with critical region $\mathcal{R}$ is

$$Q(\theta) = \mathbb{P}_\theta\{\underline{X} \in \mathcal{R}\}.$$

When $\theta \in \Theta_0$, the value of the power function $Q(\theta)$ is the probability of making type I error. When $\theta \in \Theta_1$, namely H_0 is false, the value of the power function is equal to the probability that we reject H_0, which is also equal to one minus the probability of making type II error. We want the probability of making type II error to be small, so we want $Q(\theta)$ to be large when $\theta \in \Theta_1$.

The *size* of the test is defined to be $\alpha = \sup_{\theta\in\Theta_0} Q(\theta)$, which is the maximum probability of making type I error. For a given size α test, we want to find the test with highest $Q(\theta)$ for all $\theta \in \Theta_1$. Such a test, if it exits, is called *most powerful*. We omit the investigation about the existence of most powerful tests. Instead, we give several examples of tests.

2.9.1 Wald Test

Let θ be a scalar, $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ be an estimate of θ, and $\widehat{se}$ be the estimated standard error of $\widehat{\theta}$.

Definition 2.49 (Wald Test)**.** Consider the testing

$$H_0 : \theta = \theta_0 \quad \text{versus } H_1 : \theta \neq \theta_0.$$

Assume that $\widehat{\theta}$ is asymptotically normal:

$$W = \frac{(\widehat{\theta} - \theta_0)}{\widehat{se}} \rightsquigarrow N(0, 1).$$

The size α Wald test is: reject H_0 when $|W| > z_{\alpha/2}$.

2.9.2 Pearson's χ^2 test

Pearson's χ^2 test is used for multinomial data. Recall that if $X = (X_1, \ldots, X_k)$ has a multinomial (n, p) distribution, then the MLE of p is $\widehat{p} = (\widehat{p}_1, \ldots, \widehat{p}_k) = (X_1/n, \ldots, X_k/n)$. Let $p^0 = (p_1^0, \ldots, p_k^0)$ be some fixed vector and suppose we want to test

$$H_0 : p = p^0 \quad \text{v.s.} \quad H_1 : p \neq p_0.$$

Definition 2.50. Pearson's χ^2 statistic is

$$T(X_1, \ldots, X_k) = \sum_{j=1}^{k} \frac{(X_j - n \cdot p_j^0)^2}{n \cdot p_j^0} = \sum_{j=1}^{k} \frac{(X_j - E_j)^2}{E_j}$$

where $E_j = \mathbb{E}(X_j) = n \cdot p_j^0$ is the expected value of X_j under H_0.

Under H_0, $T \rightsquigarrow \chi^2_{k-1}$. The test: reject H_0 if $T > \chi^2_{k-1}$. The p-value is $\mathbb{P}(\chi^2_{k-1} > t)$ where t is the observed value of the test statistic.

2.9.3 Permutation Test

The permutation test is a non-parametric method for testing whether two distributions are the same. Suppose $X_1, \ldots, X_m \sim F_X$ and $Y_1, \ldots, Y_n \sim F_Y$ are two independent samples. The test is

$$H_0 : F_X = F_Y \quad \text{versus} \quad H_1 : F_X \neq F_Y.$$

Let $T(x_1, \ldots, x_m, y_1, \ldots, y_n)$ be some test statistic, for example $T(X_1, \ldots, X_m, Y_1, \ldots, Y_n) = |\bar{X}_m - \bar{Y}_n|$. Let $N = m+n$ and consider forming all $N!$ permutations of the data $\{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$. For each permutation, compute the test statistic T, and denote these values by $\{T_1, \ldots, T_{N!}\}$. Under the null hypothesis, each of these values is equally likely. The distribution $\mathbb{P}_0$ that puts mass $1/N!$ on each T_j is called the permutation distribution of T. Let t_{obs} be the observed value of the test statistic. Assuming we reject when T is large, the p-value is

$$\mathbb{P}_0(T > t_{obs}) = \frac{1}{N!} \sum_{j=1}^{N!} I(T_j > t_{obs}).$$

Usually it is not practical to evaluate all $N!$ permutations. We can approximate the p-value by sampling randomly from the set of permutations.

2.9.4 The Likelihood Ratio Test

Definition 2.51. Consider the testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \notin \Theta_0$. The likelihood ratio statistic is

$$\lambda = 2 \log \left(\frac{\sup_{\theta \in \Theta} \mathcal{L}(\theta)}{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta)} \right) = 2 \log \left(\frac{\mathcal{L}(\widehat{\theta})}{\mathcal{L}(\widehat{\theta}_0)} \right).$$

If the null hypothesis is false, then the numerator ($\sup_{\theta \in \Theta} \mathcal{L}(\theta)$) will be larger than the denominator ($\sup_{\theta \in \Theta_0} \mathcal{L}(\theta)$), so that the statistic will be large. Suppose

$$\Theta_0 = \left\{ \theta : (\theta_1, \ldots, \theta_q, \theta_{q+1}, \ldots, \theta_r) = (\theta_1, \ldots, \theta_q, \theta_{q+1}^0, \ldots, \theta_r^0) \right\},$$

i.e. the set of points where some coordinates of θ are fixed. Then under H_0

$$\lambda(X_1, \ldots, X_n) \rightsquigarrow \chi^2_{r-q}.$$

The p-value for the test is $\mathbb{P}(\chi^2_{r-q} > \lambda)$.

3 The Bootstrap

Bootstrap is a method for estimating standard errors and computing confidence intervals. We let $T_n = g(X_1, \ldots, X_n)$ be a statistic. We want to know $V(T_n)$, the variance of T_n. If $T(F) = F^{-1}(1/2)$ so that $T_n = T(\hat{F}_n)$ is the median of the data, then what is the variance of T_n? As another example, if $T(F) = \int (x-\mu)^3 dF(x)/\sigma^3$ is the skewness of the distribution, then what is the variance of $T_n = T(\hat{F}_n)$? There may not exist a formula, or they can be very complicated.

Given our data points $\{x_1, \ldots, x_n\}$, we can only calculate one value of T_n. But if we want to know something about the distribution of T_n, like the variance, then we'd better have multiple values of T_n at our hands. How can we do that? Well, just re-calculate T_n using our data points! Draw with replacement $x_1^*, \ldots, x_n^*$ from $\{x_1, \ldots, x_n\}$, calculate $T_n^* = g(x_1^*, \ldots, x_n^*)$ and boo we get another T_n^*. Continue doing this way and we get many T_n^*s. Now we are able to calculate the variance. There are three ways to construct a bootstrap confidence interval:

- The *normal interval*: $T_n \pm z_{\alpha/2}\widehat{se}_{boot}$, where $\widehat{se}_{boot} = \sqrt{v_{boot}}$ is the bootstrap estimate of the standard error. This interval is not accurate unless the distribution of T_n is close to Normal. For $\alpha = 0.05$ this is approximately[1]

```
(th.hat - 2*se, th.hat + 2*se)
```

- The *pivotal interval*: $C_n = (2\hat{\theta}_n - \hat{\theta^*}_{1-\alpha/2}, 2\hat{\theta}_n - \hat{\theta^*}_{\alpha/2})$, where $\hat{\theta^*}_\beta$ is the β quantile of our bootstrap arrays. For $\alpha = 0.05$ this is

```
(2*th.hat-quantile(Tboot, .975), 2*th.hat-quantile(Tboot, .025))
```

- The *percentile interval*: $C_n = (\hat{\theta^*}_{\alpha/2}, \hat{\theta^*}_{1-\alpha/2})$. For $\alpha = 0.05$ this is

```
(quantile(Tboot, .025),quantile(Tboot, .975))
```

Q: Why sometimes we can calculate the variance of T_n analytically?

