Statistical Inference

Definition 1.1. A family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of probabilities (pmf or pdf) indexed by θ is called an ***exponential family*** if there exists $k \in \mathbb{N}$, real-valued functions $\eta_1, \ldots, \eta_k$ and B on Θ, real-valued statistics $T_1, \ldots, T_k$ and a non-negative real-valued function h on χ such that the pdf/pmfs $p(x;\theta)$ of P_θ have the form

$$p(x;\theta) = \exp\left[\sum_{i=1}^{k} \eta_i(\theta) T_i(x) - B(\theta)\right] h(x). \tag{1.1}$$

The η_i are called the ***natural*** or ***canonical*** parameters, and the $T_i(x)$ are called the ***natural*** or ***canonical*** observations.

can think of exp(-B(theta)) as a *normalisation* to get the thing to integrate to 1 over x

$$p(x;\eta) = \exp\left[\sum_{i=1}^{n} \eta_i T_i(x) - B(\eta)\right] h(x).$$

Canonical form <u>note: possible even if map theta → ni is not 1-1</u>

1.2 Parsimonious parametrization

Definition 1.2. A class of probability measures $\mathcal{P} = \{p(x;\theta) : \theta \in \Theta\}$ which is an exponential family is said to be ***strictly k-parameter*** when k is minimal.

Although $\eta = (\eta_1, \eta_2, \ldots, \eta_k)$, $T = (T_1, \ldots, T_k)$ and k are not uniquely determined we call (1.2) a ***k-dimensional family***.

Definition 1.3. The functions $T_1, \ldots, T_n$ are called ***affine independent*** (***$\mathcal{P}$-affine independent*** in Liero-Zwanzig) if for any $c_0, \ldots, c_n \in \mathbb{R}$,

$$\left(\sum_{j=1}^{n} c_j T_j(x) = c_0 \ \forall x \in \mathcal{A}\right) \implies \left(c_j = 0 \text{ for } j = 0, \ldots, k\right).$$

Similarly, the functions $\eta_1, \ldots, \eta_n$ are ***affine independent*** if $\left(\sum_{j=1}^{n} c_j \eta_j(\theta) = c_0 \ \forall \theta \in \Theta\right) \implies \left(c_j = 0 \text{ for } j = 0, \ldots, k\right)$

Proposition 1.4. *The functions T_i are $\mathcal{P}$-affine independent if $\mathrm{Cov}_\theta(T)$ is positive definite for all $\theta \in \Theta$.*

Theorem 1.5. *A family is* **strictly k-dimensional** *if in (1.2) the functions $\eta_i(\theta)$ and $T_i(x)$ are affine independent.*

1.3 Support and counterexamples

Proposition 1.6. *Two probability measures $\mathbb{P}$ and $\mathbb{Q}$ are said to be equivalent if we have $\mathbb{P}(N) = 0$ iff $\mathbb{Q}(N) = 0$. If $\mathcal{P} = \{p(x;\theta) : \theta \in \Theta\}$ is an exponential family, then all $p(\cdot;\theta)$ are equivalent.* Take two thetas, and say P(N) = 0 ⇒

$$\mathbb{P}_{\theta_1}(N) = e^{-B(\theta_1)} \int \exp\left(\sum_j \eta_j(\theta_1) T_j(x)\right) h(x) \mathbf{1}_N(x) dx = 0$$

⇒ $h1_N$ = 0 a.e(x) ⇒ P_{theta}(N) = 0 for all theta

Corollary 1.7. *In an exponential family $\mathcal{P} = \{f(x;\theta), \theta \in \Theta\}$ the support of $f(x;\theta)$ does* **not** *depend on θ. We will write $\mathcal{A}$ for the common support of the $f(x;\theta)$.*

1.4 The parameter space

Definition 1.8. The ***parameter space*** is defined to be $\Theta := \{\theta : \int h(x) \exp\left[\sum_{i=1}^n \eta_i(\theta) T_i(x)\right] \mathrm{d}x < \infty\}$

(i.e. the set of θ for which f(x; θ) can be defined.)

Definition 1.9. The ***natural parameter space*** is defined to be $\Xi := \{\eta = (\eta_1, \ldots, \eta_n) : \int h(x) \exp\left[\sum_{i=1}^n \eta_i T_i(x)\right] \mathrm{d}x < \infty\}$,

i.e. the set of η for which we can define $B(\eta) := \log \int h(x) \exp\left[\sum_{i=1}^n \eta_i T_i(x)\right] \mathrm{d}x$ so that

$\tilde{f}(x;\eta) = e^{-B(\eta)} h(x) \exp\left[\sum_{i=1}^n \eta_i T_i(x)\right]$ is a pdf/pmf on $\mathcal{X}$. Observe that you can have $\eta(\Theta) \neq \Xi$ (but $\eta(\Theta) \subseteq \Xi$).

Theorem 1.10. *The natural parameter space Ξ of a strictly k-parameter exponential family is convex and contains a non-empty k-dimensional interval.* uses Holders

Definition 1.11. If the image of the parameter space $\eta(\Theta) \subseteq \Xi$ for a strictly k-parameters exponential family contains a k-dimensional open set, then it is called ***full rank***. [a]

Theorem 1.12. *Let $\mathcal{P}$ be a strictly k-parameter exponential family with natural parameter space Ξ. Then for all $\eta \in \operatorname{Int}(\Xi)$:*

(a) all moments of T (with respect to $f(x;\eta)$) exist; *(b)* $\mathbb{E}_\eta[T_i(X)] = \frac{\partial}{\partial \eta_i} B(\eta)$ $\forall i$; *and* *(c)* $\operatorname{Cov}_\eta(T_i, T_j) = \frac{\partial^2}{\partial\eta_i \partial\eta_j} B(\eta)$ $\forall i, j$.

Chapter 2: Sufficiency and Minimality

2.1 Sufficiency

Definition 2.1. Suppose $X \sim f(x;\theta)$ for some parameter θ. A ***statistic*** $T(X)$ is a function of the data which does not depend on θ.

A statistic $T(X)$ is said to be ***sufficient*** for θ if the conditional distribution of X given T does not depend on θ. That is, $f(x \mid t, \theta) = f(x \mid t)$.

Remark. In particular, this means that for any function g the map $\theta \mapsto \mathbb{E}_\theta[g(X) \mid T = t]$ is constant.

Theorem 2.2 (The Factorisation Criterion). *Suppose $X \sim f(x;\theta)$ and let $T(X)$ be any statistic. Then a statistic $T(X)$ is sufficient for θ if and only if f can be written as*

$f(x;\theta) = g(T(x), \theta) h(x)$ *for some non-negative functions g, h.* proof: $f(x;\theta) = \mathbb{P}_\theta(X = x \mid T = t)\, \mathbb{P}_\theta(T = t)$. (t = t(x))

Using sufficiency $\mathbb{P}_\theta(X = x \mid T = t) =: h(x)$ is independent of θ, and $\mathbb{P}_\theta(T = t) =: g(t, \theta)$ only depends on t and θ.

Conversely:

$$\mathbb{P}_\theta(T = t) = \sum_{x:T(x)=t} \mathbb{P}_\theta(X = x) = \sum_{x:T(x)=t} f(x;\theta) = g(t,\theta) \sum_{x:T(x)=t} h(x).$$

Thus $\mathbb{P}_\theta(X = x \mid T = t) = \frac{\mathbb{P}_\theta(X=x, T=t)}{\mathbb{P}_\theta(T=t)} = \frac{\mathbb{P}_\theta(X=x)}{\mathbb{P}_\theta(T=t)} = \frac{h(x)}{\sum_{y:T(y)=t} h(y)}$, which has no dependence on θ!

2.2 Minimality

Definition 2.3. A statistic is ***minimal sufficient*** if it can be expressed as a function of any other sufficient statistic.

