STAT 460/560 Class 10: Sufficiency and Exponential Families

Daniel McDonald*

Reading: Chapter 9 Sections 9.13.2 and 9.13.3, [Was04]; Chapter 4.2, [van98].

1. Data Reduction and Sufficiency

There are two ways to motivate sufficient statistics: the data reduction viewpoint where we would like to discard non-informative pieces of the dataset (for storage or other benefits), and the risk reduction viewpoint where we want to construct estimators that only depend on meaningful variation in the data. We will focus on the former.

The goal in data reduction roughly is to find ways to reduce the size of a dataset without throwing away important information. We need to fix ideas more concretely to make progress on this abstract question.

Assuming a parametric model $X_1, \ldots, X_n \sim_{\text{IID}} p(x; \theta)$, our goal is to use the data to learn about θ. Suppose that we could reduce the data to a statistic T that contains all the available information in the sample for making inferences about θ.

A statistic $T = T(X_1, \ldots, X_n)$ is called *sufficient* for θ if the conditional distribution of the sample given T does not depend on θ. The is, if for all t in the range of T and all $(x_1, \ldots, x_n)$ such that $T(x_1, \ldots, x_n) = t$

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n \mid T(X_1, \ldots, X_n) = t; \theta)$$

does not depend on θ.

There is another more subtle question about what it means to condition on the sufficient statistic when it comes from a continuous distribution (since typically the conditioning set will have zero probability). It turns out that one can salvage this using something called the factorization theorem (which we will come to shortly). For now however, you should think about the discrete case and suppose that the probability that $T(X_1, \ldots, X_n) = t$ is strictly positive.

Example 10.1. Let $X_1, \ldots, X_n \sim_{\text{IID}} \text{Pois}(\theta)$. Consider $T = \sum_{i=1}^n X_i$. For values where $\sum x_i = t$, we have:

$$p_\theta(X_1 = x_1, \ldots, X_n = x_n \mid T = t) = \frac{p_\theta(X_1 = x_1, \ldots, X_n = x_n,\ T = t)}{p_\theta(T = t)}.$$

There are two possibilities, that the sum of the x_i's is equal to t and that the sum is not equal to t. In the latter case, the probability is zero so no need to consider it. In the former, we have:

$$\frac{p_\theta(X_1 = x_1, \ldots, X_n = x_n,\ T = t)}{p_\theta(T = t)} = \frac{p_\theta(X_1 = x_1, \ldots, X_n = x_n)\mathbf{I}(T = t)}{p_\theta(T = t)}$$

Since $T \sim \text{Poisson}(n\theta)$, this ratio evaluates to:

$$\frac{\prod_{i=1}^n \left(\frac{e^{-\theta}\theta^{x_i}}{x_i!}\right)}{\frac{e^{-n\theta}(n\theta)^t}{t!}} = \frac{e^{-n\theta}\theta^{\sum x_i}\mathbf{I}(T = t)(\sum x_i)!}{[\prod (x_i)!]e^{-n\theta}(n\theta)^{\sum x_i}} = \frac{\mathbf{I}(T = t)t!}{n^t[\prod_{i=1}^n (x_i)!]},$$

*Much of this material is adapted from Larry Wasserman's course notes for 36-705 at Carnegie Mellon University.

which does not depend on θ. Thus $T = \sum_{i=1}^n X_i$ is a sufficient statistic for θ.

You can similarly conclude that T/n or $3.7T$ are also sufficient statistics for θ. Furthermore, we can condition on more than T without losing information, so something like $(\sum_{i=1}^n X_i,\ X_1,\ X_2)$ is also sufficient, though somehow less "minimal". We'll come back to this idea of minimal sufficiency later.

Activity 10.1. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(\theta)$. Show that the sum $T = \sum_{i=1}^n X_i$ is sufficient.

Solution:

$$\begin{aligned} p(X_1 = x_1, \ldots, X_n = x_n \mid T = t) &= \frac{p_\theta(X_1 = x_1, \ldots, X_n = x_n,\ T = t)}{p_\theta(T = t)} \\ &= \frac{\mathbf{I}(T = t)\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\ &= \frac{\mathbf{I}(T = t)}{\binom{n}{t}}, \end{aligned}$$

which does not depend on θ.