A: We suppose our dataset $\{X_1, \ldots, X_n\}$ is generated under some distribution F. We never know F. Now, we want to calculate some statistics $T_n = g(X_1, \ldots, X_n)$ and know *its* distribution. To be less ambitious, we may want to know the mean (first moment) or variance (second moment) or confidence intervals about T_n, which all give some partial information about the distribution of T_n. T_n is a function of our data $\{X_1, \ldots, X_n\}$, and as we never know the distribution F of our data, we never know the true distribution of T_n.

[1] `th.hat` denotes the statistic (or estimator) calculated from our data. `Tboot` denotes the array of bootstrap replicates of `th.hat`. `se` is bootstrap standard error.

So, what should I do if I want to calculate the variance of T_n? Say $T_n = (\sum_{i=1}^n X_n)/n := \bar{X}_n$ is the average of the data, and I want to calculate its variance. If the variance of T_n is too large then the value of T_n may be less trustable; if the variance of T_n is very small then I can be more confident that my T_n is close to the true first moment of the distribution F. Having said the motivation, I can't proceed unless I have a distribution of the data at hands. In particular, instead of working with the unknown F, we work with $\hat{F}_n$ which closely approximates F:

$$\hat{F}_n(x) = \frac{\sum_{i=1}^n I(X_i \leq x)}{n},$$

i.e. we give every data point equal weight. If X is a random variable that has distribution $\hat{F}_n$, then $\mathbb{P}(X = X_i) = 1/n$ for each $i = 1, \ldots, n$. It is a discrete distribution that takes values from $\{X_1, \ldots, X_n\}$. We completely know the distribution $\hat{F}_n$. For example:

1. The mean of $\hat{F}_n$ is $\int x dF(x) = \dfrac{X_1 + \cdots + X_n}{n} := \bar{X}_n$
2. The variance of $\hat{F}_n$ is $\int (x - \bar{X}_n)^2 dF(x) = \dfrac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n}$

(The integrals are Riemann-Stieltjes integrals)

Note that $\hat{F}_n$ and F are different. It is important to realize that it is not $\hat{F}_n$ that generates the data, but it is F. $\hat{F}_n$ does not generate our data $\{X_1, \ldots, X_n\}$. Drawing n data points from $\hat{F}_n$ is equivalent to drawing n points with replacement at random from the pool $\{X_1, \ldots, X_n\}$. So I draw once and I get $\{X_1^*, \ldots, X_n^*\}$, which may be slightly different than $\{X_1, \ldots, X_n\}$.

The statistics now is $T_n^* = g(X_1^*, \ldots, X_n^*)$. We approximate everything we want to know about T_n by distribution of T_n^*. In essence, we are replacing F by $\hat{F}_n$ and redo everything: generate the data, calculate the statistics, calculate the distribution of the statistics etc.

$$F \text{ (unknown)} \xrightarrow{generated} \{X_1, \ldots, X_n\} \longrightarrow T_n = g(X_1, \ldots, X_n) \text{ (dist. unknown)}$$

$$\hat{F}_n \text{ (known)} \xrightarrow{generated} \{X_1^*, \ldots, X_n^*\} \longrightarrow T_n^* = g(X_1^*, \ldots, X_n^*) \text{ (dist. known or unknown)}$$

I can calculate the variance of T_n^* now. Recall $T_n^* = (X_1^* + \cdots + X_n^*)/n$, where each X_i^* is generated according to the distribution $\hat{F}_n$. Thus $V(T_n^*) = V(X_1^*) + \cdots + V(X_n^*)/n^2 = V(\hat{F}_n)/n$. What is the variance of $\hat{F}_n$? We just showed that it is $(\sum_{i=1}^n (X_i - \bar{X}_n)^2)/n$. Thus in this case, we have directly worked out the variance, which is

$$\mathbb{V}(T_n^*) = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n^2}.$$

Note that here we are treating each data point X_i^* as a random variable that has distribution $\hat{F}_n$. To recap, the statistics T_n^* is very simple, basically just the sum of those variables whose distribution is completely transparent to us. The distribution of T_n^* is thus not hard to calculate. Note that the variance of T_n^* in this case does not depend on the bootstrap data $\{X_1^*, \ldots, X_n^*\}$. It just depends on the distribution $\hat{F}_n$.

But what about this: suppose T_n is the median of our data $\{X_1, \ldots, X_n\}$. What is the variance of T_n? We generate data once (or simulate) according to $\hat{F}_n$, get $\{X_1^*, \ldots, X_n^*\}$, and then get $T_n^* = \text{median}\{X_1^*, \ldots, X_n^*\}$. What is the variance of T_n^*? I don't know. Even if I know each data point has distribution $\hat{F}_n$, I do not have a clear formula for T_n^*, so it is not obvious how should I calculate the variance as in the previous case. So I need to re-sample again, get another $T_{n,1}^*$; re-sample again, get another $T_{n,2}^*$,...until I simulated T_n^* for B times, then I can calculate the variance by

$$\mathbb{V}_{boot}(T_n^*) = \frac{1}{B}\sum_{b=1}^{B}\left(T_{n,b}^* - \frac{1}{B}\sum_{r=1}^{B} T_{n,r}^*\right)^2.$$

Summary Our data at hands $\{X_1, \ldots, X_n\}$ is generated according to an unknown distribution F. In particular, we do not know the mean and variance of F. A statistics T_n is some function of the data, which is a random variable. Given a dataset we get a concrete real number for T_n. Given another dataset we get another value for T_n, and so on. Since we do not know F, we do not know the distribution of T_n. To solve this problem, we substitute F by $\hat{F}_n$, hoping that the two do not have a large difference. Then we can *try to derive the distribution of T_n under $\hat{F}_n$ analytically*. Recall that a statistics T_n is some function of the data, so the first necessary step is to assume the data is generated according to $\hat{F}_n$. This corresponds to bootstrap just once. Note that we are still in an abstract and theoretical setting yet, and so we still see the bootstrap data as random variables, not as a concrete numerical output of our computer programs. If T_n is simple enough, for example some simple combinations of data points, then we may derive, say, the variance of T_n easily. Since the data is generated according to $\hat{F}_n$, which in turn is related to $\{X_1, \ldots, X_n\}$, the analytical expression for the approximated variance of T_n ultimately depends on $\{X_1, \ldots, X_n\}$.

However, most of the time we do bootstrap exactly because T_n is not some simple linear combinations of the data points, so it is not obvious how to calculate the variance of T_n, even if we already assumed a known distribution of the data points, namely $\hat{F}_n$. We thus re-sample the data again and again and calculate many many values of T_n, to get a sample of T_n, and use the sample variance to approximate the true variance of T_n.

