**STATS 300A: Theory of Statistics** **Fall 2015**

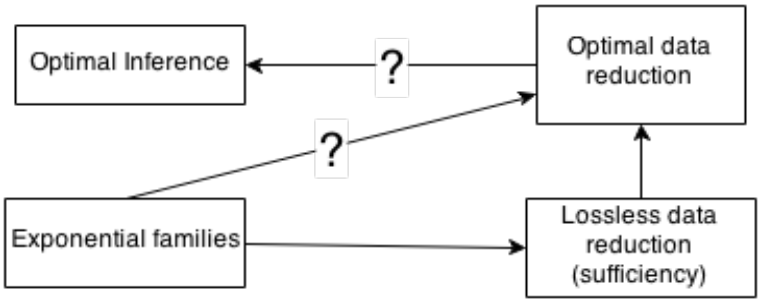
# Lecture 3 — September 29

*Lecturer: Lester Mackey* *Scribe: Konstantin Lopyrev, Karthik Rajkumar*

***Warning:*** *These notes may contain factual and/or typographic errors.*

## 3.1 Recap

Before discussing today's topic matter, let's take a step back and situate ourselves with respect to the big picture. As mentioned in Lecture 1, a primary focus of this course is optimal inference. As a first step toward reasoning about optimality, we began to examine which statistics of the data that we observe are actually relevant in a given inferential task. We learned about lossless data reduction and about the concept of sufficiency. We understood through our notions of statistical risk that lossless data reduction does just as well as the original model and that extraneous data can only hurt the model.
We then examined a broad class of distributions, viz. the exponential families, and saw how they were intimately related to our notions of sufficiency. Using the concept of minimal sufficiency, we initiated a discussion of how data could be maximally compressed without losing information relevant to the inference task.




Our present roadmap leads us to examine first how the exponential families and other distribution can be optimally reduced (this lecture), before proceeding to see how optimal data compression relates to optimal inference (next time).

## 3.2 Minimal Sufficiency

Recall that we defined a notion of maximum achievable lossless data reduction in the last lecture.

**Definition 1** (Minimal Sufficiency)**.** A sufficient statistic $T$ is **minimal** if for every sufficient statistic $T'$ and for every $x, y \in \mathcal{X}$, $T(x) = T(y)$ whenever $T'(x) = T'(y)$. In other words, $T$ is a function of $T'$ (there exists $f$ such that $T(x) = f(T'(x))$ for any $x \in \mathcal{X}$).

The following theorem provides a means for checking minimal sufficiency when our model distributions admit densities.

**Theorem 1.** Let $\{p(x;\theta), \theta \in \Omega\}$ be a family of densities with respect to some measure $\mu$.[1] Suppose that there exists a statistic $T$ such that for every $x, y \in \mathcal{X}$:

$$p(x;\theta) = C_{x,y}p(y;\theta) \quad \iff \quad T(x) = T(y)$$

for every $\theta$ and some $C_{x,y} \in \mathbb{R}$. Then $T$ is a minimal sufficient statistic.

To prove this result, we first show that $T$ is sufficient and then that it is minimal.

*Proof.* **T is sufficient:** Start with $T(\mathcal{X}) = \{t\colon t = T(x) \text{ for some } x \in \mathcal{X}\}$ = range of $T$. For each $t \in T(\mathcal{X})$, consider the preimage $A_t = \{x\colon T(x) = t\}$ and select an arbitrary representative $x_t$ from each $A_t$. Then, for any $y \in \mathcal{X}$ we have $y \in A_{T(y)}$ and $x_{T(y)} \in A_{T(y)}$. By the definition of $A_t$ this implies that $T(y) = T(x_{T(y)})$. From the assumption of the theorem,

$$\begin{aligned} p(y;\theta) &= C_{y,x_{T(y)}}p(x_{T(y)};\theta) \\ &= h(y)g_\theta(T(y)) \end{aligned}$$

which yields sufficiency of $T$ by the NFFC.

