LECTURE 2 NOTES

1. Minimal sufficient statistics.

DEFINITION 1.1 (minimal sufficient statistic). *A statistic* $\mathbf{t} := \phi(\mathbf{x})$ *is a minimial sufficient statistic for the parametric model* $\mathcal{F}$ *if*

1. *it is sufficient.*
2. *it can be expressed as a function of any other sufficient statistic; i.e. for any other sufficient statistic* $\mathbf{t}'$*, there is some function* g *such that* $\mathbf{t} = g(\mathbf{t}')$.

Recall passing a random variable $\mathbf{x}$ through a function g is a form of data reduction: a statistic $\mathbf{t}' = g(\mathbf{t})$ has no more information than $\mathbf{t}$. Thus a minimal sufficient statistic is a sufficient statistic with the least information. That is, it is an *optimal* form of data reduction. Like sufficient statistics, minimal sufficient statistics are not unique. If $\mathbf{t}$ is minimal sufficient, $g(\mathbf{t})$ for any invertible g is also minimal sufficient.

In terms of partitions of the sample space, a minimal sufficient statistic induces the coarsest (hence minimal) partition of the sample space among all sufficient statistics. Although minimal sufficient statistics are not unique, the minimal sufficient partition is unique.

Going back to the coin tossing example, consider the statistic

$$\phi'(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) = \begin{cases} 100\mathbf{x}_1 + 10\mathbf{x}_2 + \mathbf{x}_3 & \mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3 = 2 \\ \mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3 & \text{otherwise.} \end{cases} \tag{1.1}$$

Table 1 gives the conditional distribution of $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) \mid \mathbf{t}'$. It is sufficient because the conditional distribution does not depend on p. However, it is not minimal because it is not a function of $\mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3$, which we know is also sufficient.

THEOREM 1.2. *A statistic* $\mathbf{t} := \phi(\mathbf{x})$ *is a miminal sufficent statistic for a parametric model* $\mathcal{F}$ *if the ratio* $\frac{f_\theta(x_1)}{f_\theta(x_2)}$ *does not depend on* θ *if and only if* $\phi(x_1) = \phi(x_2)$.

PROOF. We shall prove the theorem when the densities in the model have common support.

TABLE 1
The conditional distribution of three coin tosses given $\mathbf{t}'$*. We see that* $\mathbf{t}'$ *is not a minimal sufficient statistic:*

$(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_2)$	$\mathbf{t}'$	$f_p(\mathbf{x} \mid \mathbf{t}')$
(0,0,0)	0	1
(1,0,0)	1	1/3
(0,1,0)	1	1/3
(0,0,1)	1	1/3
(1,1,0)	110	1
(0,1,1)	11	1
(1,0,1)	101	1
(1,1,1)	3	1

First, we show that $\mathbf{t}$ is sufficient. For each set $\mathcal{A}_t = \{x \in \mathcal{X} : \phi(x) = t\}$ in the partition of $\mathcal{X}$ induced by $\mathbf{t}$, consider a point $x_t \in \mathcal{A}_t$. The density has the form

$$f_\theta(x) = \frac{f_\theta(x)}{f_\theta(x_{\phi(x)})} f_\theta(x_{\phi(x)}).$$

By assumption, the ratio $\frac{f_\theta(x)}{f_\theta(x')}$ does not depend on θ. Further,

1. $x_{\phi(x)}$ is a function of x, so $\frac{f_\theta(x)}{f_\theta(x')}$ depends only on x.
2. $f_\theta(x_{\phi(x)})$ depends only on $\phi(x)$.

Thus, by the factorization theorem, $\mathbf{t} = \phi(\mathbf{x})$ is a sufficient statistic.

We complete the proof by showing that $\mathbf{t}$ is minimal. It suffices to show that $\phi'(x_1) = \phi'(x_2)$ for any sufficient statistic $\mathbf{t}' := \phi'(x)$ implies $\phi(x_1) = \phi(x_2)$. By the factorization theorem, the density has the form

$$f_\theta(x) = g_\theta(\phi'(x))h(x)$$

For any $x_1 \in \mathcal{X}$, choose $x_2 \in \mathcal{X}$ such that $\phi'(x_1) = \phi'(x_2)$. We observe the ratio

$$\frac{f_\theta(x_1)}{f_\theta(x_2)} = \frac{g_\theta(\phi'(x_1))h(x_1)}{g_\theta(\phi'(x_2))h(x_2)} = \frac{h(x_1)}{h(x_2)}$$