It is more illustrative to think of sufficiency in terms of partitions of the sample space.

Example 10.2. Let $X_1, X_2, X_3 \sim \text{Bern}(\theta)$. Let $T = X_1 + X_2 + X_3$. The sample space is $\{0,1\}^3$ and the statistic T partitions the sample space into four sets:

- $T = 0$: $\{(0,0,0)\}$ — $p(x|T = 0) = 1$
- $T = 1$: $\{(1,0,0), (0,1,0), (0,0,1)\}$ — $p(x|T = 1) = 1/3$
- $T = 2$: $\{(1,1,0), (1,0,1), (0,1,1)\}$ — $p(x|T = 2) = 1/3$
- $T = 3$: $\{(1,1,1)\}$ — $p(x|T = 3) = 1$

The conditional distribution of X given T does not depend on θ, so T is sufficient.

Some notes about the partition view of sufficiency:

- A partition $B_1, \ldots, B_k$ is sufficient if $p(x|x \in B_i)$ does not depend on θ for any i.
- The statistic T induces a partition of the sample space into sets $\{x : T(x) = t\}$ for each possible value t of T. T is sufficient if and only if this partition is sufficient.
- Two statistics T_1 and T_2 can generate the same partition: say $\sum X$ and $3\sum X$.
- If we split a set in the partition into two or more subsets, the resulting finer partition is also sufficient. For example, $\sum X$ and $(\sum X, X_1)$ generate different partitions, but both are sufficient.

Activity 10.2. Consider the same setting as the previous example, but with $T = X_1$. Write down the partition of the sample space induced by T and show that it is not sufficient.

2. The Factorization Theorem

Checking sufficiency via conditional probabilities directly can be tedious, particularly for continuous variables. The *Fisher-Neyman Factorization Theorem* provides a simpler characterization.

Theorem 10.1 (Neyman-Fisher Factorization)**.** *A statistic $T(X_1, \ldots, X_n)$ is sufficient for θ if and only if the joint density/pmf factors as:*

$$p(x_1, \ldots, x_n; \theta) = h(x_1, \ldots, x_n) \cdot g(T(x_1, \ldots, x_n); \theta),$$

where h depends only on the data (and not θ), and g depends on θ and on the sample only through T.

Note that the theorem doesn't involve any conditional probabilities, and thus can be applied to continuous distributions as well. The proof is not difficult, but we will only give a sketch here.

Proof. (Sketch for the discrete case) We will show that the two conditions are equivalent.

- ($\Rightarrow$) If T is sufficient, set $g(t;\theta) = P_\theta(T = t)$ and $h(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n \mid T = t)$. The product yields $P_\theta(X_1 = x_1, \ldots, X_n = x_n)$.
- ($\Leftarrow$) If the density factors

$$\begin{aligned} P_\theta(X_1 = x_1, \ldots, X_n = x_n \mid T = t) &= \frac{P_\theta(X_1 = x_1, \ldots, X_n = x_n, T = t)}{P_\theta(T = t)} \\ &= \frac{\mathbf{I}(T = t)h(x_1, \ldots, x_n)g(t;\theta)}{\displaystyle\sum_{x_1,\ldots,x_n : T(x_1,\ldots,x_n)=t} h(x_1, \ldots, x_n)g(t;\theta)} \\ &= \frac{\mathbf{I}(T = t)h(x_1, \ldots, x_n)}{\displaystyle\sum_{x_1,\ldots,x_n : T(x_1,\ldots,x_n)=t} h(x_1, \ldots, x_n)}, \end{aligned}$$

  which does not depend on θ.

□

Exercise 10.1. Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. The joint density is:

$$p(x_1, \ldots, x_n; \mu, \sigma^2) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right).$$

Use the factorization theorem to find sufficient statistics for (μ, σ^2) in the following cases:

1. Both μ and σ^2 are unknown.
2. σ^2 is known.

Solution:

1. If both μ and σ^2 are unknown, expanding $\sum(x_i - \mu)^2 = \sum x_i^2 - 2\mu\sum x_i + n\mu^2$ shows that $T(X) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ (or equivalently $(\overline{X}, S^2)$) is joint sufficient for (μ, σ^2).
2. If σ^2 is known, we can rewrite the exponent using $\sum(x_i - \mu)^2 = \sum(x_i - \overline{x})^2 + n(\overline{x} - \mu)^2$. The term involving $(x_i - \overline{x})$ doesn't depend on μ, so it can be absorbed into $h(x)$. Thus, $T(X) = \overline{X}$ is a sufficient statistic for μ.