Theorem 2.4. *A statistic T is minimal sufficient if and only if* $T(x) = T(y) \iff \frac{f(y;\theta)}{f(x;\theta)}$ *is independent of θ.*

(proof in notes)

2.3 Minimal sufficiency in exponential families

Theorem 2.5. *Suppose the functions $f(x;\theta) = \exp\left[\sum_{j=1}^k \eta_j(\theta) T_j(x) - B(\theta)\right] h(x)$ form a strictly k-parameter exponential family. Let $X = (X_1, \ldots, X_n)$ be a sample of i.i.d. random variables with distribution $f(x, \theta)$. Then: $T_{(n)} = \left(\sum_{i=1}^n T_1(X_i), \ldots, \sum_{i=1}^n T_k(X_i)\right)$ is minimal sufficient.*

proof:

$$\frac{f((x_1, \ldots, x_n);\theta)}{f((y_1, \ldots, y_n);\theta)} = \frac{\prod_{i=1}^n h(x_i)}{\prod_{i=1}^n h(y_i)} \exp\left[\sum_{j=1}^k \eta_j(\theta) \left(\sum_{i=1}^n T_j(x_i) - \sum_{i=1}^n T_j(y_i)\right)\right]$$

independent of θ if and only if $\sum_{i=1}^n T_j(x_i) = \sum_{i=1}^n T_j(y_i)$ for all $j = 1, \ldots, k$.

Chapter 3: The Fisher Information

Definition 3.1. For each $x \in \mathcal{X}$, the ***likelihood function*** $L(\cdot, x) : \Theta \to \mathbb{R}_+$ is defined by $L(\theta, x) = f(x, \theta)$.

The ***log-likelihood*** is often written $\ell(\theta, x) := \log L(\theta, x)$.

Reg 1. *The distributions $\{f(\cdot, \theta) : \theta \in \Theta\}$ have common support, so that $\mathcal{A} = \{x : f(x, \theta) > 0\}$ is independent of θ.*

1D Case

Reg 2. *$\Theta \subseteq \mathbb{R}$ is an open interval (finite or infinite).*

Reg 3. *For all $x \in \mathcal{A}$ and for all $\theta \in \Theta$, the derivative $\frac{\partial f(x,\theta)}{\partial \theta}$ exists and is finite.*

Definition 3.2. When Regs 1–3 are satisfied, for $x \in \mathcal{A}$ we define the ***score function*** $S(\theta, x) = \ell'(\theta, x) = \frac{\partial \log L(\theta, x)}{\partial \theta}$

Lemma 3.3. *Under Regs 1–3, for continuous distributions* $\frac{\partial}{\partial\theta} \int_{\mathcal{A}} f(x,\theta)\, \mathrm{d}x = \int_{\mathcal{A}} \frac{\partial}{\partial\theta} f(x,\theta)\, \mathrm{d}x$

and for discrete distributions $\frac{\partial}{\partial\theta} \sum_{x\in\mathcal{A}} f(x,\theta) = \sum_{x\in\mathcal{A}} \frac{\partial}{\partial\theta} f(x,\theta)$ (proof is LIR)

Theorem 3.4. *Under Regs 1–3,* $\mathbb{E}_\theta S(\theta, X) = 0 \ \forall \theta \in \Theta$.

Proof. In the continuous case,

$$\mathbb{E}_\theta[S(\theta,X)] = \int_{\mathcal{A}} \ell'(\theta,x) f(x,\theta)\,\mathrm{d}x = \int_{\mathcal{A}} \frac{\frac{\partial}{\partial\theta} f(x,\theta)}{f(x,\theta)} f(x,\theta)\,\mathrm{d}x = \frac{\partial}{\partial\theta}\int_{\mathcal{A}} f(x,\theta)\,\mathrm{d}x = \frac{\partial}{\partial\theta} 1 = 0.$$

The discrete case is similar. □

Definition 3.5. When Regs 1–3 are satisfied, we define the ***Fisher information*** to be $I_X(\theta) = \mathrm{Var}_\theta[S(\theta,X)] = \mathbb{E}_\theta[(\ell'(\theta,X))^2]$

Reg 4. *The log-likelihood ℓ is twice-differentiable for all $x \in \mathcal{A}, \theta \in \Theta$, and*

$$\frac{\partial^2}{\partial\theta^2}\int_{\mathcal{A}} f(x,\theta)\,\mathrm{d}x = \int_{\mathcal{A}} \frac{\partial^2}{\partial\theta^2} f(x,\theta)\,\mathrm{d}x \qquad \textit{(for continuous distributions)}$$

or

$$\frac{\partial^2}{\partial\theta^2}\sum_{x\in\mathcal{A}} f(x,\theta)\,\mathrm{d}x = \sum_{x\in\mathcal{A}} \frac{\partial^2}{\partial\theta^2} f(x,\theta)\,\mathrm{d}x \qquad \textit{(for discrete distributions)}$$

for all $\theta \in \Theta$.

Theorem 3.6. *Under Regs 1–4,*

$$I_X(\theta) = -\,\mathbb{E}_\theta[\ell''(\theta,X)].$$

$$\ell''(\theta,x) = \frac{\partial^2}{\partial\theta^2}\log f(x,\theta) = \frac{\partial}{\partial\theta}\frac{\frac{\partial}{\partial\theta} f(x,\theta)}{f(x,\theta)} = \frac{\left(\frac{\partial^2}{\partial\theta^2} f\right) f - \left(\frac{\partial}{\partial\theta} f\right)^2}{f^2} = \frac{\frac{\partial^2}{\partial\theta^2} f}{f} - \left(\frac{\frac{\partial}{\partial\theta} f}{f}\right)^2.$$

reg 4 ⇒ $\mathbb{E}_\theta\left[\left(\frac{\partial^2}{\partial\theta^2} f\right)/f\right] = 0$ …

Proposition 3.7 (Properties of the Fisher information).

1. ***(Information grows with sample size.)*** *If X and Y are independent random variables, then*
$$I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta).$$
In particular, if $Z = (X_1, \dots, X_n)$ where the X_i are i.i.d. copies of X, then $I_Z(\theta) = nI_X(\theta)$.
2. ***(Reparametrisation.)*** *If $\theta = h(\xi)$ where h is differentiable, then the Fisher information of X about ξ is*
$$I_X^*(\xi) = I_X(h(\xi))[h'(\xi)]^2.$$

Proof. (for the second point) The log-likelihood w.r.t. $\mathbb{P}_{h(\xi)}$ is $\ell^*(\xi) = \ln p(x; h(\xi))$ thus the score function is

$$S^*(\xi; x) = \frac{\partial}{\partial\xi}\ln p(x; h(\xi)) = \frac{\partial}{\partial\theta}\ln p(x;\theta)\big|_{\theta=h(\xi)} h'(\xi)$$

and so

$$\mathrm{Var}_\xi S^*(\xi, X) = \mathrm{Var}_\xi\left(S(h(\xi), X)h'(\xi)\right) = I_X(h(\xi))[h'(\xi)]^2.$$

□

3.2 The multivariate case

Reg 2′. $\Theta \subseteq \mathbb{R}^k$ *is an open set.*

Reg 3′. *For all $x \in \mathcal{A}$ and for all $\theta \in \Theta$, the partial derivatives of $L(\theta, x)$ exist and are finite.*

Reg 4′. *The log-likelihood ℓ has all its second partial derivatives, and these can all be commuted with summation/integration over $\mathcal{A}$.*

Definition 3.8. When Regs 1, 2′, 3′ are satisfied, we define the *score function* to be $S(\theta,x) = \nabla_\theta \ell(\theta,x) = \left(\frac{\partial}{\partial\theta_1}\ell(\theta,x), \dots, \frac{\partial}{\partial\theta_k}\ell(\theta,x)\right)^t$

Definition 3.9. When Regs 1, 2′, 3′ are satisfied, we define the ***Fisher information*** matrix to be

$I_X(\theta) = \mathrm{Cov}_\theta(S(\theta,X))$, so that $I_X(\theta)_{jr} = \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta_j}\ell(\theta,X)\frac{\partial}{\partial\theta_r}\ell(\theta,X)\right].$

Theorem 3.10. *Supposing Regs 1, 2′, 3′, 4′ hold, define the* **observed Fisher information** *matrix J by $J(\theta,x)_{jr} = -\dfrac{\partial^2\ell(\theta,x)}{\partial\theta_j\partial\theta_r}$ for $j, r = 1, \dots, k$. Then* $I_X(\theta) = \mathbb{E}_\theta[J(\theta,X)]$. (generalisation of 1D proof)

Chapter 4: Point estimation

Definition 4.1. For any function $g : \Theta \to \Gamma$ (for some set Γ), an ***estimator*** of $\gamma = g(\theta)$ is a function $T : \mathcal{X} \to \Gamma$. The value $T(X)$ is called the ***estimate*** of $g(\theta)$.