Do not worry about the first case: its purpose is to give us the motivation as to why we need to re-sample the data many times and calculate T_n many times. It says that only in rare cases could we not have to do the actual simulation, but most of the times we do. We can still do the simulation even if we are estimating the variance of the mean statistics. We don't have to, but there is nothing wrong in doing that.

Part III

Statistical Models and Methods

4 EM Algorithm

4.1 Foundations

In order to understand the EM algorithm, we need to be familiar with some basic concepts:

1. Conditional probability:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)} \quad \text{and} \quad \mathbb{P}(B|A) = \frac{\mathbb{P}(AB)}{\mathbb{P}(A)}.$$

   Substitute $\mathbb{P}(AB) = \mathbb{P}(B|A)\mathbb{P}(A)$ into the first equation we get the Bayes' rule

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

2. Marginal distribution: Let $Z \sim B(p)$, i.e. $Z = 1$ with probability p and $Z = 0$ with probability $1 - p$. Suppose Y has density f_1 if $Z = 1$ and it has density f_2 if $Z = 0$. This is like a coin toss. If we get a head ($Z = 1$) we draw a sample from the density f_1, and if we get a tail ($Z = 0$) we draw a sample from the density f_2. We draw many samples this way, until we get our data $\{y_1, \ldots, y_n\}$. What is the distribution of the random variable Y that generated the data? The distribution of Y given Z is

$$f(y|Z) = f_1^{[Z=1]}(y) \cdot f_2^{[Z=0]}(y).$$

   The expression [Condition] equals 1 if the Condition is true, and 0 is the Condition is false. So $f(y|Z = 1) = f_1(y)$, and $f(y|Z = 0) = f_2(y)$. This is also an interpretation of using coin toss to determine which density to draw.

   The distribution of Z is clear: the PDF of Z is: $p(1) = p$ and $p(0) = 1 - p$. Now, the distribution of Y is

$$f(y) = \int f(y|z)p(z)dz = p \cdot f_1(y) + (1 - p) \cdot f_2(y).$$

4.2 Separate Mixture of Normal Distributions

Suppose I give you the following 50 data points:

```
array([  3.82,  98.43, 100.76,   3.66, 100.48,  98.27,   3.65, 100.23,
        98.9 , 101.15,   3.4 , 101.32, 101.15, 101.76,  99.26, 100.2 ,
        99.14, 100.56, 100.24,   3.24, 101.55,  99.79,  99.91,  99.35,
         3.8 , 100.61,   2.44,   2.56,   3.08,   3.57,   3.89,  99.84,
       100.13,   2.75,  98.03,  99.7 ,  98.78,  98.74, 101.55, 100.19,
         1.27, 101.21,  98.84,   1.96, 100.71,   2.67,  99.2 , 100.28,
       101.35,  97.46])
```

and I tell you that the each data point is draw from one of the two normal densities $f(y, \mu_1)$ and $f(y, \mu_2)$ (for simplicity we fix both variances to be 1, and focus on the means). Can you tell me μ_1 and μ_2? Looking at the data, it seems that some are around 3, while others are around 100. So $\mu_1 = 3$ and $\mu_2 = 100$ is not a bad guess. What's more, we can count the total number of values that are relatively small (or close to 3) and divide by 50, to get the probability (i.e.$\mathbb{P}(Z = 1)$) that the first density is chosen.

But we need a general procedure for determine the parameters, instead of looking and guessing every time. For example, if the μ_1 and μ_2 are close, then it is hard to tell which is which:

```
array([6.1 , 5.02, 4.51, 4.4 , 3.46, 5.52, 4.44, 4.84, 5.45, 4.53, 1.78,
       4.77, 4.83, 5.17, 4.41, 2.87, 5.72, 2.61, 7.31, 4.84, 4.1 , 3.67,
       5.7 , 2.34, 3.39, 2.48, 5.26, 4.56, 4.91, 0.51, 5.17, 3.22, 3.97,
       4.49, 3.5 , 5.79, 2.84, 5.1 , 3.85, 4.84, 3.97, 4.41, 5.58, 5.49,
       4.83, 5.36, 5.21, 5.88, 2.91, 5.97])
```

(The above data is generated by $f(y, 3)$ and $f(y, 5)$) In all cases, we want to separate the data into two groups, though this task is easy for the first dataset and harder for this one. We can then calculate the means of the data for each group, to get our estimate for μ_1 and μ_2 respectively (recall that the maximum likelihood estimate for μ is the sample mean) In general, we can take the following iterative strategy: we first randomly choose some values for μ_1^0 and μ_2^0 to be our starting points. We then compare each data point with μ_1^0 and μ_2^0. If, for example, y_1 is close to μ_1^0 than μ_2^0, then it is more likely that y_1 is generated by the normal density $f(y, \mu_1^0)$. So we can say "hey, the guy y_1 belongs to $f(y, \mu_1^0)$". Similarly, We assign each data point y_i from $\{y_2, \ldots, y_n\}$ to $f(y, \mu_1^0)$ or $f(y, \mu_2^0)$ according to the distance of y_i to μ_1 or μ_2. After we do this, we have divided the dataset in to two camps. We can then calculate the average value in each camp, to get our new estimate of μ_1 and μ_2, and repeat the assignment process...

Sounds familiar? This is exactly the k-means algorithm! In each step, for each data point y_i, we assign y_i to $\mu_1^{(t)}$ or $\mu_2^{(t)}$ according to the distance of y_i to each point, where (t) denotes our current estimate of the parameters. In other words, $\mathbb{P}(y_i \sim f_1) = 1$ and $\mathbb{P}(y_i \sim f_2) = 0$ if $f(y_i, \mu_1^{(t)}) > f(y_i, \mu_2^{(t)})$, and vice versa. We designate y_i to either belong to the first density, or the second density, but not both.

The EM algorithm offers a softer approach: instead of 0/1 decision, we assign probabilities that y_i belongs to f_1 or f_2. Namely, both $\mathbb{P}(y_i \sim f_1)$ and $\mathbb{P}(y_i \sim f_2)$ can be positive, and they sum to 1. Say for our dataset $\{y_1, \ldots, y_n\}$, y_1 has 0.4 probability of being generated by f_1, y_2 has 0.65 probability of being generated by f_1, ..., and y_n has 0.1 probability of being generated by f_1. Of course, in reality, y_i is either generated by f_1 or f_2, but not both. However, we are uncertain about which density generated y_i. If we associate f_1 or f_2 directly to y_i, as in k-means algorithm, we may get it wrong. That is why we only assign probabilities here. If $\mathbb{P}(y_i \sim f_1)$ is close to 1, then it represents that we are pretty certain that y_i is generated by f_1. If $\mathbb{P}(y_i \sim f_1)$ is close to 0 then we reckon it is unlikely that y_i is generated by f_1. Having assigned the "score" for each data

point, we can do a weighted average to get our estimate for μ_1:

$$\hat{\mu}_1 = \frac{0.4y_1 + 0.65y_2 + \cdots + 0.1y_n}{0.4 + 0.65 + \cdots + 0.1}$$

(estimate for μ_2 is similar) Note that the k-means approach corresponds to the following:

$$\hat{\mu}_1 = \frac{[y_1 \sim f_1]y_1 + [y_2 \sim f_1]y_2 + \cdots + [y_n \sim f_1]y_n}{[y_1 \sim f_1] + [y_2 \sim f_1] + \cdots + [y_n \sim f_1]} \tag{2}$$

where $[y_i \sim f_1] = 1$ if we assign y_i to f_1, and 0 if not. Thus in this sense the EM algorithm can be seen as a generalization of the k-means algorithm.