3. Sufficiency and the Likelihood Function

By the Factorization Theorem, if there is a sufficient statistic T, then

$$L(\theta) = h(x) \cdot g(T(x);\theta) \propto g(T(x);\theta).$$

Thus, the likelihood function depends on the observed data solely through any sufficient statistic $T(x)$.

Thus, once we have any sufficient statistic, we have everything we need to compute the likelihood function. If we were to base our subsequent data analysis only on the likelihood function then we have lost nothing. We will see a further connection between sufficiency and the likelihood function once we define the notion of minimal sufficiency.

4. Minimal Sufficiency

As we have seen previously, sufficient statistics are not unique. Furthermore, it seems, at least intuitively, that some sufficient statistics provide much more reduction than others. For example, suppose that $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$. Then $\overline{X}_n$ is a sufficient statistic. But so is the whole data set. This motivates the following definition of minimal sufficient statistics:

A statistic $T(x_1, \ldots, x_n)$ is minimal sufficient if it is sufficient, and, furthermore, for any other sufficient statistic $S(x_1, \ldots, x_n)$ we can write $T(x_1, \ldots, x_n) = g(S(x_1, \ldots, x_n))$, i.e. T is a function of S.

Theorem 10.2. *Define*

$$R(x_1, \ldots, x_n, y_1, \ldots, y_n; \theta) = \frac{p(x_1, \ldots, x_n; \theta)}{p(y_1, \ldots, y_n; \theta)}.$$

Suppose T has the property that $R(x_1, \ldots, x_n, y_1, \ldots, y_n; \theta)$ is constant in θ if and only if $T(x_1, \ldots, x_n) = T(y_1, \ldots, y_n)$. Then T is minimal sufficient.

Example 10.3. Let $X_1, \ldots, X_n \sim \text{Pois}(\theta)$.

$$R(x_1, \ldots, x_n, y_1, \ldots, y_n; \theta) = \frac{\prod_{i=1}^n \frac{\theta^{x_i} e^{-\theta}}{x_i!}}{\prod_{i=1}^n \frac{\theta^{y_i} e^{-\theta}}{y_i!}} = \frac{\theta^{\sum x_i - \sum y_i}}{\prod x_i! / \prod y_i!}$$

This is constant in θ if and only if $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$. Thus, $\sum_{i=1}^n X_i$ is a minimal sufficient statistic.

A minimal sufficient statistic is not unique. But, the minimal sufficient *partition* is unique.

Be careful, sometimes the only minimal sufficient statistic is the whole data set.

There is an important sense in which sufficient statistics lead to better estimators. Suppose that T is sufficient for θ, and that $\hat{\theta}$ is an estimator of θ. Then, the estimator $\hat{\theta}^* = \mathbb{E}[\hat{\theta} \mid T]$ is also an estimator of θ.

Theorem 10.3 (Rao-Blackwell)**.** *If T is sufficient for θ, and $\hat{\theta}$ is an estimator of θ, then $\hat{\theta}^* = \mathbb{E}[\hat{\theta} \mid T]$ is also an estimator of θ. Furthermore,*

$$MSE(\hat{\theta}^*) \leq MSE(\hat{\theta}).$$

Example 10.4. Suppose $X_1, \ldots, X_n \sim \text{Unif}(0, \theta)$. Then $T = \max(X_1, \ldots, X_n)$ is sufficient for θ. Consider the estimator $\hat{\theta} = 2\overline{X}_n$. Note that $\overline{X}_n$ is not a function of T, that $\mathbb{E}[\hat{\theta}] = \theta$ and that $\mathbb{E}[T] < \theta$.

Then, define $\hat{\theta}^* = \mathbb{E}[2\overline{X}_n \mid T] = \frac{n+1}{n} T$ is a function of T, has the same expectation as $\hat{\theta}$, and has variance less than or equal to that of $\hat{\theta}$.