Definition 4.2. The ***bias*** of an estimator T for $\gamma = g(\theta)$ is $\text{bias}(T, \theta) = \mathbb{E}_\theta[T] - g(\theta)$.

T is called ***unbiased*** for $g(\theta)$ if $\mathbb{E}_\theta[T] = g(\theta)\ \forall\theta \in \Theta$

4.1 The method of moments

Definition 4.3. For each $k = 1, \dots, r$ let $\hat{m}_k = \frac{1}{n}\sum_{i=1}^n X_i^k$. Then the ***moment estimator*** for γ is defined as

$$\hat{\gamma}_{MME} = h(\hat{m}_1, \dots, \hat{m}_r).$$

4.2 Maximum likelihood estimators

Definition 4.4. An estimator T is called a ***maximum likelihood estimator (MLE)*** for θ if

$L(T(x), x) = \max_{\theta\in\Theta} L(\theta, x) \qquad \forall x \in \mathcal{X}$, and is denoted by $\hat{\theta}_{MLE}$.

4.3 Variance and mean squared error

Definition 4.6. The ***mean squared error (MSE)*** of an estimator T for $g(\theta)$ is defined as

$\text{MSE}_\theta(T) = \mathbb{E}_\theta[(T - g(\theta))^2]$. (This is also often called the ***quadratic loss function***.)

Proposition 4.7. *In general, for an estimator T for $g(\theta)$,* $\text{MSE}_\theta(T) = \text{Var}_\theta(T) + \underbrace{(\mathbb{E}_\theta[T] - g(\theta))^2}_{\text{bias}^2}$

In particular, if T is unbiased, $\text{MSE}_\theta(T) = \text{Var}_\theta(T)$. (proof was an exercise)

Chapter 5: MVUEs and the Cramer-Rao Lower Bound

Definition 5.1. We say T_1 is a ***uniformly better*** estimator than T_2 (or ***better in quadratic mean***) if for all $\theta \in \Theta$,

$$\text{MSE}_\theta(T_1) \leqslant \text{MSE}_\theta(T_2).$$

Remark. If $\theta = \theta_0$, then $\text{MSE}_{\theta_0}(\theta) = 0$. Hence no other estimator can be uniformly better! (so restrict to unbiased)

5.1 The CRLB in the one-dimensional case

Definition 5.2. $T = T(X_1, \dots, X_n)$ is the ***minimum variance unbiased estimator (MVUE)*** for θ (resp. for $g(\theta)$) if

- T is unbiased, and
- for all unbiased estimators $\tilde{T}$, $\text{Var}_\theta(\tilde{T}) \geqslant \text{Var}_\theta(T)\ \forall\theta \in \Theta$.

The estimator T is furthernore said to be ***regular*** if $\int_{\mathcal{A}} T(x)\frac{\partial}{\partial\theta}L(\theta; x)dx = \frac{\partial}{\partial\theta}\int_{\mathcal{A}} T(x)L(\theta; x)dx = \frac{\partial}{\partial\theta}\mathbb{E}_\theta[T(X)]$.

Theorem 5.3 (Cramer-Rao Lower Bound (CRLB) in 1 dimension). *Suppose Regs 2–4 hold and that $0 < I_X(\theta) < \infty$. Let $\gamma = g(\theta)$ where g is a continuously differentiable real-valued function with $g' \neq 0$.*

Let T be a **regular unbiased** *estimator of γ. Then* $\text{Var}_\theta(T) \geqslant \frac{|g'(\theta)|^2}{I_X(\theta)},\ \forall\theta \in \Theta$ *with equality if and only if*

$T(x) - g(\theta) = \frac{g'(\theta)S(\theta, x)}{I_X(\theta)} \quad \forall x \in \mathcal{A}\ \forall\theta \in \Theta$. In the case $g(\theta) = \theta$ the CRLB is $\text{Var}_\theta(T) \geqslant \frac{1}{I_X(\theta)}$

Remark. If T attains the CRLB, $\text{Var}_\theta(T) = \frac{|g'(\theta)|^2}{I_X(\theta)}$,

then it is clearly a MVUE. There is no guarantee that there exists an estimator which attains the bound.

(proof in notes)

Corollary 5.4. *Suppose that $\mathbb{E}_\theta[T(X)] = \theta + b(\theta)$ (so that $b(\theta)$ is the bias of T) and that T is regular. Then*

$$Var_\theta(T(X)) \geqslant \frac{|1 + b'(\theta)|^2}{I_X(\theta)}$$

5.2 Efficiency

Definition 5.5. The efficiency of an unbiased estimator T of $g(\theta)$ is the ratio of its variance and of the CRLB, that is

$$e(T,\theta) = \frac{[g'(\theta)]^2}{I_X(\theta)\mathrm{Var}_\theta T}.$$

An unbiased estimator which attains the CRLB is called ***efficient***.

Theorem 5.6. *Suppose that the distribution of $X = (X_1, \ldots, X_n)$ belongs to a one-parameter exponential family in ζ and T . Then the sufficient statistic T is an efficient estimator for the parameter $\gamma = g(\theta) = \mathbb{E}_\theta[T]$.* (proof notes)

5.3 The multivariate case

Definition 5.7. Let T, T^* be two unbiased estimators for γ. We say that T^* has a ***smaller*** covariance matrix than T at $\theta \in \Theta$ if

$u^t(\mathrm{Cov}_\theta T^* - \mathrm{Cov}_\theta T)u \leqslant 0 \quad \forall u \in \mathbb{R}^m$, and we write $\mathrm{Cov}_\theta T^* \preceq \mathrm{Cov}_\theta T$.

Theorem 5.8 (Cramer-Rao Lower Bound in m dimensions). *Suppose Regs 1, 2′, 3′, 4′ hold and that $I_X(\theta)$ is not singular. Then the CRLB is*

$$\mathrm{Cov}_\theta T \succeq (D_\theta g)(\theta) I_X(\theta)^{-1} (D_\theta g)(\theta)^t \qquad \forall \theta \in \Theta,$$

where $D_\theta g$ is the Jacobian matrix, so $(D_\theta g)(\theta)_{ij} = \dfrac{\partial g_i(\theta)}{\partial \theta_j}$.

5.4 MLEs and MVUEs

Theorem 5.9. *Under Regs 1, 2′, 3′, 4′, if $\hat{\theta}_{MLE}$ is the MLE for θ and if there exists $\tilde{\theta}$ which is unbiased and attains the CRLB, then $\tilde{\theta} = \hat{\theta}_{MLE}$ almost surely.* (proof notes)

Chapter 6: The Rao-Blackwell and Lehmann-Scheff´e theorems

Theorem 6.1 (Rao-Blackwell Theorem). *Let $X \sim P_\theta$ and let T be a sufficient statistic. Let $\hat{\gamma}$ be an unbiased estimator for $\gamma = g(\theta) \in \mathbb{R}^k$.*

Define $\hat{\gamma}_T = \mathbb{E}_\theta[\hat{\gamma} \mid T]$. Then: *1. $\hat{\gamma}_T$ is a function of T alone and does not depend on θ,* *2. $\mathbb{E}_\theta[\hat{\gamma}_T] = \gamma \ \forall \theta \in \Theta$,*

3. $\mathrm{Cov}_\theta(\hat{\gamma}_T) \preceq \mathrm{Cov}_\theta(\hat{\gamma})$ (which reduces to $\mathrm{Var}_\theta(\hat{\gamma}_T) \leqslant \mathrm{Var}_\theta(\hat{\gamma})$, in the case $k = 1$).

If $\mathrm{tr}(\mathrm{Cov}_\theta(\hat{\gamma})) < \infty$ then $\mathrm{Cov}_\theta(\hat{\gamma}_T) = \mathrm{Cov}_\theta(\hat{\gamma})$ if and only if $\hat{\gamma} = \gamma$ almost surely. (proof notes)

Definition 6.2. A statistical model $\{P_\theta : \theta \in \Theta\}$ is called ***complete*** if for any $h : \mathcal{X} \to \mathbb{R}$,

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

A statistic T is called ***complete*** if the model $\{P_\theta^T : \theta \in \Theta\}$ is complete, i.e.