How do we assign those probabilities for each data point? We use the current estimate of the parameters $(\mu_1^{(t-1)}, \mu_2^{(t-1)}, p^{(t-1)})$, as well as the data point y_i:

$$\begin{aligned}
\mathbb{P}(y_i \sim f_1) := \mathbb{P}(Z_i = 1|y_i) &= \frac{\mathbb{P}(y_i, Z_i = 1)}{\mathbb{P}(y_i)} \\
&= \frac{\mathbb{P}(y_i|Z_i = 1) \cdot \mathbb{P}(Z_i = 1)}{\mathbb{P}(y_i)} \\
&= \frac{f(y_i, \mu_1^{(t-1)}) \cdot p^{(t-1)}}{\mathbb{P}(y_i)} \\
&= \frac{f(y_i, \mu_1^{(t-1)}) \cdot p^{(t-1)}}{\mathbb{P}(y_i|Z_i = 1) \cdot \mathbb{P}(Z_i = 1) + \mathbb{P}(y_i|Z_i = 0) \cdot \mathbb{P}(Z_i = 0)} \\
&= \frac{f(y_i, \mu_1^{(t-1)}) \cdot p^{(t-1)}}{f(y_i, \mu_1^{(t-1)}) \cdot p^{(t-1)} + f(y_i, \mu_2^{(t-1)}) \cdot (1-p)^{(t-1)}}
\end{aligned}$$

($\mathbb{P}(y_i \sim f_2) := \mathbb{P}(Z_i = 0|y_i) = 1 - \mathbb{P}(y_i \sim f_1)$) Thus, the score is $\mathbb{P}(Z_i = 1|y_i)$. Given the current model (i.e. the current estimated parameters), and given the point y_i, we update our belief about how likely it is draw from the first, or the second density. I color data points by purple, and I color parameters by green. All of them are numerical values. Also recall the familiar normal density function $f(y, \mu_i) = (1/\sqrt{2\pi}) \exp\{-(y-\mu_i)^2\}$. We see from the last line that all variables are known, and so $\mathbb{P}(Z_i = 1|y_i)$ is a concrete numerical value, a number.

This, is our expectation step. Note that $\mathbb{P}(Z_i = 1|y_i) = \mathbb{E}(Z_i = 1|y_i)$. Each data point y_i is draw from either f_1 or f_2. We don't know which one. So we assign probabilities, or in other words we use $\mathbb{E}(Z_i|y_i)$ to estimate the unknown Z_i. This way we can get rid of Z_i, turn them to numerical values $\{\mathbb{P}(y_i \sim f_1)\}_{i=1}^n$, and then we can take the weighted average to get the estimate for the parameter μ_1. This can be justified by maximum likelihood estimation.

4.3 Expectation and Maximization of Likelihoods

Notation: $y = (y_1, \ldots, y_n)$ and $Z = (Z_1, \ldots, Z_n)$. To further save symbols, we use f_1 to mean $f_1(y_i, \mu_1)$ and f_2 to mean $f_1(y_i, \mu_2)$. This should be clear from the context.

The likelihood is[2]

$$\begin{aligned} L_c(\theta, Z) = p(y, Z, \theta) &= p(y|Z)p(Z) \\ &= [f(y_1|Z_1)p(Z_1)]\cdots[f(y_n|Z_n)p(Z_n)] \\ &= \prod_{i=1}^{n} f_1^{Z_i} f_2^{(1-Z_i)} p^{Z_i}(1-p)^{(1-Z_i)} \end{aligned}$$

Here $\theta = (\mu_1, \mu_2, p)$. We want to choose θ to maximize the log likelihood. The log likelihood is

$$\begin{aligned} \ell_c(\theta, Z) = \log L_c(\theta, Z) &= \sum_{i=1}^{n} \log f_1^{Z_i} f_2^{(1-Z_i)} p^{Z_i}(1-p)^{(1-Z_i)} \\ &= \sum_{i=1}^{n} [Z_i \log f_1 + (1-Z_i)\log f_2 + Z_i \log p + (1-Z_i)\log(1-p)] \end{aligned} \tag{3}$$

I use capital Z_i throughout to emphasize that Z_i is unknown to us. In this view, the log likelihood is a function of the random variables $\{Z_1, \ldots, Z_n\}$. We can then take the conditional expectation of the log likelihood given $Y_1 = y_1, \ldots, Y_n = y_n$:

$$\begin{aligned} \mathbb{E}[\ell_c(\theta, Z)|y] &= \sum_{i=1}^{n} [\tau_i \log f_1 + (1-\tau_i)\log f_2 + \tau_i \log p + (1-\tau_i)\log(1-p)] \\ &= \sum_{i=1}^{n} \left[-\tau_i(y_i-\mu_1)^2 - (1-\tau_i)(y_i-\mu_2)^2 + \tau_i \log p + (1-\tau_i)\log(1-p)\right] + \text{constant} \end{aligned}$$

where $\tau_i = \mathbb{E}(Z_i|y_i)$ is the calculated value as discussed above. We can see that the this function is composed of three sums (we omit those constants): $g(\theta) = g(\mu_1, \mu_2, p) := E[\ell_c(\theta, Z)|y] = \ell_1(\mu_1) + \ell_2(\mu_2) + \ell_3(p)$ where

$$\ell_1(\mu_1) = -\sum_{i=1}^{n} \tau_i(y_i-\mu_1)^2$$

$$\ell_2(\mu_2) = -\sum_{i=1}^{n} (1-\tau_i)(y_i-\mu_2)^2$$

$$\ell_3(p) = -\sum_{i=1}^{n} \tau_i \log p + (1-\tau_i)\log(1-p)$$

Maximizing this function is easy:

$$\frac{\partial g}{\partial \mu_1} = \frac{d\ell_1}{d\mu_1} = 2\sum_{i=1}^{n} \tau_i(y_i-\mu_1) = 0 \Longrightarrow \mu_1 = \frac{\sum_{i=1}^{n} \tau_i y_i}{\sum_{i=1}^{n} \tau_i}$$

[2]Note $[Z_i = 1] = Z_i$ and $[Z_i = 0] = 1 - Z_i$, so for example $f_1^{[Z_i=1]} f_2^{[Z_i=0]} = f_1^{Z_i} f_2^{(1-Z_i)}$.