does not depend on θ. By assumption, this implies $\phi(x_1) = \phi(x_2)$. □

EXAMPLE 1.3. *Let* $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \text{Poi}(\theta)$*. We showed that*

$$\mathbf{t} := \textstyle\sum_{i \in [n]} \mathbf{x}_i$$

is a sufficient statistic for the Poisson family. The ratio

$$\frac{f_\theta(x)}{f_\theta(x')} = \frac{\prod_{i \in [n]} x_i'!}{\prod_{i \in [n]} x_i!} \theta^{\sum_{i \in [n]} x_i - x_i'},$$

which does not depend on θ if and only if $t = t'$. By Theorem 1.2, $\mathbf{t}$ is a minimal sufficient statistic.

2. Exponential families.

DEFINITION 2.1. *A set of densities is an exponential family if the densities are of the form*

$$f_\theta(x) = \exp\left(\eta(\theta)^T \phi(x) - a(\theta)\right) h(x), \tag{2.1}$$

where

- $\eta(\theta) \in \mathbf{R}^p$ *are the* natural parameters.
- $\phi(x) \in \mathbf{R}^p$ *are sufficient statistics*
- $a(\theta) := \log \int_{\mathcal{X}} \exp\left(\eta(\theta)^T \phi(x)\right) h(x) dx$ *is the* log-partition function.
- $h : \mathcal{X} \to \mathbf{R}$ *is the* base measure.

Exponential families are of statistical relevance because many common distributions are exponential families. For example,

1. **binomial:**

$$\begin{aligned} f_p(x) &= \tbinom{n}{x} p^x (1-p)^{n-x} \\ &= \exp\big(x \log p + (n-x)\log(1-p)\big)\tbinom{n}{x} \\ &= \exp\big(x \log \tfrac{p}{1-p} + n \log(1-p)\big)\tbinom{n}{x}. \end{aligned}$$

   The natural parameter is the *logit* of p: $\log \frac{p}{1-p}$.

2. **Poisson:**

$$f_\lambda(x) = \frac{\lambda^x}{x!} e^{-\lambda} = \exp\left(x \log \lambda - \lambda\right) \frac{1}{x!}.$$

3. **Gaussian ($\sigma^2 = 1$):**

$$\begin{aligned} f_\mu(x) &= \tfrac{1}{\sqrt{2\pi}} \exp\left(\tfrac{1}{2}(x-\mu)^2\right) \\ &= \exp\left(\mu x - \tfrac{1}{2}\mu^2\right) \tfrac{e^{x^2/2}}{\sqrt{2\pi}}. \end{aligned}$$

4. **Gaussian (unknown σ^2):**

$$\begin{aligned} f_{\mu,\sigma^2}(x) &= \exp\big(-\tfrac{x^2}{2\sigma^2} + \tfrac{\mu x}{\sigma^2} - \tfrac{\mu^2}{2\sigma^2} - \log\sigma\big) \tfrac{1}{\sqrt{2\pi}} \\ &= \exp\left(\begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\sigma^{-2} \end{bmatrix}^T \begin{bmatrix} x \\ \frac{x^2}{2} \end{bmatrix} - \frac{\mu^2}{2\sigma^2} - \log\sigma\right) \frac{1}{\sqrt{2\pi}}. \end{aligned}$$

As all-inclusive as exponential families seem, there are notable exceptions. For example, the $\text{unif}(0, \theta)$ family is not an exponential family. More generally, parametric models whose support depends on the parameter are not exponential families.

Often, it is convenient to re-parametrize an exponential family in terms of its natural parameters η. Doing so gives the *canonical form* of the exponential family:

$$f_\eta(x) = \exp\left(\eta^T \phi(x) - a(\eta)\right) h(x).$$

For a set of sufficient statistics ϕ and a base measure h, the set of all valid natural parameters is the *natural parameter space*:

$$\mathcal{H} := \left\{\eta \in \mathbf{R}^p : \textstyle\int_{\mathcal{X}} \exp\left(\eta^T \phi(x)\right) h(x) dx < \infty\right\}.$$

LEMMA 2.2. *The log-partition function of an exponential family in canonical form is convex.*

PROOF. Let $\eta_1, \eta_2 \in \mathcal{H}$. For any $\alpha \in (0, 1)$,

$$\begin{aligned}
&a(\alpha\eta_1 + (1-\alpha)\eta_2) \\
&\quad = \log \int_{\mathcal{X}} \exp\left((\alpha\eta_1 + (1-\alpha)\eta_2)^T \phi(x)\right) h(x) dx \\
&\quad = \log \int_{\mathcal{X}} \exp\left(\eta_1^T \phi(x)\right)^\alpha \exp\left(\eta_2^T \phi(x)\right)^{1-\alpha} h(x) dx.
\end{aligned}$$