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

Lemma 6.3. *Let*

$$p(x;\theta) = \exp\Big[\sum_{i=1}^{k} \eta_i(\theta) T_i(x) - B(\theta)\Big] h(x), \theta \in \Theta$$

be a strictly k-parameter exponential family. The joint distribution of the natural observation vector $T(X) = (T_1(X), \ldots, T_k(X))$ belongs to a strictly k parameter exponential family with natural parameters $\eta_1(\theta), \ldots, \eta_k(\theta)$. (proof notes)

Theorem 6.4 (Completeness for exponential families). *If $\mathcal{P}$ is a full-rank strictly k-parameter exponential family then the natural observation $T(x) = (T_1(x), \ldots, T_k(x))$ is sufficient and complete.*

Theorem 6.5 (Lehman-Scheffé Theorem). *Let T be a sufficient and complete statistic for the statistical model $\mathcal{P}$ and let $\hat{\gamma}$ be an unbiased estimator for $\gamma = g(\theta) \in \mathbb{R}^k$.*

Then $\hat{\gamma}_T = \mathbb{E}_\theta[\hat{\gamma} \mid T]$ is an MVUE for γ. (proof notes: uses contradiction + Rao Blackwell theorem)

Chapter 7: Bayesian Inference: Conjugacy and Improper Priors

Theorem 7.1 (Bayes' Theorem). *Given a* **likelihood** $L(\theta, x)$ *and a* **prior** $\pi(\theta)$ *for* θ*, the* **posterior** *distribution for* θ *(the conditional distribution of* θ *given the data* X*) is given by*

$$\pi(\theta \mid x) = \frac{L(\theta, x)\pi(\theta)}{\int L(\theta', x)\pi(\theta')\,\mathrm{d}\theta'}.$$

(If π *is a mass function replace the integral with a sum.)*

We will often simply write

$$\pi(\theta \mid x) \propto L(\theta, x)\pi(\theta),$$

i.e. **posterior** $\propto$ **likelihood** $\cdot$ **prior**. *The quantity* $p(x) = \int L(\theta', x)\pi(\theta')\,\mathrm{d}\theta'$ *is called the marginal distribution of* X.

7.2 Conjugate priors

Definition 7.2. Consider a model $(L(\theta, x))_{\theta \in \Theta, x \in \mathcal{X}}$. We say that a family of prior distributions $(\pi_\gamma)_{\gamma \in \Gamma}$ is ***conjugate*** if

$$\forall \gamma \in \Gamma, x \in \mathcal{X}\ \exists \gamma(x) \text{ s.t. } \pi_\gamma(\cdot \mid x) = \pi_{\gamma(x)}(\cdot).$$

We say the prior and the posterior are ***conjugate distributions***, and the prior is a ***conjugate prior*** for the likelihood L.

Proposition 7.3 (Conjugate priors for exponential families). *Suppose*

$$L(\theta, x) = h(x)\exp\left\{\sum_{i=1}^{k} \eta_i(\theta)T_i(x) - B(\theta)\right\}$$

defines a k*-parameter exponential family. Then the distributions of the form*

$$\pi_\gamma(\theta) \propto \exp\left\{\gamma_0 B(\theta) + \sum_{i=1}^{k} \gamma_i \eta_i(\theta)\right\}$$

for parameters $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_k)$ *are a conjugate prior family.* (proof exercise)

7.3 Improper priors

Definition 7.4. We say that a pdf/pmf π is an ***improper prior*** if it has infinite mass:

$$\int_\Theta \pi(\theta)\,\mathrm{d}\theta = \infty, \quad \pi(\theta) \geqslant 0\ \forall \theta \in \Theta$$

A posterior distribution $\pi(\theta \mid x)$ can be defined as usual as soon as $\int_\Theta f(x, \theta)\pi(\theta)\,\mathrm{d}\theta < \infty.$

7.4 Predictive Distributions

Definition 7.5. If $X_1, \dots, X_n, X_{n+1}$ are i.i.d. obsevations from the distribution $f(x, \theta)$, with prior $\pi(\theta)$, then the ***posterior predictive distribution*** is

$$f(x_{n+1} \mid x) = \int_\Theta f(x_{n+1}, \theta)\pi(\theta \mid x)\,\mathrm{d}\theta$$

where here $x = (x_1, \dots, x_n)$.

Chapter 8: Non-Informative Priors

8.1 Uniform priors

Definition 8.1. The ***uniform prior*** or ***flat prior*** is the prior $\pi(\theta) \propto 1$.

problem: not flat under reparameterization

8.2 Jeffreys prior

Definition 8.2. ***Jeffrey's prior*** is given, in the one-dimensional case, by $\pi(\theta) \propto \sqrt{I_\theta}$

where $I_\theta = \mathbb{E}_\theta[\frac{\partial^2}{\partial\theta^2}\ell(\theta, x)]$ is the Fisher information. benefit: is invariant under reparameterization

Jeffreys prior in higher dimensions

Definition 8.3. The ***k-dimensional Jeffrey's prior*** is given by

$$\pi(\theta) \propto |I_\theta|^{1/2},$$

where $|I_\theta| = \det I_\theta$ and I_θ is the Fisher information matrix, so under the standard regularity assumptions $(I_\theta)_{ij} = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(\theta, x)\right]$.

8.3 Maximum entropy prior

Definition 8.4. The *entropy* of a pdf/pmf π is defined as $\mathrm{Ent}[\pi] = -\int_\Theta \pi(\theta)\log\pi(\theta)\,\mathrm{d}\theta.$

A maximum entropy probability distribution has entropy that is at least as great as that of all other members of a specified class of probability distributions. According to the principle of maximum entropy, if nothing is known about a distribution except that it belongs to a certain class (usually defined in terms of specified properties or measures), then the distribution with the largest entropy should be chosen as the least-informative default.

Theorem 8.5. *The density $\pi(\theta)$ that maximises $\mathrm{Ent}[\pi]$ subject to $\mathbb{E}[T_j(\theta)] = t_j$ for $j = 1, \dots, p$ takes the p-parameter exponential family form*

$$\pi(\theta) \propto \exp\left[\sum_{i=1}^{p} \lambda_i T_i(\theta)\right] \quad \forall\theta \in \Theta,$$

where $\lambda_1, \dots, \lambda_p$ are determined by the constraints

Chapter 9 Hierarchical Models

Definition 9.1. The building blocks of a *hierarchical Bayesian model* for the observations $Y_1, \dots, Y_n$ with parameters $\theta_1, \dots, \theta_n$ and *hyperparameter* ϕ are

- $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ a family of probability distributions on $\mathcal{A}$. We write $p(y|\theta)$ for the pmf/pdf of P_θ.
- $\{\pi_\phi, \phi \in \Phi\}$ a family of probability distributions on Θ (the parametrized priors). We write $p(\theta|\phi)$ for the pdf/pmf of π_ϕ.
- and P be a distribution on Φ (the *hyperprior distribution*). We write $p(\phi)$ for its pdf/pmf.

Then the corresponding hierarchical model is the following joint distribution of the Y_j, θ_i and ϕ.

I: $y_j|\theta_j, \phi \sim p(y_j|\theta_j)$ independently for each j, (note this does not depend on ϕ)

II: $\theta_j|\phi \sim p(\theta_j|\phi)$

III: $\phi \sim p(\phi)$

The *joint prior* distribution is $p(\theta, \phi) = p(\theta \mid \phi)p(\phi)$ and the *joint posterior* distribution is $p(\theta, \phi \mid y) \propto p(y \mid \theta, \phi)p(\theta, \phi) = p(y \mid \theta)p(\theta \mid \phi)p(\phi)$.

