

Additional Notes on Parameter Estimation

These notes summarize some material presented in class, including some examples.

Consider a family $\{p_\theta(y), \theta \in \Lambda, y \in \Gamma\}$ of pdf's and an estimator $\hat{\theta}(y)$, viewed as a mapping from the parameter space Λ to the observation space Γ . The expectation of a function $f(y)$ with respect to p_θ is denoted by

$$\mathbb{E}_\theta[f(Y)] = \int_\Gamma f(y) p_\theta(y) dy.$$

Consider a cost function $C(\hat{\theta}, \theta)$, viewed as a mapping from $\Lambda \times \Lambda$ to Γ . The conditional risk for the estimator $\hat{\theta}(Y)$ is

$$R_\theta(\hat{\theta}) = \mathbb{E}_\theta[C(\hat{\theta}(Y), \theta)] = \int_\Gamma C(\hat{\theta}(y), \theta) p_\theta(y) dy.$$

1 Minimum Variance Unbiased Estimators

In the absence of a prior distribution on θ , there exists generally no estimator that minimizes $R_\theta(\hat{\theta})$ uniformly over all $\theta \in \Lambda$. For instance, the trivial estimator $\hat{\theta}(Y) = \theta_0$ (for some arbitrary $\theta_0 \in \Lambda$) is perfect when $\theta = \theta_0$ but is poor otherwise.

To avoid such pathological situations, one may require that the estimator be **unbiased**:

$$\mathbb{E}_\theta[\hat{\theta}(Y)] = \theta, \quad \forall \theta \in \Lambda.$$

If the cost function is squared-error, $C(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$, then the conditional risk is mean-squared error. For an unbiased estimator, we obtain

$$R_\theta(\hat{\theta}) = \int_\Gamma (\hat{\theta}(y) - \theta)^2 p_\theta(y) dy = \text{MSE}_\theta(\hat{\theta}) = \text{Var}_\theta(\hat{\theta}).$$

A **Minimum Variance Unbiased Estimator** (MVUE) is an estimator that is unbiased and minimizes $\text{Var}_\theta(\hat{\theta})$ uniformly over all $\theta \in \Lambda$. MVUE estimators exist for some parametric families. However, even when they exist, they may be hard to find.

We next introduce some tools that will be used to derive MVUEs (when they exist!)

2 Sufficient Statistics

A function $T : \Gamma \rightarrow \mathcal{T}$ is a **sufficient statistic** for the family $\{p_\theta, \theta \in \Lambda\}$ if the distribution of Y conditioned on $T(Y)$ does not depend on θ . One may say more briefly that “ T is

sufficient for θ .” As we shall see, a sufficient statistic “compresses” the data in a lossless way, i.e., in a way that does not lose information for solving the estimation problem.

A function $T : \Gamma \rightarrow \mathcal{T}$ is a **minimal sufficient statistic** for the family $\{p_\theta, \theta \in \Lambda\}$ if it is also a function of *every other sufficient statistic* for θ .

Example. Consider n iid Bernoulli random variables with parameter θ . The probability of a binary sequence $\underline{y} \in \{0, 1\}^n$ is

$$p_\theta(\underline{y}) = \theta^{n_1}(1 - \theta)^{n - n_1},$$

where $n_1(\underline{y}) = \sum_{i=1}^n 1_{\{y_i=1\}}$ is the number of 1’s in the sequence.

We see that n_1 is sufficient for θ because the distribution of \underline{Y} conditioned on $N_1 = n_1$ is uniform (does not depend on θ) over the set of $\binom{n}{n_1} = \frac{n!}{n_1!(n - n_1)!}$ sequences that have n_1 1’s and $n - n_1$ 0’s. It may also be formally shown that n_1 is a minimal sufficient statistic for θ .

Existence of sufficient statistics. The trivial choice $T(Y) = Y$ is a sufficient statistic. The more interesting question is whether nontrivial sufficient statistics can be found. In the Bernoulli example above, n_1 was a sufficient statistic – in fact, a minimal sufficient statistic. That is significant compression, because there are only $n + 1$ possible values for n_1 , while there are 2^n possible values for \underline{y} . Once a sufficient statistic $T(Y)$ is identified, any one-to-one transformation of $T(Y)$ is also a sufficient statistic. In the above example, $n - n_1$ is a sufficient statistic, and so is $5n_1$, etc. In general, minimal sufficient statistics don’t always exist, and when they do, they may be hard to find.

Factorization Theorem. Denote by q_θ the distribution of the sufficient statistic $T(Y)$ induced by T and p_θ . We may factor $p_\theta(y)$ as

$$\begin{aligned} p_\theta(y) &= p(y|T(y), \theta) q_\theta(T(y)) \\ &= p(y|T(y)) q_\theta(T(y)) \end{aligned} \tag{1}$$

where by definition of a sufficient statistic we have dropped θ from the conditioning in the last line. So $p_\theta(y)$ factors in the form

$$p_\theta(y) = h(y) g_\theta(T(y)). \tag{2}$$

It turns out that this factorization is necessary and sufficient for T to be sufficient for θ . (The word “sufficient” is used twice but with different meanings in the above sentence!) The formula (2) is slightly more general than (1) in that neither $h(y)$ nor $g_\theta(T(y))$ is required to be a pdf.