Proof.

$$\text{MSE}(\hat{\theta}^*) = \mathbb{E}[(\mathbb{E}[\hat{\theta}^* \mid T] - \theta)^2] = \mathbb{E}[(\mathbb{E}[\hat{\theta}^* - \theta \mid T])^2] \leq \mathbb{E}[\mathbb{E}[(\hat{\theta} - \theta)^2 \mid T]] = \text{MSE}(\hat{\theta}),$$

where the inequality follows from Jensen's inequality. □

5. Exponential Families

A family $\{P_\theta\}$ of distributions forms an s-dimensional exponential family if the distributions P_θ have densities of the form:

$$p(x; \theta) = \exp\left[\sum_{i=1}^s \eta_i(\theta) T_i(x) - A(\theta)\right] h(x),$$

where η_i, A are functions which map θ to $\mathbb{R}$. You can check that $(T_1(x), \ldots, T_s(x))$ is a sufficient statistic. The term $A(\theta)$ is known as the log-normalization constant or the log-partition function. Let $\mathcal{X}$ be the set of possible values of x.

Remark: You can ignore this: As a technical note, exponential families can be defined with respect to the Lebesgue measure (as we did implicitly above) or with respect to any other measure (for instance, the discrete measure on $\{1, \ldots, k\}$). We will continue to simply think of $\mathcal{X}$ as a subset of $\mathbb{R}$ and the measure as the Lebesgue measure.

Although thinking of the above form is standard, it is usually much more convenient to parametrize the distribution in what is known as its *canonical parametrization*, where we simply take $\eta_i(\theta)$ to be the parameters. In this case, we can more compactly write:

$$p(x;\theta) = \exp\left[\sum_{i=1}^{s} \theta_i T_i(x) - A(\theta)\right] h(x).$$

In this case, we refer to θ as the *natural parameters* of the distribution. Notice that none of these parametrizations are unique, we can replace T_i by cT_i and θ_i by θ_i/c and obtain the same distribution.

The term $A(\theta)$ is what makes the distribution integrate to 1, i.e.

$$A(\theta) = \log\left[\int_{\mathcal{X}} \exp\left[\sum_{i=1}^{s} \theta_i T_i(x)\right] h(x)dx\right].$$

The set of θ's for which $A(\theta) < \infty$ constitute the natural parameter space.

Several distributions you have or will encounter are exponential family distributions (Wikipedia has a long list). We will do a couple of examples here.

Example 10.5. The Normal family of distributions has density,

$$p(x;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2}\right),$$

which is a 2-parameter exponential family, with natural parameters $(\theta_1, \theta_2) = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)$, and sufficient statistics (x, x^2). One can verify that the natural parameter space is $\mathbb{R} \times (-\infty, 0)$.

Discrete distributions can similarly belong to an exponential family (you have to replace all the integrals with sums and so on).

Exercise 10.2. The Binomial distribution (with n fixed)has pmf,

$$p(x;\theta) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x \in \{0, 1, \ldots, n\}.$$

Show that this is an exponential family and identify its sufficient statistic and natural parameter.
Solution:

We can re-write this as:

$$p(x;\theta) = \binom{n}{x} \exp\left(x \log\frac{p}{1-p} + n\log(1-p)\right), \quad x \in \{0, 1, \ldots, n\}.$$

This shows that it is in an exponential family with sufficient statistic x (number of successes), and natural parameter,

$$\theta = \log\frac{p}{1-p}.$$

Exercise 10.3. Show that the Pois(λ) is an exponential family and identify its sufficient statistic and natural parameter.

Wikipedia has a long list of exponential family distributions, their natural parameters, sufficient statistics and other useful information. It is good practice to try to derive the natural parameters for some popular distributions.