9.3 Exchangeability

Definition 9.2. The distribution of a random vector $\theta = (\theta_1, \dots, \theta_I)$ is *symmetric*, or *exchangeable*, if

$$(\theta_1, \dots, \theta_I) \overset{d}{=} (\theta_{\sigma(1)}, \dots, \theta_{\sigma(I)}) \quad \text{for any permutation } \sigma.$$

Proposition 9.3. *If $\theta = (\theta_1, \dots, \theta_I)$ has (prior) distribution*

$$p(\theta) = \int \left[\prod_{i=1}^{I} \pi(\theta_i \mid \psi)\right] g(\psi)\,\mathrm{d}\psi$$

for some ψ with distribution $g(\psi)$, i.e. the θ_i are conditionally independent given ψ, then the distribution of θ is exchangeable (symmetric). (proof exercise)

Theorem 9.4 (De Finetti). *All exchangeable sequences are of the above form in the large sample limit.*

Chapter 10 Decision Theory

As usual, we will assume a data *model* $X \mid \theta \sim f(x, \theta)$ for some parametric family $\{f(x, \theta) : \theta \in \Theta\}$, where Θ is our *parameter space*.

- An *action (or decision) space* $\mathcal{A}$. Typical examples include $\mathcal{A} = \{0, 1\}$ for selecting a hypothesis, or $\mathcal{A} = g(\Theta)$ for estimating a function $g(\theta)$ of a parameter.
- A *loss function* $L : \Theta \times \mathcal{A} \to \mathbb{R}_+$. Given an action $a \in \mathcal{A}$, if the true parameter is $\theta \in \Theta$ we incur loss $L(\theta, a)$ (don't confuse this with the Likelihood).

- A *set of decision rules* $\mathcal{D} \subseteq \{\delta : \mathcal{X} \to \mathcal{A}\}$. A decision rule δ specifies which action we take given observation $x \in \mathcal{X}$.

Definition 10.1. For a given rule $\delta \in \mathcal{D}$ and parameter $\theta \in \Theta$, the *(frequentist) risk* is

$R(\theta, \delta) = \mathbb{E}_\theta[L(\theta, \delta(X))] = \int_{\mathcal{X}} L(\theta, \delta(x)) f(x, \theta)\, dx.$ This is the expected loss assuming the true parameter is θ.

10.2 Admissibility

Definition 10.2. We say that δ_2 *strictly dominates* δ_1 if $R(\theta, \delta_1) \geqslant R(\theta, \delta_2)\ \forall \theta \in \Theta$

and $R(\theta, \delta_1) > R(\theta, \delta_2)$ for at least some θ. A procedure δ_1 is *inadmissible* if there exists δ_2 such that δ_2 strictly dominates δ_1.

We define *admissible* to simply mean *not inadmissible*.

10.3 Minimax rules and Bayes rules

Definition 10.3. A rule δ is a *minimax rule* if $\sup_\theta R(\theta, \delta) \leqslant \sup_\theta R(\theta, \delta')\ \forall \delta' \in \mathcal{D}$.

It minimises the maximum risk: $\delta^* = \text{argmin}_{\delta \in \mathcal{D}} \sup_{\theta \in \Theta} R(\theta, \delta)$.

Definition 10.4. The *Bayes integrated risk* (or simply *Bayes risk*)for a decision rule δ and a prior $\pi(\theta)$ is

$$r(\pi, \delta) := \int_\Theta R(\theta, \delta)\pi(\theta)\, d\theta.$$

A decision rule δ is said to be a *Bayes rule* w.r.t. π if it minimises the Bayes risk: $r(\pi, \delta) = \inf_{\delta' \in \mathcal{D}} r(\pi, \delta') =: r_\pi$.

Definition 10.5. A prior distribution π is least favorable if $r_\pi \geqslant r_{\pi'}$ for all prior distrbutions π'.

Theorem 10.6. *Suppose that π is a prior distribution on Θ and that δ_{Bayes} is the Bayes estimator for π with*

$$r(\pi, \delta_{Bayes}) = r_\pi.$$

If δ_0 is a rule such that $\sup_\theta R(\theta, \delta_0) \leqslant r_\pi$

then δ_0 is minimax, and, furthermore, if δ_{Bayes} is the unique Bayes estimator for π then δ_0 is the unique minimax procedure.

Proof. Let δ be any other rule. Then $\sup_\theta R(\theta, \delta) \geqslant \int R(\theta, \delta)\pi(\theta) d\theta \geqslant \int R(\theta, \delta_{\text{Bayes}})\pi(\theta) d\theta = r_\pi \geqslant \sup_\theta R(\theta, \delta_0)$.

The second inequality is strict if there is a unique Bayes estimator which gives the second point.

Theorem 10.7. *Let δ_{Bayes} be the Bayes estimator for some prior π. If* $R(\theta, \delta_{Bayes}) \leqslant r_\pi$ *for all θ then δ_{Bayes} is minimax and π is a least favorable prior.*

Proof. The first part is simply an application of Theorem 10.6.

Let π' be some other distribution. Then, writing δ'_{Bayes} for the Bayes estimator with respect to π' we have

$r_{\pi'} = \int R(\theta, \delta'_{\text{Bayes}})\pi'(\theta) d\theta \leqslant \int R(\theta, \delta_{\text{Bayes}})\pi'(\theta) d\theta \leqslant \sup_\theta R(\theta, \delta_{\text{Bayes}}) = r_\pi.$

Corollary 10.8. *If a Bayes rule δ_{Bayes} has constant Risk, then it is minimax.* <u>very useful</u>

Corollary 10.9. *Let $\omega_\pi \subset \Theta$ be the set of θ at which the risk function of δ_{Bayes} achieves its maximum, i.e.*

$$\omega_\pi = \{\theta : R(\theta, \delta_{Bayes}) = \sup_{\theta'} R(\theta', \delta_{Bayes}\}.$$

Then δ_{Bayes} is minimax if and only if

$$\pi(\omega_\pi) = 1.$$

10.4 Bayes rule and posterior risk

Definition 10.10. The *expected posterior loss* of a rule δ w.r.t. a prior π is

$$\Lambda(x,\delta) = \mathbb{E}\left[L[\theta,\delta(x)] \mid X = x\right] = \int_\Theta L(\theta,\delta(x))\pi(\theta \mid x)\,d\theta.$$

Theorem 10.11. *Suppose that $X \mid \theta \sim P_\theta$ and that $\theta \sim \pi$. Suppose in addition that the following hypothesis hold for the problem of estimating $g(\theta)$ with non-negative loss function $L(\theta, d)$.*

(a) There exists an estimator (a rule) δ_0 with finite risk. (b) For almost all x, there exists a value $c(x)$ which minimizes $y \mapsto \Lambda(x,y)$ Then $\delta(x) = c(x)$ is a Bayes estimator. (proof notes)

Proposition 10.12 (Bayes rules and admissibility). *Let δ^π be a Bayes rule w.r.t. π with finite Bayes risk. Then*

1. If δ^π is unique then it is admissible 2. If $\theta \mapsto R(\theta,\delta)$ is continuous for all δ and π has a positive density w.r.t. the Lebesgue measure, then δ^π is admissible.

1. If δ^π is not admissible then there is some δ such that $R(\theta,\delta) \leqslant R(\theta,\delta^\pi)$ $\forall\theta\in\Theta$ and $R(\theta,\delta) < R(\theta,\delta^\pi)$ for some θ. This implies $r(\pi,\delta) \leqslant r(\pi,\delta^\pi)$, so δ must also be Bayes, so by uniqueness $\delta = \delta^\pi$, contradicting the definition of δ. So δ^π is admissible.

2. As above, if δ^π is not admissible then there is some δ such that $R(\theta,\delta) \leqslant R(\theta,\delta^\pi)$ $\forall\theta\in\Theta$ and $A_\delta \neq \emptyset$, where $A_\delta := \{\theta : R(\theta,\delta) < R(\theta,\delta^\pi)\}$.

   Since $\theta \mapsto R(\theta,\delta) - R(\theta,\delta^\pi)$ is continuous, A_δ must contain an open set. So $\pi(A_\delta) > 0$. A contradiction!

10.5 Point estimation

Definition 10.13. The *zero-one loss* is of the form $L(\theta,\hat\theta) = \begin{cases} a & \text{if } |\theta-\hat\theta| > b, \\ 0 & \text{otherwise} \end{cases}$ where a, b are positive constants.

The *absolute error loss* is of the form $L(\theta,\hat\theta) = k|\hat\theta - \theta|$ where k is a positive constant.