Example. For the Bernoulli example above, we have $T(y) = n_1$ and

$$\begin{aligned} \underbrace{p(y|T(Y) = n_1)}_{\text{Uniform}} &= 1_{\{T(y)=n_1\}} / \binom{n}{n_1}, \\ \underbrace{q_\theta(n_1)}_{\text{Binomial}} &= \binom{n}{n_1} \theta^{n_1} (1 - \theta)^{n - n_1} \end{aligned}$$

3 Rao-Blackwell Theorem

Rao-Blackwell Theorem. Consider an *arbitrary* unbiased estimator $\hat{g}(Y)$ of $g(\theta)$ and a sufficient statistic $T(Y)$ for θ . Define the following estimator of $g(\theta)$ based on $T(Y)$:

$$\tilde{g}(T(y)) = \mathbb{E}_\theta[\hat{g}(Y) \mid T(Y) = T(y)]. \quad (3)$$

Then we have

- (a) $\tilde{g}(T(y))$ is also an unbiased estimator of θ , and
- (b) $\text{Var}_\theta[\tilde{g}(T(Y))] \leq \text{Var}_\theta[\hat{g}(Y)]$ with equality iff $\tilde{g}(T(Y)) = \hat{g}(Y)$ a.s. p_θ .

Thus “averaging” the initial estimator $\hat{g}(Y)$ by conditioning on a sufficient statistic can only reduce the variance the estimator. It is remarkable that this Rao-Blackwellization procedure often works well even if the initial estimator $\hat{g}(Y)$ is a poor one, as shown by the example below.

Example. Returning to the Bernoulli example, let $g(\theta) = \theta$, and consider the poor (but unbiased) estimator $\hat{g}(Y) = Y_1$, i.e., the estimator discards all samples of the observed sequence, except for the first one. Applying (3) with $T(y) = n_1$, we obtain

$$\tilde{g}(n_1) = \mathbb{E}_\theta \left[Y_1 \mid \sum_{i=1}^n Y_i = n_1 \right]$$

By symmetry, we have

$$\mathbb{E}_\theta \left[Y_1 \mid \sum_{i=1}^n Y_i = n_1 \right] = \mathbb{E}_\theta \left[Y_i \mid \sum_{i=1}^n Y_i = n_1 \right] \quad \forall i = 1, 2, \dots, n$$

whence

$$\tilde{g}(n_1) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\theta \left[Y_i \mid \sum_{i=1}^n Y_i = n_1 \right] = \frac{1}{n} \mathbb{E}_\theta \left[\sum_{i=1}^n Y_i \mid \sum_{i=1}^n Y_i = n_1 \right] = \frac{n_1}{n}.$$

The variance of the initial estimator $\hat{g}(Y) = Y_1$ is $\theta(1-\theta)$, independently of n . The variance of the Rao-Blackwellized estimator $\tilde{g}(N_1)$ is only $\frac{\theta(1-\theta)}{n}$.

Note that the Rao-Blackwellized estimator $\tilde{g}(T(Y))$ may depend on the choice of the initial estimator $\hat{g}(Y)$. This was not the case in the example above (for reasons that are given below), but consider the trivial sufficient statistic $T(Y) = Y$. Applying (3), we obtain

$$\tilde{g}(y) = \mathbb{E}_\theta[\hat{g}(Y) \mid Y = y] = \hat{g}(y),$$

i.e, Rao-Blackwellization trivially reproduces the initial estimator.

Corollary to Rao-Blackwell Theorem. If T is sufficient for θ and $g^*(T(Y))$ is the only function of $T(Y)$ that is an unbiased estimator of $g(\theta)$, then g^* is a MVUE for $g(\theta)$.

4 Completeness

A family $\{p_\theta, \theta \in \Lambda\}$ is **complete** if, for any function $f : \Gamma \rightarrow \mathbb{R}$, we have

$$\mathbb{E}_\theta[f(Y)] = 0 \quad \forall \theta \in \Lambda \quad \Rightarrow \quad f(Y) = 0 \text{ a.s. } p_\theta, \quad \forall \theta \in \Lambda.$$

Example. Consider $\Gamma = \{0, 1, 2\}$ and the family

$$p_\theta(y) = \begin{cases} \theta^2 & : y = 0 \\ 2\theta & : y = 1 \\ 1 - \theta^2 - 2\theta & : y = 2 \end{cases}$$

where $\theta \in \Lambda = [0, \sqrt{2} - 1]$ (ensuring that p_θ is a valid pmf). The family $\{p_\theta\}$ is represented by a curve (parameterized by θ) in the 3-D probability simplex. For any function $f(y)$, we have

$$\begin{aligned} \mathbb{E}_\theta[f(Y)] &= \theta^2 f(0) + 2\theta f(1) + (1 - \theta^2 - 2\theta)f(2) \\ &= \theta^2(f(0) - f(2)) + 2\theta(f(1) - f(2)) + f(2) \end{aligned}$$

which is a second-order polynomial in θ . To have $\mathbb{E}_\theta[f(Y)] = 0$ for all $\theta \in \Lambda$, we need to have all three coefficients $f(0) - f(2)$, $f(1) - f(2)$, $f(2)$, equal to zero. Equivalently, $f(0) = f(1) = f(2) = 0$, i.e., the function $f(y)$ is zero for all $y \in \Gamma$. Hence the family $\{p_\theta\}$ is complete.