$$\frac{\partial g}{\partial \mu_2} = \frac{d\ell_2}{d\mu_2} = 2\sum_{i=1}^{n}(1-\tau_i)(y_i-\mu_2) = 0 \Longrightarrow \mu_2 = \frac{\sum_{i=1}^{n}(1-\tau_i)y_i}{\sum_{i=1}^{n}(1-\tau_i)}$$

$$\frac{\partial g}{\partial p} = \frac{d\ell_3}{dp} = \sum_{i=1}^{n}\left(\frac{\tau_i}{p} + \frac{\tau_i - 1}{1-p}\right) = \sum_{i=1}^{n}\frac{\tau_i - p}{p(1-p)} = 0 \Rightarrow \sum_{i=1}^{n}(\tau_i - p) = 0 \Rightarrow p = \sum_{i=1}^{n}\tau_i/n$$

Note that in Eq. (3), if we assign Z_i to be either 1 or 0 by some rules, as in the k-means algorithm, then the ith term in the summation would become either $[\log f_1 + \log p]$ (when $Z_i = 1$) or $[\log f_2 + \log(1-p)]$ (when $Z_i = 0$). Maximizing the resulting log-likelihood function would be equally easy. In this case, the MLE for μ_1 would be the mean value of all y_is for which $Z_i = 1$, the MLE for μ_2 would be the mean value of all y_is for which $Z_i = 0$, as in Eq. (2). Similarly, the MLE for p would be the proportion of is for which $Z_i = 1$. Instead of assign Z_i to either 1 or 0, the EM algorithm substitute Z_i by $\mathbb{E}(Z_i|y_i)$, and the resulting log-likelihood function is as well easy to maximize. The output of MLE are new parameters μ_1 and μ_2, along with p. We can then use the parameters those parameters to further calculate $\mathbb{E}(Z_i|y_i)$. This is like in k-means algorithm, after we assigned each point to the nearest center, we calculate new mean value for each group (i.e. new parameters), and then we use the new means to further assign each point to a class.

I have used the mixture of two normal densities as example. But we can see that the approach can be easily generalized: Z_i can take values from $\{1, \ldots, k\}$. If $Z_i = j$ then the point is draw from some density f_j. The log-likelihood function, in terms of the random variable Z_i, is still a linear combination of $[Z_i = 1], \ldots, [Z_i = k-1]$ ($[Z_i = j]$ is a random variable that is equal to 1 if $Z_i = j$ and 0 otherwise). We can then substitute $[Z_i = j]$ by $\mathbb{E}[[Z_i = j]|y_i]$. The resulting log-likelihood function can be separated, and we can choose each parameter separately, as above.

The general procedure is

EM Algorithm

1. Initialize parameters θ_0.
   Repeat until convergence: {
2. Calculate $\mathbb{E}(Z_i|y_i)$ for each i. This is like a soft assignment to clusters.
3. $\theta = \text{argmax}_\theta\, \mathbb{E}[\ell_c(\theta, Z)|y]$ where $\mathbb{E}(Z_i|y_i)$s calculated from step 2 appear in the function. }

The next natural question is, how do we know that this iterative procedure can truly maximize the complete log-likelihood? To simplify notation, we focus on each individual data point x. The log-likelihood of the data is the sum of the log-likelihood of each individual data point.

The log-likelihood is

$$\begin{aligned}
\log p(x;\theta) &= \log \sum_z p(x,z;\theta) \\
&= \log \sum_z p(x,z;\theta)\frac{q(z)}{q(z)} \\
&= \log \mathbb{E}_q \left[\frac{p(x,z;\theta)}{q(z)}\right] \\
&\geq \mathbb{E}_q \log \left[\frac{p(x,z;\theta)}{q(z)}\right] \\
&:= \mathcal{L}(q,\theta)
\end{aligned}$$

where $q(z)$ denotes any distribution of Z. The fourth line is derived from Jensen's inequality $f(\mathbb{E}X) \geq \mathbb{E}f(X)$ for concave function f and random variable X, and from the fact that $\log(x)$ is a concave function. We can see that $\mathcal{L}(q,\theta)$ gives a lower bound to the log-likelihood function. We want the bound to be as tight as possible, so we would like to shift q so as to make $\mathcal{L}(q,\theta)$ to be as close to $p(x;\theta)$ as possible. When does the equality holds? It is when $\mathbb{E}X = X$, i.e. X is a constant. In order for the expression inside the square bracket to be a constant, we'd like to have

$$q(z) \propto p(x,z,\theta)$$

And since we also require $q(z)$ to be a probability distribution ($\sum_z q(z) = 1$), the best $q(z)$ is now

$$q^*(z) = \frac{p(x,z;\theta)}{\sum_z p(x,z;\theta)} = \frac{p(x,z;\theta)}{p(x;\theta)} = p(z|x;\theta).$$

Now we see why we take the conditional expectation in the E-step: this way $\mathcal{L}(q,\theta)$ would touch the log-likelihood function at current interaction of the parameters $\theta^{(t)}$, so we have a tight bound, and it is $\mathcal{L}(q^*(z),\theta)$ that we are trying to maximize in the M-step. To put it differently, the EM algorithm can be seen as *coordinate ascent*: the E-step maximizes $\mathcal{L}(q,\theta)$ from the first coordinate, and the M-step maximizes $\mathcal{L}(q,\theta)$ from the second coordinate. We can imagine that we follow a zigzag path to reach a (local) maximum of $\mathcal{L}(q,\theta)$. Thus, both the E-step and the M-step contribute to the increase of the objective function. Both of them are equally important: without the E-step we do not know where we are going, and without the M-step we also have no point in doing anything else.

Fig. 2[3] provides a visualization:

[3]The figure is taken from the Internet.

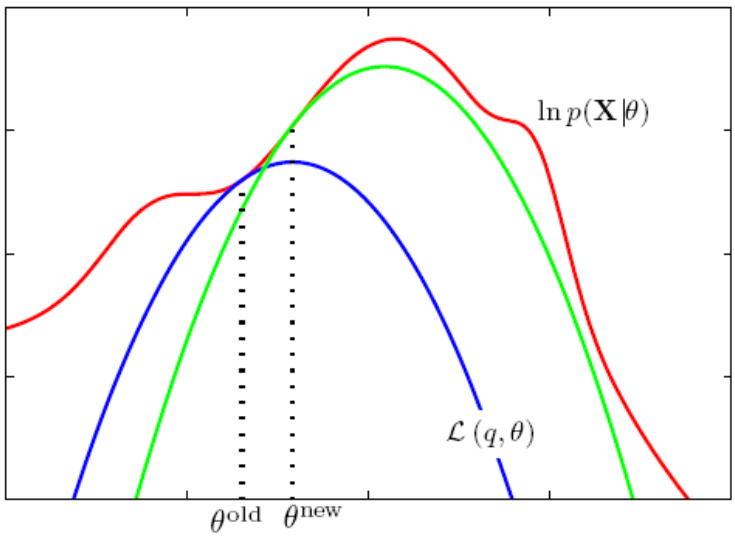

Figure 2: visualization of EM algorithm

5 Simulation Methods

5.1 Monte Carlo Integration

Monte Carlo integration is a numerical technique for calculating integrals and summations. Suppose we want to evaluate the integral

$$I = \int_a^b h(x)dx$$

for some complicated function $h : [a, b] \to \mathbb{R}$. We can rewrite the above as

$$I = \int_a^b [h(x)(b-a)] \cdot \frac{1}{b-a} dx = \int_a^b w(x) f(x) dx$$

with $w(x) = h(x)(b-a)$ and $f(x) = 1/(b-a)$. We see that f is the probability density for the uniform distribution $U(a, b)$. Hence

$$I = \mathbb{E}_f(w(X))$$

with $X \sim U(a, b)$. We transformed the problem of calculating an integral to calculating the expectation of some (simple) random variable. We approximate this expectation by its empirical counterpart, namely we draw samples $\{X_1, \ldots, X_N\}$ from the distribution of the random variable and then take arithmetic average

$$\widehat{I} = \frac{1}{N} \sum_{i=1}^{N} w(X_i).$$

By law of large numbers (Theorem 1.7), $\widehat{I}$ converges to the true theoretical value I as $N \to \infty$. To emphasize again, in Monte Carlo method we transform an integration problem to a probability problem (calculating expectations), and it is at this point that we resort to numerical techniques. We use tools from probability theory to guarantee theoretical convergence of the approximation to the true value.