The *quadratic loss* is of the form $L(\hat\theta,\theta) = k(\hat\theta-\theta)^2$ where k is a positive constant.

Proposition 10.14. *The Bayes estimate under the:*

1. zero-one loss with interval radius b tends to the posterior mode as $b \to 0$;

2. absolute error loss is the posterior median; 3. quadratic loss is the posterior mean. (proof notes)

10.6 Finite decision problems

Definition 10.15. A decision problem is said to be finite when Θ is finite. We write $\Theta = (\theta_1,\dots,\theta_k)$.

Definition 10.16. The *risk set* $S \subseteq \mathbb{R}^k$ is the set of points $\{(R(\theta_1,\delta),\dots,R(\theta_k,\delta)) : \delta \in \mathcal{D}\}$.

Lemma 10.17. *S is a convex set.*

Proof. Let $\delta_1, \delta_2 \in \mathcal{D}$ be two rules. Take $\alpha \in (0,1)$. Then define a randomized rule as follows:

$$\delta'(x) = \begin{cases} \delta_1(x) & \text{with prob } \alpha, \\ \delta_2(x) & \text{with prob } 1-\alpha. \end{cases}$$

Then $R(\theta,\delta') = \alpha R(\theta,\delta_1) + (1-\alpha)R(\theta,\delta_2)$. So the convex combination is a valid decision rule. □

Chapter 11: The James-Stein Estimator

Theorem 11.1 (Stein's Paradox). *The* **James-Stein estimator** $\hat\mu_{\text{JSE}} := \left(1 - \frac{p-2}{\sum_{i=1}^p X_i^2}\right) X$ *strictly dominates $\hat\mu_{\text{MLE}}$ for quadratic loss.* **Corollary 11.2.** *If $p \geqslant 3$, $\hat\mu_{\text{MLE}}$ is inadmissible for quadratic loss.*

Remark. This is *very surprising*! For instance, suppose you take measurements to estimate:

1. The average weight K of a kiwi at Tesco;
2. The average height G of a blade of grass in University Parks;
3. The average speed S of a bike going down Cornmarket Street.

These are totally unrelated quantities; but Stein's paradox tells us that we get better estimates (on average) for the vector (K, G, S) by simultaneously using the three measurements![1]

Lemma 11.3 (Stein's Lemma). *For independent Gaussian random variables* $X = (X_1, \ldots, X_p)$ *with* $X_i \sim \mathcal{N}(\mu_i, 1)$ *for each* i*, then for each* i *and for any bounded differentiable function* h*,*

$$\mathbb{E}[(X_i - \mu_i)h(X)] = \mathbb{E}\left[\frac{\partial h(X)}{\partial X_i}\right].$$

(proof of this uses Tower Law, proof Stein's Paradox also in notes)

Proof. By the Tower Law, $\mathbb{E}[(X_i - \mu_i)h(X)] = \mathbb{E}\left[\mathbb{E}[(X_i - \mu_i)h(X) \mid \{X_j : j \neq i\}]\right]$ Using integration by parts, S

$$\mathbb{E}[(X_i - \mu_i)h(X) \mid \{X_j : j \neq i\}] = \int_{-\infty}^{\infty} (x_i - \mu_i)h(x)e^{-(x_i-\mu_i)^2/2}\,dx_i = \left[-e^{-(x_i-\mu_i)^2/2}h(x)\right]_{x_i=-\infty}^{x_i=\infty} + \int_{-\infty}^{\infty} \frac{\partial h(x)}{\partial x_i} e^{-(x_i-\mu_i)^2/2}\,dx_i$$

$$= 0 + \mathbb{E}\left[\frac{\partial h(X)}{\partial X_i} \mid X_j : j \neq i\right]$$

since h is bounded. Applying the tower property of conditional expectations again gives the result. □

Proof of Stein's Paradox. Consider the family of estimators $\hat{\mu}_{\mathrm{JSE}} = \left(1 - \frac{a}{\sum X_i^2}\right)X$ indexed by the parameter a. These are called the ***James-Stein estimators***.

Recalling that $\hat{\mu}_{\mathrm{MLE}} = X$, we get $R(\mu, \hat{\mu}_{\mathrm{MLE}}) = \sum_{i=1}^{p} \mathbb{E}[(\mu_i - X_i)^2] = p$ (since $\mathrm{Var}(X_i) = 1$).

On the other hand, writing $\hat{\mu}_i := \left(1 - \frac{a}{\sum_j X_j^2}\right)X_i$, $R(\mu, \hat{\mu}_{\mathrm{JSE}}) = \sum_{i=1}^{p} \mathbb{E}[(\mu_i - \hat{\mu}_i)^2]$

$$= \sum_{i=1}^{p}\left(\mathbb{E}[(\mu_i - X_i)^2] - 2a\,\mathbb{E}\left[\frac{(X_i - \mu_i)X_i}{\sum_j X_j^2}\right] + a^2\,\mathbb{E}\left[\frac{X_i^2}{\left(\sum_j X_j^2\right)^2}\right]\right)$$

Now the first term is just 1, since $\mathrm{Var}(X_i) = 1$, and by Stein's Lemma,

$$\mathbb{E}\left[\frac{(X_i - \mu_i)X_i}{\sum_j X_j^2}\right] = \mathbb{E}\left[\frac{\partial}{\partial X_i}\frac{X_i}{\sum_j X_j^2}\right] = \mathbb{E}\left[\frac{\sum_j X_j^2 - 2X_i^2}{\left(\sum_j X_j^2\right)^2}\right] = \mathbb{E}\left[\frac{1}{\sum_j X_j^2} - 2\frac{X_i^2}{\left(\sum_j X_j^2\right)^2}\right]$$

Putting this all together, we get $R(\mu, \hat{\mu}_{\mathrm{JSE}}) = p - (2ap - 4a)\,\mathbb{E}\left[\frac{1}{\sum X_j^2}\right] + a^2\,\mathbb{E}\left[\frac{1}{\sum X_j^2}\right] = p - (2a(p-2) - a^2)\,\mathbb{E}\left[\frac{1}{\sum X_j^2}\right]$

This is minimised at $a = p - 2$, and is less than p for this value; this concludes the proof.

Chapter 12: Empirical Bayes Methods

Definition 12.1. ***Empirical Bayes*** methods adapt the hierarchical Bayesian model by replacing the hyperparameter vector ψ with a point-estimate $\hat{\psi}$ derived from the data.

So we now just have the likelihood $X \sim f(x, \theta)$ and the prior $\theta \sim \hat{\psi}(\theta) = \pi(\theta, \hat{\psi})$.

The reduced model has posterior

$$\hat{\pi}(\theta \mid x) \alpha L(\theta, x)\pi(\theta, \hat{\psi})$$

and a ***Bayes estimator*** $\hat{\theta}_{\mathrm{EB}}$ can be calculated using $\hat{\pi}(\theta \mid x)$. So for quadratic loss, we have $\hat{\theta}_{\mathrm{EB}} = \int \theta\hat{\pi}(\theta \mid x)\,d\theta$, the posterior mean.

Remark. In this setting, the Bayes estimator is called an ***empirical Bayes estimator***, or an ***EB estimator***.

12.2 Choice of point estimate

- Use the MLE $\hat{\psi} = \mathrm{argmax}_{\psi}\, p(x \mid \psi)$ where $p(x \mid \psi) = \int L(\theta, x)\pi(\theta, \psi)\,d\theta$ is the marginal likelihood.
- Use the method of moments: choose $\hat{\psi}$ such that $\pi(\theta, \hat{\psi})$ has the same mean and variance as the *sample mean* and *sample variance* of the MLEs of the θ_i.

12.3 James-Stein and empirical Bayes

Proposition 12.2. *The James-Stein estimator can be intepreted as an empirical Bayes estimator.*

(Specifically, for $a = p$ it's the EB estimator for quadratic loss when using a mean-zero Gaussian prior whose variance is estimated using maximum likelihood.)

Proof. We wish to construct an EB estimator for quadratic loss. There is some freedom of choice of prior, but we will assume as our prior that θ_i are drawn independently from a $\mathcal{N}(0, \tau^2)$ distribution.

Given τ, then, we have $\theta_i \mid (x_i, \tau^2) \sim \mathcal{N}\left(x_i \frac{\tau^2}{1+\tau^2}, \frac{\tau^2}{1+\tau^2}\right)$. This can be calculated by completing the square.