By Hölder's inequality,

$$\begin{aligned}
&\leq \log \left(\textstyle\int_{\mathcal{X}} \left(\exp\left(\eta_1^T \phi(x)\right) h(x)\right)^{\frac{\alpha}{\alpha}} dx \right)^\alpha \left(\textstyle\int_{\mathcal{X}} \left(\exp\left(\eta_2^T \phi(x)\right) h(x)\right)^{\frac{1-\alpha}{1-\alpha}} dx \right)^{1-\alpha} \\
&= \alpha a(\eta_1) + (1-\alpha) a(\eta_2).
\end{aligned}$$

Since $a(\eta_1)$ and $a(\eta_2)$ are finite by assumption, so is $a(\alpha\eta_1 + (1-\alpha)\eta_2)$. □

LEMMA 2.3. *The log-partition function $a : \Theta \to \mathbf{R}$ is convex and infinitely differentiable. Further, its derivatives are*

$$\nabla a(\theta) = \mathbf{E}_\theta\big[\phi(\mathbf{x})\big], \tag{2.2}$$

$$\nabla^2 a(\theta) = \mathbf{cov}_\theta\big[\phi(\mathbf{x})\big]. \tag{2.3}$$

PROOF. The proof of Lemma 2.2 actually shows a is convex. To complete the proof, we assume exchanging expectation and differentiation is permitted; verifying the validity of the assumption is done by a tedious argument

that appeals to the dominated convergence theorem. We defer the details to the appendix.

We exchange expectation and differentiation to obtain

$$\begin{aligned}\nabla_i a(\theta) &= \frac{\int_{\mathcal{X}} \phi_i(x) \exp\left(\theta^T \phi(x)\right) h(x) dx}{\int_{\mathcal{X}} \exp\left(\theta^T \phi(x)\right) h(x) dx} \\ &= \int_{\mathcal{X}} \phi_i(x) \exp\left(\theta^T \phi(x) - a(\theta)\right) h(x) dx \\ &= \mathbf{E}_\theta\big[\phi_i(\mathbf{x})\big]\end{aligned}$$

for any $i \in [p]$, which estabishes (2.2). Exchanging expectation and differentiation again,

$$\begin{aligned}\nabla^2_{i,j} a(\theta) &= \int_{\mathcal{X}} \phi_i(x)(\phi_j(x) - \nabla_j a(\theta)) \exp\left(\theta^T \phi(x) - a(\theta)\right) h(x) dx \\ &= \int_{\mathcal{X}} \phi_i(x)\phi_j(x) \exp\left(\theta^T \phi(x) - a(\theta)\right) h(x) dx \\ &\quad - \nabla_j a(\theta) \int_{\mathcal{X}} \phi_i(x) \exp\left(\theta^T \phi(x) - a(\theta)\right) h(x) dx \\ &= \mathbf{E}_\theta\big[\phi_i(\mathbf{x})\phi_j(\mathbf{x})\big] - \mathbf{E}_\theta\big[\phi_i(\mathbf{x})\big]\, \mathbf{E}_\theta\big[\phi_j(\mathbf{x})\big],\end{aligned}$$

for any $i, j \in [p]$, which estabishes (2.3). □

The superficial dimension of an exponential family may be reduced by re-parametrization in two cases:

1. $\mathcal{T} := \phi(\mathcal{X})$ is a subset of an affine set (in $\mathbf{R}^p$): there are $a \in \mathbf{R}^p$ and $b \in \mathbf{R}$ such that $a^T \phi(x) = b$ for any $x \in \mathcal{X}$.[1]
2. $\mathcal{H}$ is a subset of an affine set (in $\mathbf{R}^p$).

In the first case, two parameters $\eta_1, \eta_2 \in \mathcal{H}$ may correspond to the same density, making the model unidentifiable. Indeed, the densities that correspond to η and $\eta + a$ are proportional to each other:

$$\begin{aligned}f_{\eta+a}(x) &\propto \exp\left((\eta + a)^T \phi(x)\right) h(x) \\ &= \exp\left(\eta^T \phi(x) + b\right) h(x) \\ &\propto \exp\left(\eta^T \phi(x)\right) h(x).\end{aligned}$$

In fact, any parameter $\eta + \gamma a$ for some $\gamma \in \mathbf{R}$ is associated with the same density. In the second case, η obeys linear inequality constraints, the values

[1] An affine set is the set of solutions to a linear system: $\{x \in \mathbf{R}^n : Ax = b\}$ for some $A \in \mathbf{R}^{m \times n}$ and $b \in \mathbf{R}^m$.

of some parameters are determined by the values of other parameters. Thus some of the parameters are extraneous. If neither case holds, the exponential family is *minimal*.