It is also possible to calculate the standard error of the estimate $\widehat{I}$. It is

$$\widehat{se} = \frac{s}{\sqrt{N}}$$

where $s^2 = \sum_{i=1}^{N} (w(X_i) - \widehat{I})^2/(N-1)$. A $1-\alpha$ confidence interval for I is $\widehat{I} \pm z_{\alpha/2}\widehat{se}$.

Example 5.1. Let $h(x) = x^3$. Then $I = \int_0^1 x^3 dx = 1/4 = 0.25$. To evaluate the integral using Monte Carlo integration method, we just need to draw samples $\{x_1, \ldots, x_N\}$ from the uniform distribution on $[0, 1]$, and then calculate $\sum_{i=1}^{N} x_i^3/N$. For $N = 10,000$ we get $\widehat{I} = 0.248$ with a standard error of 0.0028.

Example 5.2. Let

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

be the standard normal density. Suppose we want to compute the CDF at some point x:

$$I = \int_{-\infty}^{x} f(s)ds = \Phi(x).$$

We can write the integration as

$$I = \int_{-\infty}^{\infty} h(s) f(s)ds$$

where

$$h(s) = \begin{cases} 1 & s < x \\ 0 & s \geq x. \end{cases}$$

This is $\mathbb{E}[h(X)]$ with $X \sim N(0,1)$. Thus we can draw samples $\{x_1, \ldots, x_N\}$ from the standard normal distribution $N(0,1)$, calculate $\{h(x_1), \ldots, h(x_N)\}$, and then calculate arithmetic average. It is the number of observations that are less or equal to x divided by N.

Example 5.3. consider estimating the area of the unit circle (i.e.π) $A = \{(x,y) \mid x^2 + y^2 \leq 1\}$ in $\mathbb{R}^2$. We can place dots randomly in the square $S = \{(x,y) \mid |x| \leq 1, |y| \leq 1\}$, count the number of dots in the circle, and then divide the count by the total number of dots, finally multiply 4. To see this, recall that the uniform distribution on S has density

$$f(x,y) = \begin{cases} \frac{1}{4} & \text{for } (x,y) \in S, \\ 0 & \text{otherwise.} \end{cases}$$

And so

$$\begin{aligned} \text{Area of } A &= \int_A 1 dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbb{1}_A(x,y)dxdy = \int_S (\mathbb{1}_A(x,y)\cdot 4)\cdot\frac{1}{4}dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (\mathbb{1}_A(x,y)\cdot 4) f(x,y)dxdy = \mathbb{E}_f[\mathbb{1}_A(X)\cdot 4] \end{aligned}$$

where $X \in \mathbb{R}^2$ is uniformly distributed on S. Thus, to approximate $\mathbb{E}_f[\mathbb{1}_A(X)\cdot 4]$, we uniformly draw many samples from the square S, and for each sample we calculate if it is in the circle. Then the approximation is given by

$$\frac{\sum_{i=1}^N \mathbb{1}(x_i, y_i)\cdot 4}{N} = 4 \cdot \frac{\sum_{i=1}^N \mathbb{1}(x_i, y_i)}{N}.$$

5.2 Importance Sampling

Importance sampling is a variance reduction technique that can be used in the Monte Carlo method. The idea behind importance sampling is that certain values of the input random variables in a simulation have more impact on the parameter being estimated than others. If these "important" values are emphasized by sampling more frequently, then the estimator variance can be reduced. Hence, the basic methodology in importance sampling is to choose a distribution which "encourages" the important values. Let g be a density that we know how to simulate from.

$$I = \int h(x) f(x) dx = \int \frac{h(x) f(x)}{g(x)} g(x) dx = \mathbb{E}_g(w(X))$$

with $w = hf/g$. We can simulate $X_1, \ldots, X_N$ from g and estimate I by

$$\widehat{I} = \frac{1}{N} \sum_{i=1}^{N} w(X_i) = \frac{1}{N} \sum_{i=1}^{N} \frac{h(X_i) f(X_i)}{g(X_i)}.$$

There is a potential problem: $\widehat{I}$ may have infinite standard error, as can be seen from the second moment:

$$\mathbb{E}_g(w^2(X)) = \int \left(\frac{h(x) f(x)}{g(x)} \right)^2 g(x) dx = \int \frac{h(x)^2 f(x)^2}{g(x)} dx.$$

If g has thinner tails than f, then this integral may be infinite.

5.3 Accept-Reject Algorithm

The accept-reject algorithm is a method for sampling from a distribution F with density f. If the inverse of F can be worked out easily, then we can first draw random samples $\{u_1, \ldots, u_n\}$ from the uniform distribution $\mathcal{U}(0, 1)$. Then $\{x_1 = F^{-1}(u_1), \ldots, x_n = F^{-1}(u_n)\}$ should be a random sample from the distribution F. To see this, let $X = F^{-1}(U)$ and note $\mathbb{P}\{X \leq x\} = \mathbb{P}\{F^{-1}(U) \leq x\} = \mathbb{P}\{U \leq F(x)\} = F(x)$, so that the distribution of X is indeed F. However, if the inverse of F cannot be worked out analytically, then we can resort to the accept-reject algorithm. The idea of the algorithm is to enclose the density f by some density g that we know how to sample from. Then place a bunch of dots on the graph beneath g. The x coordinate of each dot is placed according to the density g, and the y coordinate of each dot is placed randomly beneath g (note that if we enclose f by a rectangle, i.e. use the uniform distribution, the dots are placed randomly in the rectangle). Then remove the dots that fall out of the graph of f. The x coordinates of the remaining dots should be a random sample from the distribution f.

Let $g(x)$ be the density of some random variable Y that we know how to sample. Let $\alpha = \text{argmax}_{x \in S}\{f(x)/g(x)\}$, where S is the support. Then $f(x) \leq \alpha g(x)$ for all $x \in S$.
Repeat:

- Draw x from g and u from $\mathcal{U}(0,1)$. If

$$u \leq \frac{f(x)}{\alpha g(x)}$$

then we accept x, otherwise we reject x.

The output is a sample $\{x_1, \ldots, x_n\}$ from the distribution f.