To estimate τ, then, we can compute the marginal likelihood of X_i given τ:

$$X_i \mid \tau^2 \sim \mathcal{N}(0, \tau^2 + 1) \text{ independently for each } i.$$

This is maximised by $\hat{\tau}^2 = \frac{1}{p}\sum_{j=1}^p (X_j^2 - 1)$. (This is from the standard result for the MLE for the variance of a Gaussian distribution).

So the estimated posterior distribution is $\theta_i \mid x_i \sim \mathcal{N}\left(x_i \frac{\hat{\tau}^2}{1+\hat{\tau}^2}, \frac{\hat{\tau}^2}{1+\hat{\tau}^2}\right)$. Thus the Bayes estimator for quadratic loss, i.e. the posterior mean, is

$$\hat{\theta}_{\mathrm{EB},i} = X_i \frac{\hat{\tau}^2}{1+\hat{\tau}^2} = X_i \frac{\left(\frac{1}{p}\sum_{j=1}^p X_j^2\right) - 1}{\frac{1}{p}\sum_{j=1}^p X_j^2} = X_i \left(1 - \frac{p}{\sum X_j^2}\right).$$

12.4 Non-parametric empirical Bayes

So far we have estimated a hyperprior distribution by finding a point estimate for the hyperparameter. We could instead estimate the hyperprior (or marginal) distribution *directly* from the data. This is known as ***non-parametric empirical Bayes***. One such method is illustrated below.

Chapter 13: Hypothesis Tests

Let $X_1, \ldots, X_n$ be a random sample from $f(x;\theta)$ where $\theta \in \Theta$ is a scalar or vector parameter. Suppose we are interested in testing

The null hypotehsis $H_0 : \theta \in \Theta_0$ against the alternative $H_1 : \theta \in \Theta_1$. Unless specified otherwise we assume that $\Theta_0 \cap \Theta_1 = \emptyset$

If a hypothesis consists of a single point in Θ so that $\Theta_0 = \{\theta_0\}$ say, we say that it is a ***simple*** hypotehsis Otherwise it is called a ***composite*** hypothesis.

In general a test consists of a ***critical region*** C such that we reject H_0 if and only if $X \in C$. We reformulate this slightly by introducing the concept of the ***test function*** $\phi : \mathcal{X} \mapsto \{0,1\}$

$$\phi(x) = \begin{cases} 1 \text{ if } x \in C \\ 0 \text{ if } x \notin C \end{cases} \qquad \phi(x) = \begin{cases} 1 \text{ if } x \in C_1 \\ \gamma \text{ if } x \in C_= \\ 0 \text{ if } x \in C_0 \end{cases}$$

We will sometimes simply say ***the test*** ϕ. We will also sometimes need the notion of a randomized test. Suppose that $\mathcal{X} = C_1 \cup C_0 \cup C_=$ where $C_1, C_0, C_=$ are pairwise disjoint. Fix $\gamma \in [0,1]$. Then the we generalize the notion of test function by saying that

is the test where we ***reject*** H_0 when $x \in C_1$, ***accept*** H_0 when $x \in C_0$, and ***reject*** H_0 ***with probability*** γ if $x \in C_=$ (by flipping a coin). Such a test ϕ is called a ***randomized test***.

Definition 13.1. • The ***power function*** of a test is defined to be $w(\theta) = \mathbb{P}_\theta(\text{Reject}H_0) = \mathbb{E}_\theta[\phi(X)]$.

• The ***size*** of a test is often denoted α and is defined to be $\alpha := \sup_{\theta \in \Theta_0} w(\theta).$

Within this framework we can consider various classes of problems: 1. Simple H_0 vs simple H_1

2. Simple H_0 vs composite H_1: 3. Composite H_0 vs composite H_1:

13.1.2 Neyman-Pearson Theorem

Consider a test of a simple null hypothesis $H_0 : \theta = \theta_0$ agains a simple alternative $H_1 : \theta = \theta_1$. Define the ***likelihood ratio***:

$$\Lambda(x) = \frac{f(x, \theta_1)}{f(x, \theta_0)}.$$

Theorem 13.2. *Define the critical region*

$$C = \{x : \Lambda(x) \geqslant k\}$$

and suppose that the constants k and α are such that $\mathbb{P}_{\theta_0}(X \in C) = \alpha$. Then among all tests of H_0 against H_1 of size α, the test with critical region C has **maximum power**.

The tests with critical regions such as C are called ***Neyman-Pearson test*** or ***likelihood ratio test*** (LRT).

13.1.3 Uniformly most powerful tests

Definition 13.3. A ***uniformly most powerfull test*** or UMP test of size α is a test function ϕ_0 such that

1. $\mathbb{E}_\theta(\phi_0(X)) \leqslant \alpha$ for all $\theta \in \Theta_0$,
2. Given any other test ϕ for which $\mathbb{E}_\theta(\phi(X)) \leqslant \alpha$ for all $\theta \in \Theta_0$, we have $\mathbb{E}_\theta(\phi_0(X)) \geqslant \mathbb{E}_\theta(\phi(X))$ for all $\theta \in \Theta_1$.

Definition 13.4. A family of densities $\{f(x, \theta), \theta \in \Theta \subseteq \mathbb{R}\}$ with real scalar variable x is said to be of ***monotone likelihood ratio*** or MLR for short if there exists a function $t(x)$ such that the likelihood ratio

$$x \mapsto \frac{f(x, \theta_2)}{f(x, \theta_1)}$$

is a non-decreasing function of $t(x)$ whenever $\theta_1 \leqslant \theta_2$.

Theorem 13.5. *Suppose that X has a distribution from a family which is MLR with respect to a statistic $t(X)$ and that we wish to test $H_0 : \theta \leqslant \theta_0$ against $H_1 : \theta > \theta_0$. Suppose that the distribution of $t(X)$ is continuous. Then*

1. *The test with critical region*
$$C = \{x : t(x) > t_0\}$$
is UMP among all test of size at most $\mathbb{P}_{\theta_0}(X \in C)$.
2. *Given α, there exists some t_0 such that the test above has size α.*

Proof. For any $\theta_1 > \theta_0$ the Neyman Pearson test of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ has a critical region of the form $C = \{x : t(x) > t_0\}$ for some t_0 which is chosen so that $\mathbb{P}_{\theta_0}(T(X) > t_0) = \alpha$. Note that t_0 does not depend on θ_1 and so the critical region C is the same for all values of θ_1. Thus, we see that this test is UMP for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$.

Next, we claim that for any critical region of the form $C = \{x : t(x) > t_0\}$ the map

$$\theta \mapsto \mathbb{P}_\theta(X \in C)$$

is non-decreasing. This can be seen using a argument involving randomized test procedures and the optimality of the LRT (see Young and Smith p72).

It follows that if $\mathbb{P}_{\theta_0}(X \in C) = \alpha$ then $\sup_{\theta \leqslant \theta_0} \mathbb{P}_\theta(X \in C) \leqslant \alpha$. Suppose that C' is another critical region such that $\sup_{\theta \leqslant \theta_0} \mathbb{P}_\theta(X \in C') \leqslant \alpha$ as well. This implies trivially that $\mathbb{P}_{\theta_0}(X \in C') \leqslant \alpha$ and thus by optimality of the LRT that for all $\theta_1 > \theta_0$ we have

$$\mathbb{P}_{\theta_1}(X \in C') \leqslant \mathbb{P}_{\theta_1}(X \in C)$$

This shows that C is UMP among all tests of its size.

The second statement in the Theorem is clear by continuity. $\square$

13.2 Bayes factors

13.2.1 Bayes factors for simple hypotheses

Bayes' rule tells us that (writing f_i for the density of X under H_i)

$$\mathbb{P}(H_0 \text{ is true} \mid X_x) = \frac{\pi_0 f_0(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)}$$

which can also be written as

$$\frac{\mathbb{P}(H_0 \text{ is true} \mid X = x)}{\mathbb{P}(H_1 \text{ is true} \mid X_x)} = \frac{\pi_0}{\pi_1}\frac{f_0(x)}{f_1(x)}.$$

posterior odds = prior odds $\times$ Bayes factor.