Denote by $\{q_\theta(t), t \in \mathcal{T}\}$ the distribution of $T(Y)$ when $Y \sim p_\theta$. We say that $T(Y)$ is a **complete sufficient statistic** if the family $\{q_\theta, \theta \in \Lambda\}$ is complete.

5 Finding MVUEs

Property 1. If a family $\{p_\theta, \theta \in \Lambda\}$ is complete then all sufficient statistics for θ are trivial, i.e., any such $T(Y)$ is an invertible function of Y . Compressing Y would destroy information about θ .

Property 2. A complete sufficient statistic is always minimal.

Property 3. Let T be a complete sufficient statistic and $g^*(T(Y))$ be an unbiased estimator of $g(\theta)$. Then g^* is the unique MVUE.

Property 3 has a direct, useful implication. If we are able to find a complete sufficient statistic $T(Y)$ for θ and *any* unbiased estimator $\hat{g}(Y)$ of $g(\theta)$, then we obtain the MVUE by Rao-Blackwellization of $\hat{g}(Y)$.

6 Exponential Families

We say that $\{p_\theta(y), \theta \in \Lambda, y \in \Gamma\}$ is an exponential family if

$$p_\theta(y) = h(y)C(\theta)e^{\sum_{i=1}^m Q_i(\theta)T_i(y)}, \quad \theta \in \Lambda, y \in \Gamma$$

for real-valued functions $h(y)$, $C(\theta)$, $Q_i(\theta)$, $T_i(y)$, $1 \leq i \leq m$. Examples include Gaussian, Poisson, Laplacian, and binomial distributions. For instance, the Poisson distribution with parameter $\theta > 0$ may be written in the form

$$p_\theta(y) = e^{y \ln \theta} \frac{1}{y!} e^{-\theta}, \quad y = 0, 1, 2, \dots$$

which is a 1-D exponential family with functions $h(y) = \frac{1}{y!}$, $C(\theta) = e^{-\theta}$, $Q(\theta) = \ln \theta$, and $T(y) = y$.

Completeness theorem for exponential families. If the parameter set Λ contains a m -dimensional rectangle, then the \mathbb{R}^m -valued function $T(y) = [T_l(y), 1 \leq l \leq m]$ is a complete sufficient statistic for θ .

Example. Consider a homogeneous Poisson process with intensity $\lambda > 0$. The number of arrivals in the interval $[0, T]$ is a Poisson random variable with parameter λT :

$$p_\lambda(n) = e^{-\lambda T} \frac{(\lambda T)^n}{n!}.$$

Denote by $\underline{Y} = [Y_1, \dots, Y_n]$ the vector made of the n ordered arrival times. Conditioned on n , the vector \underline{Y} is distributed as

$$p(\underline{y}|n) = \frac{n!}{T^n} 1_{\{0 \leq y_1 \leq \dots \leq y_n \leq T\}}.$$

The distribution of the data (n, \underline{y}) takes the form

$$\begin{aligned} p(n, \underline{y}) &= p_\lambda(n) p(\underline{y}|n) \\ &= e^{-\lambda T} \frac{(\lambda T)^n}{n!} \frac{n!}{T^n} 1_{\{0 \leq y_1 \leq \dots \leq y_n \leq T\}} \\ &= \lambda^n e^{-\lambda T} 1_{\{0 \leq y_1 \leq \dots \leq y_n \leq T\}} \\ &= e^{n \ln \lambda} e^{-\lambda T} 1_{\{0 \leq y_1 \leq \dots \leq y_n \leq T\}}. \end{aligned}$$

By the factorization theorem, N is a sufficient statistic for λ . By the completeness theorem for exponential families, N is a complete sufficient statistic. Since $\mathbb{E}_\lambda[N] = \lambda T$, an unbiased estimator of λ is

$$\hat{\lambda}(n) = \frac{n}{T}.$$

Since this estimator is a function of a complete sufficient statistic, it is necessarily MVUE.

7 Summary

We can identify sufficient statistics by looking for factorizations of $p_\theta(y)$. Given a sufficient statistic $T(Y)$ and an arbitrary unbiased estimator, Rao-Blackwellization of that estimator can only improve it.

While it may be difficult to check whether a sufficient statistic is complete, the task is much easier for exponential families. If a complete sufficient statistic $T(Y)$ is identified, then finding an unbiased estimator based on $T(Y)$ yields the unique MVUE.