We can show the correctness of the algorithm as follows:

$$\mathbb{P}\{X \leq x\} = \mathbb{P}\left\{Y \leq y \mid U \leq \frac{f(Y)}{\alpha g(Y)}\right\} = \frac{\mathbb{P}\left\{Y \leq x, U \leq \frac{f(Y)}{\alpha g(Y)}\right\}}{\mathbb{P}\left\{U \leq \frac{f(Y)}{\alpha g(Y)}\right\}}$$

$$= \frac{\int_{-\infty}^{x} \int_{0}^{\frac{f(z)}{\alpha g(z)}} du \cdot g(z) dz}{\int_{-\infty}^{\infty} \int_{0}^{\frac{f(z)}{\alpha g(z)}} du \cdot g(z) dz} = \int_{-\infty}^{x} f(z) dz = F(x).$$

5.4 Markov Chain Monte Carlo (MCMC)

Consider again the integration problem

$$I = \int h(x) f(x) dx.$$

The *Markov Chain Monte Carlo (MCMC)* method constructs a Markov chain $\{X_i\}_{i=1}^{\infty}$ whose stationary distribution is f. Under certain conditions it will follow that

$$\frac{1}{N} \sum_{i=1}^{N} h(X_i) \xrightarrow{p} \mathbb{E}_f[h(X)] = I.$$

We have the following theorem

Theorem 5.4. An irreducible, ergodic Markov chain has a unique stationary distribution π. The limiting distribution exists and is equal to π. If g is any bounded function, then with probability 1

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} g(X_n) \to \mathbb{E}_\pi(g(X)) = \sum_j g(j) \pi_j.$$

5.4.1 The Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm is a specific MCMC method. Let $q(y \mid x)$ be an arbitrary, friendly distribution (we know how to sample from $q(y \mid x)$). It is called the *proposal distribution*.

Choose an arbitrary X_0. Suppose we have generated $X_0, X_1, \ldots, X_i$. To generate X_{i+1},

1. Generate a *proposal* or *candidate* $Y \sim q(y \mid X_i)$.
2. Evaluate $r = r(X_i, Y)$ where
$$r(x, y) = \min\left\{\frac{f(y)}{f(x)}\frac{q(x \mid y)}{q(y \mid x)}, 1\right\}.$$
3. Set
$$X_{i+1} = \begin{cases} Y & \text{with probability } r \\ X_i & \text{with probability } 1 - r. \end{cases}$$

A simple way to execute step 3 is to generate u from $U(0,1)$ and set X_{i+1} to Y if $u < r$ and X_i otherwise. A common choice for $q(y \mid x)$ is $N(x, b^2)$ for some $b > 0$. This means that the proposal is draw from a normal distribution centered at the current value. In this case the proposal density q is symmetric and r simplifies to

$$r = \min\left\{\frac{f(y)}{f(x)}, 1\right\}.$$

The transition function $p(x, y)$ that the algorithm designs is

$$p(x, y) = r(x, y) \cdot q(y \mid x). \tag{4}$$

We show that it satisfies the *detailed balance* property

$$f(x)p(x, y) = f(y)p(y, x) \tag{5}$$

so that

$$\int f(y)p(y, x)dy = \int f(x)p(x, y)dy = f(x)\int p(x, y)dy = f(x),$$

which implies that f is a stationary distribution of the Markov chain with transition function $p(x, y)$.

Consider two states x and y. Either

$$f(x)q(y \mid x) < f(y)q(x \mid y) \quad \text{or} \quad f(x)q(y \mid x) > f(y)q(x \mid y).$$

In the first (second) case, the probability of flowing from x to y is smaller (larger) than the probability of flowing from y to x. Without loss of generality assume the later, i.e. there is too much flow from x to y and too little from y to x. This implies that we should set

$$r(x,y) = \min\left\{\frac{f(y)}{f(x)}\frac{q(x \mid y)}{q(y \mid x)}, 1\right\} = \frac{f(y)}{f(x)}\frac{q(x \mid y)}{q(y \mid x)},$$
$$r(y,x) = \min\left\{\frac{f(x)}{f(y)}\frac{q(y \mid x)}{q(x \mid y)}, 1\right\} = 1.$$

Thus we should reduce the probability of jumping from x to y and increase the probability of jumping from y to x. The transition probability according to Eq. (4) is

$$p(x,y) = r(x,y) \cdot q(y \mid x) = \frac{f(y)}{f(x)}\frac{q(x \mid y)}{q(y \mid x)} \cdot q(y \mid x) = \frac{f(y)}{f(x)} q(x \mid y). \tag{6}$$

On the other hand,

$$p(y,x) = r(y,x) \cdot q(x \mid y) = q(x \mid y). \tag{7}$$

From Eq. (6) and Eq. (7), we see that indeed

$$f(x)p(x,y) = f(y)p(y,x).$$

5.4.2 Gibbs Sampling

Gibbs sampling is a way to turn high-dimensional problem into several one-dimensional problem. Let's illustrate this with two dimensional data. Suppose we want to sample from $f_{X,Y}(x,y)$, and suppose we know how o simulate from $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$. Let (X_0, Y_0) be starting values. After we drawn $(X_0, Y_0), \ldots, (X_n, Y_n)$, the next sample (X_{n+1}, Y_{n+1}) is determined by

$$X_{n+1} \sim f_{X|Y}(x \mid Y_n),$$
$$Y_{n+1} \sim f_{Y|X}(x \mid X_{n+1}).$$

6 Selected Topics

6.1 Principal Component Analysis

Suppose we have a data matrix $X_{n\times k}$, namely n data points of k random variables $(X_1, \ldots, X_k)$. Suppose k is very large, e.g. 1000.....then in this case it may be very difficult to do data analysis with this original data. For example, it is hard to make sense of linear regressions like this:

$$y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{1000} X_{1000} + \varepsilon,$$

or to visualize clustering of data points in 1000 dimensional space. Also, many variables may be correlated, so there may be redundancy in the data. Thus we may want to project our data into smaller spaces (e.g. 2 or 3 dimensional space), with as little distortion as possible. For example, if two points x and y are far apart in the original data set, then we would like the two projected points x' and y' to be far apart. In general, we are trying to represent the variations in the data, originally spread across thousands of dimensions, by variations in terms of two or three variables (so that we can, for example, visualize the high dimensional data without losing much information)

To transform our data into new space, we use k orthonormal unit vectors $u_1, \ldots, u_k$. We let $u = (u_1, \ldots, u_k)$ denote the $k \times k$ matrix. Our new data matrix would be Xu. Each row r (data point) is transformed into the new row (coordinate)

$$[\langle r, u_1\rangle, \langle r, u_2\rangle, \ldots, \langle r, u_k\rangle] = [r \cdot u_1, r \cdot u_2, \ldots, r \cdot u_k],$$

so that each point d is represented in the new coordinate system as

$$d = \langle d, u_1\rangle u_1 + \langle d, u_2\rangle u_2 + \ldots + \langle d, u_k\rangle u_k.$$

Note that $X_1' = \begin{pmatrix} d_1 \cdot u_1 \\ \cdots \\ d_n \cdot u_1 \end{pmatrix} = Xu_1$ and similarly each vector of realizations of variable X_i is transformed into $X_i' = Xu_i$.