**T is minimal:** Consider another sufficient statistic $T'$. By the NFFC

$$p(x;\theta) = \tilde{g}_\theta(T'(x))\tilde{h}(x).$$

Take any $x, y$ such that $T'(x) = T'(y)$. Then

$$\begin{aligned} p(x;\theta) &= \tilde{g}_\theta(T'(x))\tilde{h}(x) \\ &= \tilde{g}_\theta(T'(y))\tilde{h}(y)\frac{\tilde{h}(x)}{\tilde{h}(y)} \\ &= p(y;\theta)C_{x,y} \end{aligned}$$

Hence, $T(x) = T(y)$ by the assumption of the theorem. So, $T'(x) = T'(y)$ implies $T(x) = T(y)$ for any sufficient statistic $T'$ and any $x, y$. As a result, $T$ is a minimal sufficient statistic. □

Interestingly, minimal sufficient statistics are quite easy to find when working with minimal exponential families.

**Remark 1.** For any minimal $s$-dimensional exponential family the statistic $(\sum_i T_1(X_i), \ldots, \sum_i T_s(X_i))$ is a minimal sufficient statistic. (See Keener Ex. 3.12)

[1] Note that in this class, $\mu$ will typically be Lebesgue measure for continuous distributions or counting measure for discrete distributions.

**Example 1** (Curved Exponential Family)**.** Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\sigma, \sigma^2)$, $\theta = \sigma > 0$. Then

$$\begin{aligned}\frac{p(x;\theta)}{p(y;\theta)} &= \frac{\exp\left(-\frac{1}{2\sigma^2}\sum_i x_i^2 + \frac{\sigma}{\sigma^2}\sum_i x_i - \frac{n\sigma^2}{2\sigma^2}\right)}{\exp\left(-\frac{1}{2\sigma^2}\sum_i y_i^2 + \frac{\sigma}{\sigma^2}\sum_i y_i - \frac{n\sigma^2}{2\sigma^2}\right)} \\ &= \exp\left(-\frac{1}{2\sigma^2}\left(\sum_i x_i^2 - \sum_i y_i^2\right) + \frac{1}{\sigma}\left(\sum_i x_i - \sum_i y_i\right)\right).\end{aligned}$$

Is $T(X) = (T_1(X), T_2(X)) = (\sum_i X_i^2, \sum_i X_i)$ minimal sufficient?

First, if $T(x) = T(y)$ for some $x, y \in \mathcal{X}$, then their ratio is equal to 1 and hence does not depend on $\theta$. This means $T$ is sufficient.

Second, if for some $x, y$, the ratio is independent of $\theta$, notice that the ratio $\to 1$ as $\sigma \to \infty$ (log of the ratio $\to 0$). Therefore $C_{x,y} = 1$ and $\log C_{x,y} = 0 = \log\left(\frac{p(x;\theta)}{p(y\theta)}\right)$. This implies

$$\frac{1}{2\sigma^2}(T_1(y) - T_1(x)) + \frac{1}{\sigma}(T_2(x) - T_2(y)) = 0 \qquad \forall\sigma$$

Multiplying $2\sigma^2$ through, we get

$$T_1(y) - T_1(x) = 2\sigma(T_2(y) - T_2(x) \qquad \forall\sigma$$

We see that the RHS $\to 0$ as $\sigma \to 0$. So

$$\begin{aligned} & T_1(y) - T_1(x) = 0 \\ \implies\ & T_2(y) = T_2(x)\end{aligned}$$

Consequently, $T$ is a minimal sufficient statistic.

**Remark 2.** What if the support of $X$, i.e., the set $\{x \in \mathcal{X} \colon p(x;\theta) > 0\}$, depends on $\theta$? Then if $p(x;\theta) = C_{x,y}p(y;\theta)$, $x$ and $y$ must be supported by (exactly) the same $\theta$'s. Otherwise there would be a $\theta$ dependence, which we assumed we did not have.