DEFINITION 2.4. *An exponential family in canonical form is minimal if*

1. *the image of $\mathcal{X}$ under ϕ is not a subset of an affine set.*
2. *$\mathcal{H}$ is not a subset of an affine set.*

Minimal exponential families are classified into two types: full-rank and curved. A minimal exponential family is full-rank when its natural parameter space $\mathcal{H}$ contains an open set. Otherwise, a minimal exponential family is curved.

By the factorization theorem, $\phi(\mathbf{x})$ is a sufficient statistic for densities of the form

$$f_\eta(x) = \exp\left(\eta^T \phi(x) - a(\eta)\right) h(x).$$

If the exponential family is minimal, then $\phi(\mathbf{x})$ is a minimal sufficient statistic.

THEOREM 2.5. *If $\mathcal{F}$ is a minimal exponential family, then $\mathbf{t} = \phi(\mathbf{x})$ is a minimal sufficient statistic.*

PROOF. By Theorem 3.2, it suffices to show the ratio $\frac{f_\eta(x_1)}{f_\eta(x_2)}$ is constant in η if and only if $\phi(x_1) = \phi(x_2)$. If $\phi(x_1) = \phi(x_2)$, the ratio is

$$\frac{f_\eta(x_1)}{f_\eta(x_2)} = \exp\left(\eta^T(\phi(x_1) - \phi(x_2))\right) = 1.$$

Assume $\frac{f_\eta(x_1)}{f_\eta(x_2)}$ is constant in η. That is

$$\frac{f_{\eta_1}(x_1)}{f_{\eta_1}(x_2)} = \frac{f_{\eta_2}(x_1)}{f_{\eta_2}(x_2)} \text{ for any } \eta_1, \eta_2 \in \mathcal{H}.$$

We simplify to obtain

$$(\eta_1 - \eta_2)^T(\phi(x_1) - \phi(x_2)) = 0.$$

By the definition of minimal exponential family, $\mathcal{H}$ is not a subset of an affine set. Since subspaces are affine sets, Θ is not a subset of any subspace, which implies $\text{span}(\Theta) = \mathbf{R}^p$.

Since

$$\begin{aligned}&\left\{(\eta_1-\eta_2)^T\big(\phi(x_i)-\phi(x_i')\big)=0 \text{ for all } \eta_1,\eta_2\in\mathcal{H}\right\}\\ &\quad\iff \left\{\eta^T\big(\phi(x_i)-\phi(x_i')\big)=0 \text{ for all } \eta\in \operatorname{span}(\mathcal{H})\right\},\end{aligned}$$

we deduce

$$\eta^T\big(\phi(x_i)-\phi(x_i')\big)=0 \text{ for all } \eta\in\mathbf{R}^p,$$

which allows us to conclude $\phi(x_i)=\phi(x_i')=0$. □

The joint distribution of *i.i.d.* samples from a p-dimensional exponential family is another p-dimensional exponential family:

$$\begin{aligned}\prod_{i\in[n]} f_\theta(x_i) &= \prod_{i\in[n]} \exp\left(\eta(\theta)^T\phi(x_i)-a(\theta)\right)h(x_i)\\ &= \exp\left(\eta(\theta)^T\left(\textstyle\sum_{i\in[n]}\phi(x_i)\right)-a(\theta)\right)\prod_{i\in[n]}h(x_i)\end{aligned}$$

By the factorization theorem, $\sum_{i\in[n]}\phi(x_i)$ is a sufficient statistic. We remark that its dimension does not grow with the sample size. In fact, *i.i.d.* exponential families are the only parametric models that admit a sufficient statistic whose dimension does not depend on the sample size. The Pitman-Koopman-Darmois[2] theorem says that if $\{\mathbf{x}_i\}_{i\in[n]}$ are *i.i.d.* samples from a continuous density f_θ in a parametric model that consists of densities whose support does not depend on the parameter, then there is a sufficient statistic whose dimension does not depend on the sample size if and only if the parametric model is an exponential family.

