STATS 300A - Lecture 2

Michael Howes

October 25, 2021

1 Recap

Last lecture we introduced decision theory and the concept of “inadmissability” as a way to compare two decision procedures. We quickly saw that very rarely will we be able to compare two decision processes by inadmissability. We learnt that there were two ways to work around this

(a) We could restrict the class of decision procedures. For instance we could focus on unbiased estimators rather than all estimators.

(b) We could collapse the risk function to a single number (rather than a function of θ). We saw the ideas of Bayes risk and minimax risk.

We also introduced the concept of sufficiency as a means of data reduction and stated the NFFC. This is where we will start today’s class.

2 Neymann-Fisher Factorisation Condition

Recall that a statistic $T(X)$ is sufficient for a family $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ if the distribution of $X|T = t$ does not depend on θ. We stated the NFFC as a way of checking and finding sufficient statistics.

Theorem 1. *Suppose each P_θ has a density $p(x|\theta)$ (wrt a common measure μ), then $T(X)$ is sufficient for X if and only if*

$$p(x|\theta) = g_\theta(T(x))h(x),$$

for all $x \in X$ and $\theta \in \Omega$.

Proof. We will prove the discete case (ie μ is the counting measure on a countable set). First assume that we have the factorisation $p(x|\theta) = g_\theta(T(x))h(x)$. Then for all $x \in \mathcal{X}$ and $\theta \in \Omega$, we have

$$\begin{aligned}
P_\theta(X = x|T(X) = t) &= \frac{P_\theta(X = x, T(X) = t)}{P_\theta(T(X) = t)} \\
&= \frac{\mathbf{1}_{T(x)=(t}(x,t))P_\theta(X = x)}{\sum_{x' \in \mathcal{X}, T(x')=t} P_\theta(X = x')} \\
&= \frac{\mathbf{1}_{T(x)=t}(x,t)p(x|\theta)}{\sum_{x' \in \mathcal{X}, T(x')=t} p(x|\theta)} \\
&= \frac{\mathbf{1}_{T(x)=t} g_\theta(T(x))h(x)}{\sum_{x' \in \mathcal{X}, T(x')=t} g_\theta(T(x'))h(x')} \\
&= \frac{g_\theta(t)\mathbf{1}_{T(x)=t}(x,t)h(x)}{g_\theta(t)\sum_{x' \in \mathcal{X}, T(x')=t} h(x')} \\
&= \frac{\mathbf{1}_{T(x)=t}(x,t)h(x)}{\sum_{x' \in \mathcal{X}, T(x')=t} h(x')}.
\end{aligned}$$

The above expression does not depend on θ and hence T is sufficient.

Now suppose that T is a sufficient statistic for X. Fix $\theta_0 \in \Omega$ and define $g_\theta(t) = P_\theta(T(X) = t)$ and define $h(x) = P_{\theta_0}(X = x|T(X) = T(x))$. Note that by sufficiency $h(x) = P_\theta(X = x|T(X) = x)$ for all $\theta \in \Omega$. We thus have that

$$\begin{aligned} p(x|\theta) &= P_\theta(X = x) \\ &= P_\theta(X = x, T(X) = T(x)) \\ &= P_\theta(T(X) = T(x))P_\theta(X = x|T(X) = T(x)) \\ &= g_\theta(T(x))h(x). \end{aligned}$$

Thus $p(x|\theta)$ factorises in the required way. □

For the rest of today's lecture we will focus on two topics.

(a) Exponential families.

(b) The concept of minimal sufficeny and optimal data reduction.

3 Exponential Families

Definition 1. A collection of distributions $\mathcal{P} = \{P_\theta|\theta \in \Omega\}$ is an *s-dimensional exponential family* if each distribution P_θ has a density $p(x|\theta)$ (wrt to a common reference measure μ) such that

$$p(x|\theta) = h(x)\exp\left\{\sum_{j=1}^{s}\eta_j(\theta)T_j(x) - B(\theta)\right\}.$$

The function $h(x)$ is called the *base measure*. The function $\eta_j(\theta)$ are called *natural parameters* and $B(\theta)$ is called the *log partition function*.

Note that $(T_1(X), \ldots, T_s(X))$ is a sufficient statistic for P_θ by the NFFC. Also note that $B(\theta)$ is determined by h, η_j and T_j since we must have $\int_{\mathcal{X}} p(x|\theta)d\mu(x) = 1$ and thus

$$B(\theta) = \log\left(\int_{\mathcal{X}} h(x)\exp\left\{\sum_{j=1}^{s}\eta_j(\theta)T_j(x)\right\}d\mu(x)\right).$$

Definition 2. The *canonical form* of an exponential family is

$$p(x|\eta) = h(x)\exp\left\{\sum_{j=1}^{s}\eta_j T_j(x) - A(\eta)\right\}.$$

Example 1. Suppose that $P_\theta = \text{Binom}(n, \theta)$ where n is fixed and $\theta \in [0, 1]$. Then

$$\begin{aligned} p(x|\theta) &= \binom{n}{x}\theta^x(1-\theta)^{n-x} \\ &= \binom{n}{x}\exp\{\log(\theta)x + \log(1-\theta)(n-x)\} \\ &= \binom{n}{x}\exp\left\{\frac{\log(\theta)}{\log(1-\theta)}x + \log(1-\theta)n\right\}. \end{aligned}$$

Thus the binomial family is a 1-dimensional exponential family with $h(x) = \binom{n}{x}$, $T_i(x) = x$, $\eta_i(\theta) = \frac{\log(\theta)}{\log(1-\theta)}$ and $B(\theta) = -\log(1-\theta)n$.

