Statistics 210A Lecture 4 Notes

Daniel Raban

September 7, 2021

1 Sufficient Statistics

1.1 Recap: differential identities for exponential families

Last time, we were talking about exponential families $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ with densities

$$p_\theta(x) = e^{\eta(\theta)^\top T(x) - B(\theta)} h(x).$$

In natural parameters, we have

$$p_\eta(x) = e^{\eta^\top T(x) - A(\eta)} h(x).$$

Last time, we proved some differential identities by starting with the equation

$$e^{A(\eta)} = \int e^{\eta^\top T(x)} h(x) \, d\mu(x)$$

and differentiating with respect to η_j. We saw that

$$\nabla A(\eta) = \mathbb{E}_\eta[T(X)], \qquad \nabla^2 A(\eta) = \operatorname{Var}_\eta(T(X)).$$

In general, we have

$$e^{-A(\eta)} \frac{\partial^{k_1 + \cdots + k_s}}{\partial_{\eta_1}^{k_1} \cdots \partial_{\eta_s}^{k_s}} (e^{A(\eta)}) = \mathbb{E}_\eta[T_1^{k_1} \cdots T_s^{k_s}].$$

This is saying that $e^{A(\eta+u) - A(\eta)}$ is the **moment generating function** of T:

$$\frac{\partial}{\partial u_j} e^{A(\eta+u) - A(\eta)}|_{u=0} = \left(\frac{\partial}{\partial \eta_j} e^{A(\eta)} \right) \cdot e^{-A(\eta)}.$$

If we take logs, we get that $A(\eta + u) - A(\eta)$ is the **cumulant generating function** of $T(X)$.[1]

[1] We have been calling $A(\eta)$ the CGF, but technically that is only the case where $\eta = 0$.

Here is another calculation of the MGF for $T(X)$ in an exponential family:

$$\begin{aligned} M_\eta^{T(X)}(u) &= \mathbb{E}_\eta[e^{u^\top T(X)}] \\ &= \int e^{u^\top T} e^{\eta^\top T - A(\eta)} h \, d\mu \\ &= e^{-A(\eta)} e^{A(u+\eta)}. \end{aligned}$$

1.2 Sufficiency

Our motivation is going to be the example of coin flipping.

Example 1.1. Suppose $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Ber}(\theta)$, so our data is $X \sim \prod_i \theta^{x_i}(1-\theta)^{x_i}$ on $\{0,1\}^n$. Instead of observing the whole sequence, we can observe a summary statistic $T(X) = \sum_i X_i \sim \text{Binom}(n,\theta) = \theta^t(1-\theta)^{n-t}\binom{n}{t}$ on $\{0, 1 \dots, n\}$ which only records the total number of heads. This is a lossy compression of the data $(X_1, \dots, X_n) \mapsto T(X)$. Why can we justify this?

We can think of the information in $(X_1, \dots, X_n)$ as coming in two parts: the first part is $T(X)$, which is the part relevant to estimating θ, and the second part is the ordering, which doesn't depend on θ. The reason that $T(X)$ is the important part for estimating θ is that $T(X)$ is the only part that depends on θ.

Definition 1.1. Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a statistical model for data X. $T(X)$ is **sufficient** for the model $\mathcal{P}$ if $P_\theta(X \mid T)$ does not depend on θ.

Example 1.2. Continuing our coin flipping example,

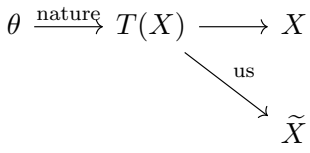
$$\begin{aligned} \mathbb{P}_\theta(X = x \mid T = t) &= \frac{\mathbb{P}_\theta(X = x, T = t)}{\mathbb{P}_\theta(T = t)} \\ &= \frac{\theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}}{\theta^t(1-\theta)^{n-t}\binom{n}{t}} \mathbb{1}_{\{\sum_i x_i = t\}} \\ &= \frac{1}{\binom{n}{t}} \mathbb{1}_{\{\sum_i x_i = t\}}. \end{aligned}$$

The interpretation is that we can think of Nature as generating the data in 2 steps:

1. Generate $T(X) \sim P_\theta(T(X))$, dependent on θ.
2. Generate $X \sim P(X \mid T)$, not dependent on θ.

Sufficiency principle: If $T(X)$ is sufficient, then any statistical procedure should depend on the data X only through T.