Definition 13.6. We call $\frac{\pi_0}{\pi_1}$ the ***prior odds*** in favor of H_0 and $B = \frac{f_0(x)}{f_1(x)}$ is the ***Bayes factor***.

13.2.2 Bayes factors for composite hypothesis

Definition 13.7. The ***Bayes factor*** in the composite-composite case is defined to be

$$B = \frac{\int_{\Theta_0} f(x,\theta)g_0(\theta)\,\mathrm{d}\theta}{\int_{\Theta_1} f(x,\theta)g_1(\theta)\,\mathrm{d}\theta}.$$

The ***Bayes factor*** in the simple-composite case is defined to be

$$B = \frac{f(x,\theta_0)}{\int_{\Theta_1} f(x,\theta)g_1(\theta)\,\mathrm{d}\theta}.$$

More generally, there is nothing here that requires the same parametrization under the two hypothesis. Suppose that we have two candidate parametric models M_1 and M_2 for data X, and the two models have respective parameter vectors θ_1 and θ_2. Under prior densities $\pi_1(\theta_1)$ and $\pi_2(\theta_2)$, the marginal distribution for X under each models are found as

$$p(x \mid M_i) = \int f(x, \theta_i, M_i)\pi_i(\theta_i)\,\mathrm{d}\theta_i$$

and the ***Bayes factor*** is just their ratio

$$B = \frac{p(x \mid M_1)}{p(x \mid M_2)}.$$

Note that form this point of view, what we have is really a hierarchical Bayesian model where where the model correspond to the hyperparameter.

13.3 Hypothesis testing in the context of decision theory

Suppose we wish to test the hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta = \theta_1$ and consider the (non-random) test ϕ with critical region C

$$\phi(x) = \begin{cases} 1 \text{ if } x \in C \\ 0 \text{ if } x \notin C \end{cases}$$

A generic loss function can be written:

$$L(\theta, \phi(x)) = \begin{cases} a\phi(x) & \text{if } \theta = \theta_0 \\ b(1-\phi(x)) & \text{if } \theta = \theta_1. \end{cases}$$

Lemma 13.8. *The rule ϕ has risk $R(\theta_0, \phi) = a\alpha$ and $R(\theta_1, \phi) = b\beta$ where $\beta = 1 - w(\theta_1)$.*

Proof. We have $R(\theta_0, \phi) = \mathbb{E}_{\theta_0}[a\phi(X)] = a\alpha$ $\quad R(\theta_1, \phi) = \mathbb{E}_{\theta_1}[b(1-\phi(X))] = b(1 - w(\theta_1))$. ☐

Lemma 13.9. *The Bayes risk for ϕ under the prior π is*

$$r(\pi, \delta_C) = p_0 a\alpha(C) + p_1 b\beta(C).$$

proof trivial, get expected risk

Definition 13.10. The ***Bayes test*** is the rule δ_C with the critical region C chosen to minimise the Bayes risk (under the loss function defined above).

Theorem 13.11 (Bayes test for simple hypotheses). *The critical region for the Bayes test with prior π and loss L is*

$$C = \left\{x : \frac{f(x,\theta_1)}{f(x,\theta_0)} \geqslant A\right\}$$

where $A = \frac{p_0 a}{p_1 b}$.

(proof notes)

Corollary 13.12. *The Bayes test is a likelihood ratio test with $A = \frac{p_0 a}{p_1 b}$.*

Corollary 13.13. *Every likelihood ratio test is a Bayes test for some prior probabilities p_0, p_1.*

13.3.2 The case of the 0–1 loss function

In the case that L is the 0–1 loss, so $a = b = 1$ and

$$L(\theta, \delta_C(x)) = \begin{cases} 1 & \text{if } \theta = \theta_0 \text{ and } x \in C, \\ 1 & \text{if } \theta = \theta_1 \text{ and } x \notin C, \\ 0 & \text{otherwise,} \end{cases}$$

Definition 13.14. The ***maximum a posteriori (MAP) test*** chooses the hypothesis with the highest posterior probability $\mathbb{P}(H_i \mid X = x)$.

Theorem 13.15. *The MAP test is the Bayes test under the 0–1 loss.* (proof exercise)

Proposition 13.16. *The Bayes test for the 0–1 loss (i.e. the MAP test) rejects H_0 iff*

$$\frac{f(x, \theta_0)}{\int_{\Theta_1} f(x, \theta) g_1(\theta)\, \mathrm{d}\theta} < \frac{\pi_1}{\pi_0}.$$

(application of Thm 13.11 with a=b=1, need only check that it's a MAP test)

13.5 Two sided hypothesis tests

We now consider in more details situations in which $H_0 : \theta \in \Theta_0$ is either $\Theta_0 = [\theta_1, \theta_2]$ or $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \mathbb{R} \setminus \Theta_0$. In this situation we cannot expect to find a UMP test, even for nice families such as exponentials or MLR. The reason is obvious: if we construct a Neyman–Pearson test of say $\theta = \theta_0$ against $\theta = \theta_1$ for some $\theta_1 \neq \theta_0$, the test takes quite a different form when $\theta_1 > \theta_0$ from when $\theta_1 < \theta_0$. We simply cannot expect one test to be most powerful in both cases simultaneously. However, if we have an exponential family with natural statistic $T = t(X)$, or a family with MLR with respect to $t(X)$, we might still expect tests of the form

$$\phi(x) = \begin{cases} 1 \text{ if } x \in t(x) \notin [t_1, t_2] \\ \gamma(x) \text{ if } t(x) = t_1 \text{ or } t_2 \\ 0 \text{ if } x \in (t_1, t_2). \end{cases}$$

where $t_1 < t_2$ to have good properties. Such tests are called ***two sided tests*** based on T.

Definition 13.17. A test of $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ is called ***unbiased*** of size α if

$$\mathbb{P}_\theta(X \in C) \leqslant \alpha \quad \forall \theta \in \Theta_0 \qquad \text{but} \qquad \mathbb{P}_\theta(X \in C) \geqslant \alpha \quad \forall \theta \in \Theta_1.$$

A test which is uniformly most powerful amongst the class of all unbiased tests is called ***uniformly most powerful unbiased***, abbreviated UMPU.

13.5.1 UMPU tests for one-parameter exponential families

Consider an exponential family of the form

$$f(x, \theta) = h(x) \exp\{\theta t(x) - B(\theta)\}$$

with $\theta \in \mathbb{R}$. Let $T = t(X)$ be the natural observation.

Remember that T itself also belongs to an exponential family with density form

$$f_T(t, \theta) = h_T(t) \exp\{\theta t - B(\theta)\}.$$

We shall assume that T is a continuous random variable with $h_T > 0$ on the open set that defines the range of T. This avoids the need for randomised tests and this makes our proofs less technical at the cost of very little loss of generality.

Theorem 13.18. *For any α there exists a UMPU test of size α which is of the two-sided form in T.*

need following Lemmas

Lemma 13.19. *Let $f_0, f_1, \dots, f_m$ be $m+1$ probability densities, and let $\alpha_1, \dots, \alpha_m$ be constants such that the class $\mathcal{C}$*

$$\mathcal{C} = \{\phi : \int \phi(x) f_i(x)\,\mathrm{d}x = \alpha_i,\ i = 1, \dots, m\}$$

is non-empty. Then

1. *There is one member of $\mathcal{C}$ that maximizes $\int f_0(x)\phi(x)\,\mathrm{d}x >$*
2. *A necessary and sufficient condition for $\phi^* \in \mathcal{C}$ to be a maximizer is that there exists constants $k_1, \dots, k_m$*

$$\phi(x) = \begin{cases} 1 \text{ if } f_0(x) > \sum_{i=1}^m k_i f_i(x) \\ 0 \text{ if } f_0(x) < \sum_{i=1}^m k_i f_i(x) \end{cases} . \tag{13.1}$$

3. *If $\phi \in \mathcal{C}$ satisfies (13.1) with $k_1, \dots, k_m \geqslant 0$ then it maximises $\int f_0(x)\phi(x)\,\mathrm{d}x$ among all functions satisfying*

$$\int \phi(x) f_i(x)\,\mathrm{d}x \leqslant \alpha_i,\ i = 1, \dots, m$$