36-705: Intermediate Statistics **Fall 2019**

Lecture 13: September 27

Lecturer: Siva Balakrishnan

Before we turn our attention to estimation more formally, we are going to take one final detour through what are called *exponential families*. Exponential family distributions possess many useful and pleasant properties, and give us a somewhat unified way to think about "nice" distributions.

Good references for this material (since it has not been taught in the last few versions of this course) include Chapter 3 of Martin Wainwright and Michael Jordan's monograph on exponential families, and Lehmann and Casella's Theory of Point Estimation book.

13.1 Exponential Families - Canonical Parametrization

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 $\mathbb{R}$, and the $T_i(x)$ are known as the *sufficient statistics* (it should be clear to you why this is). The term $A(\theta)$ is known as the log-normalization constant or the log-partition function (the former terminology will be clearer in a second). We will assume that $x \in \mathcal{X}$, where $\mathcal{X}$ is just some set.

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 θs 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 1: 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).

Example 2: The Binomial distribution has pmf,

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

We can re-write this as:

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

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

$$\theta = \log\left(\frac{p}{1-p}\right).$$

Example 3: The Poisson(λ) distribution has pmf,

$$p(x;\theta) = \frac{\exp(-\lambda)\lambda^x}{x!} = \frac{1}{x!}\exp(x\log\lambda - \lambda),$$

which shows that it is an exponential family with sufficient statistic x, and natural parameter $\theta = \log(\lambda)$.

If you have not seen these before, 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.

13.2 Properties of Exponential Families

13.2.1 Random sampling

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:

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

13.2.2 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,

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

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 but you can look up the distinction), i.e.

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

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. If this did not make sense ignore it but remember the conclusion: *A is a convex function of θ.*

13.2.3 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 simply:

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

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

13.2.4 Minimal representations and minimal sufficiency

An exponential family representation is said to be minimal if the sufficient statistics are not redundant, i.e. there is no set of coefficients $a \in \mathbb{R}^s, a \neq 0$ such that,

$$\sum_{i=1}^{s} a_i T_i(x) = \text{const},$$

for all $x \in \mathcal{X}$. If the representation is not minimal then essentially one can eliminate some of the sufficient statistics from the representation to obtain a minimal representation. Non-minimal exponential families are sometimes called *over-complete* exponential families. Over-complete exponential families are not statistically identifiable (while minimal ones are), i.e. there can be two different parameter vectors $\theta^1 \neq \theta^2$, such that, $p(X;\theta^1) = p(X;\theta^2)$. This effectively means, even if I gave you infinite data from the model, you cannot meaningfully estimate the parameter θ.

An exponential family where the spaced of allowed parameters θ_i is s-dimensional is called a full-rank family. On the other hand if there are relationships between the θ_i (for instance, $\theta_2 = \theta_1^2$) then the exponential family is *curved.*. For a full-rank exponential family, the sufficient statistics turn out to be minimal sufficient, i.e. the statistic

$$T(X_1, \ldots, X_n) = \left(\sum_{i=1}^{n} T_1(X_i), \ldots, \sum_{i=1}^{n} T_s(X_i)\right),$$

is minimal sufficient.

13.2.5 The mean parameterization

You should skip this section unless exponential families really piqued you curiosity (in which case, you should really read some of the references).

We have been discussing the canonical parametrization of exponential families. It turns out that an equivalent way to parameterize the distribution is via what are called its mean parameters. We will not show this equivalence (it is not difficult) but rather just introduce the terminology here.

Suppose we define:

$$\mu_i = \mathbb{E}[T_i(X)] = \int_{x \in \mathcal{X}} T_i(x) \exp\left[\sum_{i=1}^{s} \eta_i(\theta) T_i(x) - A(\theta)\right] h(x) dx,$$

then it turns out that the collection $(\mu_1, \ldots, \mu_s)$ is in 1-1 correspondence with the natural parameters of the exponential family.

The estimation problem (i.e. given samples from $p(X;\theta)$ trying to figure out θ) can be viewed as trying to find the natural parameters given the mean parameters. More broadly, if you take a graphical models class you might learn that many tasks in exponential families (computing probabilities etc.) can be framed as trying to map between natural and mean parameters.

13.2.6 The maximum entropy duality

The classical motivation for exponential families comes from what is called the *principle of maximum entropy.* The idea is that, we suppose that we are given a random sample $\{X_1, \ldots, X_n\}$ from some distribution, and we compute the empirical expectations of certain functions that we choose:

$$\widehat{\mu}_i = \frac{1}{n} \sum_{j=1}^{n} T_i(X_j) \quad \text{for} \;\; i \in \{1, \ldots, s\}.$$

For simplicity, you could imagine the case when $T(X) = (X, X^2, \ldots, X^s)$, i.e. where the statistics we are interested in are just moments, but everything we are discussing is much more general. Based on just these empirical expectations we want to infer a full probability distribution on the samples. A distribution p is *consistent* with the data we observe if it is the case that,

$$\widehat{\mu}_i = \mathbb{E}_p[T_i(X)] \quad \text{for} \;\; i \in \{1, \ldots, s\}.$$

We of course would like to pick a consistent distribution. It turns out that in most interesting cases, if we constrain a small number of statistics in this fashion there are infinitely many consistent distributions, so we need to come up with a way to choose between them.

The principle of maximum entropy suggests to pick the distribution that has the largest (Shannon) entropy. The entropy of a distribution is:

$$H(p) = -\int_{x \in \mathcal{X}} p(x) \log(p(x)) dx.$$

Roughly, the entropy measures the complexity of a distribution (i.e. the average number of bits needed to encode samples from a distribution). The principle of maximum entropy says that one should be "maximally agnostic" about all aspects of the distribution that are not explicitly constrained. If this does not make sense, then just think about the principle as giving a possibly "natural" way to choose a distribution from a collection.

So we could imagine trying to find the distribution p^* that,

$$p^* = \arg\max_p H(p)$$

subject to the constraints that,

$$\widehat{\mu}_i = \mathbb{E}_p[T_i(X)] \quad \text{for} \quad i \in \{1, \ldots, s\}.$$

The solution to this problem can be computed using the calculus of variations, and is always an exponential family distribution, i.e. there exist some parameters θ such that the distribution p^* has the form:

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

In this case, the θ_i are what are called Lagrange parameters. They are equivalent to the maximum likelihood estimates for the parameters of this distribution (we will see what this means in a little bit).

The main take home is that an alternate way to think about exponential families, is that they arise naturally from trying to constrain a few simple statistics of a distribution using the data and then choosing a distribution that maximizes the entropy subject to those constraints.

13.2.7 Bregman Divergences and KL Divergences

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

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

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.

13.2.8 Parameter Estimation - Maximum Likelihood and the Method of Moments

This will be something we will much more slowly, but lets try to understand the main ideas here in the context of exponential families.

One of the dominant strategies of parameter estimation is to compute a value of the parameter that maximizes the likelihood of the observed data. We have seen that the likelihood in an exponential family is concave and given by

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

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:

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

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.

Usually we cannot compute this estimator in closed form so we use an iterative algorithm (like gradient ascent) to maximize the likelihood. However, you should remember that exponential families have concave likelihoods so this is usually a tractable endeavour (at least for simple enough families).

The alternative way to estimate parameters of a distribution is known as the method of moments. Here the idea is to pick some statistics of the data, and the try to find parameters for your distribution so that the empirical average of the statistics are equal to their expected values under the estimated model. For exponential families as we can see above these two methods of estimation coincide.