Why should we believe in this sufficiency principle? Suppose we generate $\widetilde{X} \sim \mathbb{P}(X \mid T)$.



Then $\widetilde{X} \stackrel{d}{=} X$, so any estimator gives $\delta(\widetilde{X}) \stackrel{d}{=} \delta(X)$. So we should always be fine using $T(X)$, since we don't really lose any information by using it. Later, we will see that using sufficient statistics can reduce the loss we incur in estimation.

1.3 Factorization theorem for sufficient statistics

Theorem 1.1 (Fischer-Neyman)**.** *Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a statistical model with densities $p_\theta(x)$ with respect to a common dominating measure μ. Then T is sufficient for $\mathcal{P}$ if and only if there exist nonnegative functions g_θ, h such that $p_\theta(x) = g_\theta(T(x))h(x)$ for μ-a.e. x.*

Here is a "physics proof." For a careful proof, check Keener.

Proof. ($\Longleftarrow$):

$$\begin{aligned} p_\theta(x \mid T = t) &= \mathbb{1}_{\{T(x)=t\}} \cdot \frac{g_\theta(t)h(x)}{\int_{T(z)=t} g_\theta(t)h(z)\, d\mu(z)} \\ &= \mathbb{1}_{\{T(x)=t\}} \cdot \frac{h(x)}{\int_{T(z)=t} h(z)\, d\mu(z)}. \end{aligned}$$

($\Longrightarrow$): Take

$$g_\theta(t) = \int_{T(x)=t} p_\theta(x)\, d\mu(x) = \mathbb{P}_\theta(T(X) = t),$$

$$h(x) = \frac{p_{\theta_0}(x)}{\int_{T(z)=T(x)} p_{\theta_0}(z)\, d\mu(z)} = \mathbb{P}_{\theta_0}(X = x \mid T(X) = T(x)).$$

for any fixed $\theta_0 \in \Theta$. Then

$$\begin{aligned} g_\theta(T(x))h(x) &= \mathbb{P}(T(X) = T(x))\mathbb{P}_\theta(X = x \mid T(X) = T(x)) \\ &= p_\theta(x). \end{aligned}$$

□

Example 1.3. For exponential families,

$$p_\theta(x) = \underbrace{e^{\eta(\theta)^\top T(x) - B(\theta)}}_{g_\theta(T(x))} \underbrace{h(x)}_{h(x)},$$

so T is sufficient for θ.

Example 1.4. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P_\theta^{(1)}$ for any model $\mathcal{P}^{(1)} = \{P_\theta^{(1)} : \theta \in \Theta\}$ on $\mathcal{X} \subseteq \mathbb{R}$. $P_\theta^{(1)}$ is invariant to permuting $X = (X_1, \ldots, X_n)$. The **order statistics** $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ are defined by $X_{(k)} =$ the k-th smallest value (counting repeats). For example, if $X = (1, 3, 3, -1)$, then $X_{(1)} = -1, X_{(2)} = 1, X_{(3)} = 3, X_{(4)} = 3$.

If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P_\theta^{(1)}$ is any univariate model $\mathcal{P}^{(1)}$, then the order statistics are sufficient. For a more general $\mathcal{X}$, we can say the **empirical distribution**

$$\widehat{P}_n(\cdot) = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}(\cdot)$$

is sufficient.

1.4 Minimal sufficiency

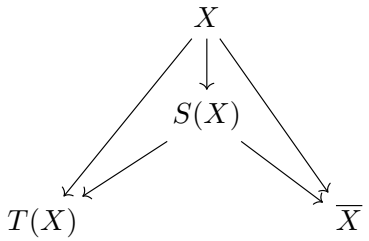
Example 1.5. Consider $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, 1)$. The following statistics are sufficient:

$$T(X) = \sum_i X_i, \qquad \overline{X} = \frac{1}{n} \sum_i X_i,$$

$$S(X) = (X_{(1)}, \ldots, X_{(n)}), \qquad X = (X_1, \ldots, X_n).$$

It seems like the latter two statistics have more information than $T(X)$ or $\overline{X}$. These are all sufficient statistics (and in fact the data itself is always sufficient), so what should we do with regards to the sufficiency principle? The idea is to find sufficient statistics with the least amount of information, i.e. the ones that cannot recover the others.

Here is a diagram that expresses which statistics have more information than others:



Next time, we will talk about minimal sufficient statistics, which have minimal information while remaining sufficient.