APPENDIX A

LEMMA A.1. *If two functions f_1 and f_2 are*

1. *proportional to each other: $f_1(x)=cf_2(x)$ for any $x\in\mathcal{X}$*
2. *$\int_{\mathcal{X}} f_1(x)dx=\int_{\mathcal{X}} f_2(x)dx$,*

then they are equal: $f_1(x)=f_2(x)$ for any $x\in\mathcal{X}$.

PROOF. Indeed, by integrating $f_1(x)=cf_2(x)$ and rearranging, we have

$$\frac{\int_{\mathcal{X}} f_1(x)dx}{\int_{\mathcal{X}} f_2(x)dx}=c.$$

By the assumption that their integrals agree, we deduce $c=1$, which shows that $f_1(x)=f_2(x)$. □

[2]One of the namesakes of the theorem is Dr. Jim Pitman's father, Dr. Edwin Pitman.

LEMMA A.2. *Let $f_\theta(x)$ be a member of a canonical exponential family. We have*

$$\nabla_i z(\eta) = \int_{\mathcal{X}} \nabla_i \big[\exp(\eta^T \phi(x))\big] h(x) dx$$

at any $\eta \in \operatorname{int}(\mathcal{H})$.

PROOF. We show that exchanging differentiation and integration is justified at the origin.

By the definition of the derivative,

$$\begin{aligned}\nabla_i z(0) &= \lim_{n\to\infty} \frac{z(e_i\delta_n) - z(0)}{\delta_n} \\ &= \lim_{n\to\infty} \int_{\mathcal{X}} \frac{e^{(e_j\delta_n)^T\phi(x)} - 1}{\delta_n} h(x) dx,\end{aligned}$$

where $\{\delta_n\}$ is any decreasing sequence that converges to zero. We assume $d_0 := \sup_n \delta_n$ is small enough so that $e_j\delta_0 \in \operatorname{int}(\mathcal{H})$. We recognize the limit of the integrand is

$$\nabla_i \big[\exp(\eta^T \phi(x))\big] h(x).$$

To justify exchanging $\lim_{n\to\infty}$ and integration, we appeal to the dominated convergence theorem.

Let $d_0 := \sup_n \delta_n$. The (absolute value of the) integrand is at most

$$\begin{aligned}&\frac{\left|e^{(e_j\delta_n)^T\phi(x)} - 1\right|}{\delta_n} h(x) \\ &\quad\le \frac{e^{\left|(e_j\delta_n)^T\phi(x)\right|}\left|(e_j\delta_n)^T\phi(x)\right|}{\delta_n} h(x) && \left(|e^x - 1| \le e^{|x|}\,|x|\right) \\ &\quad\le \frac{e^{\left|(e_j\delta_0)^T\phi(x)\right|}\left|(e_j\delta_0)^T\phi(x)\right|}{\delta_0} h(x) \\ &\quad\le \frac{\exp^{2\left|(e_j\delta_0)^T\phi(x)\right|}}{\delta_0} h(x) && \left(|x| \le e^{|x|}\right) \\ &\quad\le \frac{\exp^{2\left|(e_j\delta_0)^T\phi(x)\right|} + \exp^{-2\left|(e_j\delta_0)^T\phi(x)\right|}}{\delta_0} h(x) && \left(e^{|x|} \le e^x + e^{-x}\right)\end{aligned}$$

We check that the bound is integrable:

$$\begin{aligned}&\int_{\mathcal{X}} \frac{\exp^{2\left|(e_j\delta_0)^T\phi(x)\right|} + \exp^{-2\left|(e_j\delta_0)^T\phi(x)\right|}}{\delta_0} h(x) dx \\ &\quad= z(2e_j\delta_0) + z(-2e_j\delta_0),\end{aligned}$$

which, by the assumption that $e_j\delta_0 \in \text{int}(\mathcal{H})$ is finite. Thus we are free to exchange $\lim_{n\to\infty}$ and integration to conclude

$$\begin{aligned}\nabla_i z(0) &= \lim_{n\to\infty} \int_{\mathcal{X}} \frac{e^{(e_j\delta_n)^T\phi(x)} - 1}{\delta_n} h(x)dx \\ &= \int_{\mathcal{X}} \lim_{n\to\infty} \frac{e^{(e_j\delta_n)^T\phi(x)} - 1}{\delta_n} h(x)dx \\ &= \int_{\mathcal{X}} \nabla_i \big[\exp(\eta^T\phi(x))\big]_0 h(x)dx.\end{aligned}$$

□

Yuekai Sun
Berkeley, California
November 29, 2015