6. Random sampling and exponential families

The exponential family structure is preserved for an i.i.d. sample, i.e. if $\{X_1, \ldots, X_n\}$ are i.i.d from some exponential family distribution $p(x;\theta)$ then the joint distribution:

$$p(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} h(x_i) \exp\left[\sum_{i=1}^{s} \theta_i \sum_{j=1}^{n} T_i(x_j) - nA(\theta)\right],$$

is in an exponential family with the same natural parameters but with sufficient statistics:

$$T_i(x_1, \ldots, x_n) = \sum_{j=1}^{n} T_i(x_j).$$

7. Log-partition generates moments

Recall that,

$$A(\theta) = \log\left[\int_{\mathcal{X}} \exp\left[\sum_{i=1}^{s} \theta_i T_i(x)\right] h(x)dx\right],$$

so taking the derivatives of A with respect to θ we obtain that,

$$\frac{\partial A(\theta)}{\partial \theta_i} = \frac{\int_{\mathcal{X}} T_i(x) \exp\left[\sum_{i=1}^{s} \theta_i T_i(x)\right] h(x)dx}{\int_{\mathcal{X}} \exp\left[\sum_{i=1}^{s} \theta_i T_i(x)\right] h(x)dx} = \mathbb{E}[T_i(X)].$$

You might wonder why we can switch derivatives and integrals - this is done rigorously using the dominated convergence theorem. Similarly, you can easily verify that higher derivatives lead to (functions of) higher moments (technically cumulants and not moments), i.e.

$$\frac{\partial^2 A(\theta)}{\partial \theta_i \partial \theta_j} = \mathbb{E}\left[(T_i(X) - \mathbb{E}[T_i(X)])(T_j(X) - \mathbb{E}[T_j(X)])\right] = \mathrm{Cov}(T_i(X), T_j(X)).$$

This is why the function $A(\theta)$ is classically known as the *cumulant function*.

This latter property also reveals that A is a convex function of θ, i.e. it is bowl-shaped. Convexity is implied by the fact that the second-derivative matrix (i.e. the Hessian matrix) is positive semi-definite. For exponential families, the Hessian matrix is the covariance matrix of the sufficient statistics T_i, and covariance matrices are always positive semi-definite.

Remember the conclusion: A is a convex function of θ.

8. The likelihood function in exponential families

When we observe a random sample $X_1, \ldots, X_n \sim p(x;\theta)$ from an exponential family distribution, the log-likelihood function is:

$$\ell(\theta; x_1, \ldots, x_n) \propto \left[\sum_{i=1}^{s} \theta_i \sum_{j=1}^{n} T_i(x_j) - nA(\theta)\right].$$

The log-likelihood function in an exponential family is concave. To see this compute the Hessian of $L(\theta; x_1, \ldots, x_n)$ and observe that this is $-n$ times the Hessian of A. Since A is convex, its negation is concave.

9. Bregman Divergences and KL Divergences

Given a (strictly) convex function A we can define a divergence between parameters by:

$$\rho(\theta_1, \theta_2) = A(\theta_2) - A(\theta_1) - \langle \nabla A(\theta_1), \theta_2 - \theta_1 \rangle.$$

For a pair of distributions we can define the KL divergence (assuming everything below is finite):

$$\mathrm{KL}(p, q) = \int p(x) \log(p(x)/q(x)) dx.$$

It is easy to see that for exponential families – the Bregman divergence between parameters (using the log-partition as the convex function) is exactly equal to the KL divergence between the corresponding distributions.

10. Parameter Estimation

Recall maximum likelihood: choose the estimate $\widehat{\theta}$ to be the value of θ that maximizes the likelihood function. We have seen that the log-likelihood in an exponential family is concave and given by

$$\ell(\theta; x_1, \ldots, x_n) \propto \left[\sum_{i=1}^{s} \theta_i \sum_{j=1}^{n} T_i(x_j) - nA(\theta) \right],$$

so we can simply take the derivative with respect to θ and set this equal to 0. Using the facts we have seen earlier about the derivative of A, we can see that this amounts to solving the following system of equations for θ:

$$\mathbb{E}_\theta[T_i(X)] = \frac{1}{n} \sum_{j=1}^{n} T_i(x_j) \quad \text{for } i \in \{1, \ldots, s\}$$

where

$$\mathbb{E}_\theta[T_i(X)] = \int T_i(x) p_\theta(x) dx.$$

So the maximum likelihood estimator simply picks the parameters θ to match the empirical expectations of the sufficient statistics to the expected value of the sufficient statistics under the distribution. Since exponential families have concave likelihoods this is usually a tractable problem. What is more, for exponential families as we can see above these two methods of estimation coincide.