**Example 2.** Let $X_1, \ldots, X_n \overset{iid}{\sim} U(0,\theta)$ and $T(X) = \max\{X_1, \ldots, X_n\}$. In that case for $x = (x_1, ..., x_n)$ such that $x_i > 0, i = 1, ..., n$ (in short, $x > 0$)

$$p(x;\theta) = \prod_{i=1}^{n} \frac{1}{\theta}\mathbb{I}(x_i < \theta) = \frac{1}{\theta^n}\mathbb{I}(T(x) < \theta).$$

If $T(x) = T(y)$ then $p(x;\theta) = 1 \times p(y;\theta)$. That scale ratio between the distributions does not depend on $\theta$ and so $T$ is sufficient.

Conversely, if $x, y > 0$ are supported by the same $\theta$'s, then $\{\theta$ supporting $x\} = (T(x), \infty) = (T(y), \infty) = \{\theta$ supporting $y\}$. Therefore $T(x) = T(y)$ and $T$ is a minimal sufficient statistic.

## 3.3 Ancillarity and Completeness

In each of the examples we encountered, we were able to achieve significant data compression without losing any information on our statistical model. But is this always the case? Are there ever instances where sufficient statistics, or even minimal sufficient statistics, don't reduce the data in any significant way?
The answer is a resounding yes.

**Example 3.** Consider $X_1, \ldots, X_n \overset{iid}{\sim} \text{CauchyLoc}(\theta)$, whose distribution is given by

$$p(x;\theta) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2} = f(x-\theta),$$

then $(X_{(1)}, \ldots, X_{(n)})$ is minimal sufficient. (See TPE 1.5.)

This is also true for the double exponential location model, $p(x;\theta) \propto \exp(|x-\theta|)$.

So how can we explain this drastic difference in compressibility? It turns out that the lossless compressibility of data drawn from a model is related to the amount of ancillary information present in its minimal sufficient statistics.

**Definition 2.** A statistic $A$ is **ancillary** for $X \sim \mathbb{P}_\theta \in \mathcal{P}$ if the distribution of $A(X)$ does not depend on $\theta$.

**Example 4.** Consider again $X_1, \ldots, X_n \overset{iid}{\sim} \text{CauchyLoc}(\theta)$. Then $A(X) = X_{(n)} - X_{(1)}$ is ancillary even though $(X_{(1)}, \ldots, X_{(n)})$ is minimal sufficient. To see this, note that $X_i = Z_i + \theta$ for $Z_i \overset{iid}{\sim} \text{CauchyLoc}(0)$, so $X_{(i)} = Z_{(i)} + \theta$, and $A(X) = A(Z)$. This last quantity does not depend on $\theta$.

Ideally, the statistics that we make use of will include as little ancillary information as possible. In fact, we will demand even more than this. To do this we introduce a slightly weaker notion of ancillarity:

**Definition 3.** A statistic $A$ is **first-order ancillary** for $X \sim \mathbb{P}_\theta \in \mathcal{P}$ if $\mathbb{E}_\theta[A(X)]$ does not depend on $\theta$.

From this we define the concept of complete statistics.

**Definition 4.** A statistic $T$ is **complete** for $X \sim \mathbb{P}_\theta \in \mathcal{P}$ if no non-constant function of $T$ is first-order ancillary. In other words, if $\mathbb{E}_\theta[f(T(X))] = 0$ for all $\theta$, then $f(T(X)) = 0$ with probability 1 for all $\theta$.

Completeness formalizes our ideal notion of optimal data reduction, whereas minimal sufficiency is our achievable notion of optimal data reduction. We now examine some properties of complete statistics.