At this point, we note that the data matrix has to be standardized, so that $\mathbb{E}[X_i] = 0$ for each $i = 1, \ldots, k$. This ensures that the covariance matrix is $X^T X/(n-1)$.

Now, we would like to maximize the variance of the first component. The variance is

$$\begin{aligned}(n-1) \cdot var(Xu_1) &= (Xu_1)^T (Xu_1) \\ &= \|Xu_1\|^2 = \langle Xu_1, Xu_1\rangle \\ &= \langle u_1, X^T X u_1\rangle = \langle u_1, u_1\rangle_{X^T X} \\ &= \|u_1\|^2_{X^T X}\end{aligned}$$

Namely, we want to maximize $\langle u_1, u_1 \rangle_{X^T X}$ subject to the constraint that $\|u_1\|^2 = 1$. Since $X^T X$ is symmetric and positive-definite, by the spectral theorem there exists an orthonormal basis such that $X^T X$ is

$$\begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_k \end{pmatrix}.$$

We assume the eigenvalues has been arranged so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$. Suppose u_1 has coordinate $(u_{11}, u_{21}, \ldots, u_{k1})$ in this basis. Then

$$\langle u_1, u_1 \rangle_{X^T X} = \lambda_1 u_{11}^2 + \lambda_2 u_{21}^2 + \ldots + \lambda_k u_{k1}^2,$$

where $u_{11}^2 + u_{21}^2 + \cdots + u_{k1}^2 = 1$. The coordinate acts like a weight, so it is obvious that to reach maximum u_1 has to be $(1, 0, \ldots, 0)$, i.e. the unit length eigenvector corresponding to the first eigenvalue λ_1. By the same token, to

$$\max_{u_i} var(Xu_i) = \langle u_i, u_i \rangle_{X^T X} \quad \text{s.t.} \quad \|u_i\|^2 = 1$$

we just need to let $u_i^* = (0, 0, \ldots, 1, \ldots, 0)$, i.e. the unit length eigenvector corresponding to λ_i.

From our (optimal) solution $u_1, \ldots, u_k$, the (optimal) values for $(n-1) \cdot var(Xu_1), \ldots, (n-1) \cdot var(Xu_k)$ are exactly the eigenvalues $\lambda_1, \ldots, \lambda_k$. We have

$$(n-1) \cdot (var(Xu_1) + \cdots + var(Xu_k)) = \lambda_1 + \cdots + \lambda_k = trace(X^T X)$$

and also

$$(n-1) \cdot (var(X_1) + \cdots + var(X_k)) = trace(X^T X)$$

so that

$$var(Xu_1) + \cdots + var(Xu_k) = var(X_1) + \cdots + var(X_k),$$

i.e. our transformation does not lose information. The first transformed variable $X_1' = Xu_1$ has the largest variance, the second X_2' has the second largest variance etc. Furthermore,

$$\begin{aligned} Cov(X_i', X_j') &= Cov(Xu_i, Xu_j) = (Xu_i)^T (Xu_j)/(n-1) \\ &= u_i^T X^T X u_j/(n-1) \\ &= \lambda_j u_i^T u_j/(n-1) \\ &= \lambda_j \cdot 0 \\ &= 0. \end{aligned}$$

so that the new variables are all uncorrelated.

6.2 Clustering

Given a set of data points, clustering gives a *partition*. We define the class function

$$C : X \to \{1, \ldots, k\}$$

that maps each data point to its class. We use $C_1, \ldots, C_k$ to denote the pre-image of the range, i.e. the partition.

A popular clustering algorithm is K-means clustering. Given a set of observations $X = \{x_1, \ldots, x_n\}$, where each observation is d-dimensional real vector, k-means clustering aims to partition the n-observations into k ($k \leq n$) sets $C = \{C_1, \ldots, C_k\}$ so as to minimize the within-cluster variance, i.e. minimize

$$V(C_1) + V(C_2) + \cdots + V(C_k) = \sum_{j=1}^{k} \sum_{x \in C_j} \|x - \bar{C}_j\|^2$$

Note that the centroids $\{\bar{C}_j\}_{j=1,\ldots,k}$ are in general not points in the original data sets. The problem is in general NP-Hard. Instead, we shall use a heuristic algorithm, often called k-means algorithm or Lloyd's algorithm. It has three steps: initialization, partition, and update.

1. Initialize the mean of each class:

$$C_1.mean = c_1^0, \ldots, C_k.mean = c_k^0.$$

2. Repeat until convergence:

   `for each` $x \in X$:

   $x \in C_j$ `if` $C_j = \underset{C}{\operatorname{argmin}}\, d(x, C) = \underset{C}{\operatorname{argmin}}\, d(x, C.mean)$

The running time should be obvious: At each iteration we compute k distances $\{d(x, C_j)\}_{j=1,\ldots,k}$ for each data point x. There are n data points, so each iteration takes a total of $n \cdot k$ computations. Each computation of distance takes d summations. So the total running time is $O(n \cdot k \cdot d \cdot T)$, where T is the number of iterations.

There are also metrics for evaluating clustering results, for example the silhouette coefficient. A higher silhouette coefficient score relates to a model with better clusters. It is defined for each sample and is composed of two scores:

- $a(x)$: The mean distance between x and all other points in the same class.

$$a(x) = \mathbb{E}[d(x, y) \mid y \in C(x)]$$

- $b(x)$: The mean distance between x and all other points in the next nearest cluster.

$$b(x) = \mathbb{E}[d(x, y) \mid y \in \operatorname*{argmin}_{C \neq C(x)} d(x, C)]$$

The silhouette coefficient function $s : X \to \mathbb{R}$ is defined as

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}.$$

Normally we would expect $a(x) \leq b(x)$ for each x so that $s(x) \geq 0$. If $b(x) < a(x)$ then the classification for x would be bad and we have $s(x) < 0$. In the extreme case that $a(x) = 0$, i.e. $C(x) = \{x\}$, we have $s(x) = b(x)/b(x) = 1$. In the other extreme case $b(x)$ would be close to 0, so that $s(x)$ will be close to -1. Note that by definition $b(x)$ can never be 0, so $s(x)$ can never reach -1. We thus have $-1 < s(x) \leq 1$.

References

Ash, R. B. (1999). *Probability and Measure Theory*. 2nd ed. San Diego, California: Academic Press (cit. on p. 5).

Chung, K. L. and F. AitSahlia (1974). *Elementary Probability Theory*. 4th ed. Undergraduate Texts in Mathematics. New York: Springer-Verlag (cit. on p. 5).

Durrett, R. (2010). *Probability: Theory and Examples*. 4th ed. Cambridge Series in Statistical and Probabilistic Mathematics. New York: Cambridge University Press (cit. on p. 5).

Resnick, S. I. (2014). *A Probability Path*. 1st ed. Modern Birkhäuser Classics. Birkhäuser Basel (cit. on p. 5).

Shreve, S. E. (2004). *Stochastic Calculus for Finance II: Continuous-Time Models*. 1st ed. Springer Finance. New York: Springer-Verlag (cit. on p. 5).

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. New York: Springer (cit. on p. 5).