The Poission, Beta, Gamma, Gaussian, Dirichlet, Pareto and Wishart distribution are all also exponential families.

Definition 3. Suppose we have a fixed choice of $h(x)$ and $T_j(x)$ as in the definition of an exponential family in canonical form. Then the *natural parameter space* is the set Θ of all η such that

$$0 < \int_{\mathcal{X}} h(x) \exp\left\{ \sum_{j=1}^{s} \eta_j T_j(x) \right\} d\mu(x) < \infty.$$

This is the space where the normalisation constant is in $(0, \infty)$. An application of Hölder's inequality proves that Θ is convex.

The following definitions are intended to rule out circumstances where our model has been over parameterised.

Definition 4. A canonical exponential family $\mathcal{P} = \{P_\eta : \eta \in H\}$ is minimal if

- For all $(\lambda_j)_{j=0}^{s}$, if $\sum_{j=1}^{s} \lambda_j T_j(x) = \lambda_0$ for all $x \in \mathcal{X}$, then $\lambda_j = 0$ for all $j = 0, 1, \ldots, s$. (That is the function (T_j) do not satisfy any affine constraints).
- For all $(\lambda_j)_{j=0}^{s}$, if $\sum_{j=1}^{s} \lambda_j \eta_j = \lambda_0$ for all $\eta \in H$, then $\lambda_j = 0$ for all $j = 0, 1, \ldots, s$. (That is the set H does not satisfy any affine constraints).

Definition 5. Suppose $\mathcal{P} = \{P_\eta : \eta \in H\}$ is an s-dimensional exponential family in canonical form. If H contains an open s-dimensional rectangle, then $\mathcal{P}$ is called *full rank*. If H does not contain any open s-dimensional rectangles, then $\mathcal{P}$ is said to be curved.

Exponential famillies have the following nice properties

(a) An iid sampple from an exponential family is again an exponential family.

(b) If $f(X)$ is integrable (ie $\mathbb{E}_\eta[|f(X)|] < \infty$) for all η in an open subset of H, then the function

$$G(f, \eta) = \int f(x) \exp\left\{ \sum_{j=1}^{s} \eta_j T_j(x) \right\} h(x) d\mu(x),$$

is infinitely continuously differentiable on said open subset and integration and differentiation can be exchanged (see TSH 2.7.1 for details).

This second point is powerful and can be used for moment calculations. Let's take $f(x) = 1$, then

$$G(f, \eta) = \int \exp\left\{ \sum_{i=1}^{s} \eta_j T_j(x) \right\} h(x) d\mu(x) = \exp(A(\eta)).$$

Thus

$$\begin{aligned} \frac{\partial}{\partial \eta_i} A(\eta) &= \frac{\frac{\partial}{\partial \eta_i} \exp(A(\eta))}{\exp(A(\eta))} \\ &= \frac{\int T_j(x) h(x) \exp\left\{ \sum_{j=1}^{s} \eta_j T_j(x) \right\} d\mu(x)}{\exp(A\eta)} \\ &= \mathbb{E}_\eta[T_i(X)]. \end{aligned}$$

Higher moments of T can be calculated by taking higher derivatives. We now turn to minimal sufficiency.

4 Optimal data reduction

Definition 6. A sufficient statistic T is *minimal* if for every sufficient statistics T', T is a function of T'. That is, if $T'(x) = T'(y)$, then $T(x) = T(y)$.

This means that T is the "coarest" minimal statistic. The following theorem gives a sufficient condition for a statistic to minimal sufficient.

Theorem 2. *Suppose that P_θ has a density $p(x|\theta)$ for all $\theta \in \Omega$ (with respect to a common measure μ). A statistic T is minimal sufficient if the following condition holds. For all $x, y \in \mathcal{X}$, $T(x) = T(y)$ if and only if there exists $C_{x,y}$ such that*

$$p(x|\theta) = C_{x,y} p(y|\theta),$$

for all $\theta \in \Omega$.

If the support of X does not depend on θ, then the condition that there exists $C_{x,y}$ such that $p(x|\theta) = C_{x,y} p(y|\theta)$ is equivalent to saying that the ratio

$$\frac{p(x|\theta)}{p(y|\theta)},$$

does not depend on θ. We will prove this theorem in the next lecture.