1. If $T$ is complete sufficient, then $T$ is minimal sufficient. This is known as Bahadur's theorem.

2. Complete sufficient statistics yield optimal unbiased estimators. (This will be studied in the next lecture.)

**Example 5.** Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Ber}(\theta)$, $\theta \in (0,1)$. Then $T(X) = \sum_i X_i$ is sufficient. Suppose $\mathbb{E}_\theta[f(T(X))] = 0$ for all $\theta \in (0,1)$. This means

$$\sum_{j=0}^{n} f(j)\binom{n}{j}\theta^j(1-\theta)^{n-j} = 0, \qquad \forall \theta \in (0,1).$$

Dividing through by $\theta^n$ and using $\beta = \frac{\theta}{1-\theta}$, we get

$$\sum_{j=1}^{n} f(j)\binom{n}{j}\beta^j = 0, \qquad \forall \beta > 0.$$

If $f$ are non-zero, then the quantity on the left is a polynomial of degree at most $n$. However, an $n$th-degree polynomial can have at most $n$ roots. Hence it impossible for the LHS to equal 0 for every $\beta > 0$ unless $f = 0$. Hence $T$ is complete.

**Example 6.** Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\theta, \sigma^2)$ with unknown $\mu \in \mathbb{R}$ and known $\sigma^2 > 0$. Is $\overline{X}_n = \frac{1}{n}\sum_i X_i$ complete for this model? (We already know that it is minimal sufficient.) The answer is yes, but to keep the algebra simple we will show that it is complete in the special case of $n = 1$ and $\sigma = 1$, so that $T(X) = X \sim \mathcal{N}(\theta, 1)$. Suppose

$$\mathbb{E}_\theta[f(X)] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(x)\exp\left(-\frac{(x-\theta)^2}{2}\right)dx = 0 \qquad \forall \theta \in \mathbb{R}$$

Multiplying through by $\sqrt{2\pi}e^{\frac{\theta^2}{2}}$ gives

$$\int_{-\infty}^{\infty} f(x)\exp\left(-\frac{x^2}{2}\right)\exp(\theta x)\,dx = 0 \qquad \forall \theta \tag{3.1}$$

Now decompose $f$ into its positive and negative part as $f(x) = f_+(x) - f_-(x)$, where $f_+(x) = \max(f(x), 0)$ and $f_-(x) = \max(-f(x), 0)$. Then $f_+(x) \geq 0$ and $f_-(x) \geq 0$ for all $x \in \mathbb{R}$, and $f_+(x) = f_-(x)$ if and only if $f_+(x) = f_-(x) = 0$.

If $f(x) \geq 0$ a.e. or $f(x) \leq 0$ a.e., then (3.1) implies that $f(x) = 0$ a.e. because setting $\theta = 0$ gives us an integral of a nonnegative (or nonpositive) function being zero. This is completeness.

In the other case, $f_+$ and $f_-$ have non-zero components and we may write

$$\frac{\int_{-\infty}^{\infty} f_+(x)e^{\frac{-x^2}{2}}e^{\theta x}dx}{\int_{-\infty}^{\infty} f_+(x)e^{-\frac{x^2}{2}}dx} = \frac{\int_{-\infty}^{\infty} f_-(x)e^{\frac{-x^2}{2}}e^{\theta x}dx}{\int_{-\infty}^{\infty} f_-(x)e^{-\frac{x^2}{2}}dx}, \tag{3.2}$$

where the equality of the numerators follows from (3.1), and the equality of the denominators follows from setting $\theta = 0$. The quantity

$$\frac{f_+(x)e^{-\frac{x^2}{2}}}{\int_{-\infty}^{\infty} f_+(x)e^{-\frac{x^2}{2}}dx}$$

defines a probability density, and the left-hand-side of (3.2) is the moment generating function of this density. Similarly, the right-hand-side is the moment generating function of the density

$$\frac{f_-(x)e^{-\frac{x^2}{2}}}{\int_{-\infty}^{\infty} f_-(x)e^{-\frac{x^2}{2}}dx}.$$

The equality of the moment generating functions implies equality of the densities, which in turn implies $f_+(x) = f_-(x)$ a.e. Then $f_+(x) = f_-(x) = 0$ a.e. so $f(x) = 0$ a.e. Hence $T$ is complete.