Exponential families

Patrick Breheny

September 27

Introduction

- We now turn to the middle part of this course, where we will take these tools that we have learned apply them to prove various theoretical properties of likelihood
- For the most part, we will try to make as few assumptions as possible about the probability model we are using
- However, the theoretical properties of likelihood turn out to be particularly simple and straightforward if the probability model falls into a class of models known as exponential families
- Today we will cover the idea behind exponential families, see why they are particularly convenient for likelihood, and discuss some extensions of the family

History

- First, a bit of history
- In the 19th and early 20th centuries, statistical theory and practice was almost exclusively focused on classical parametric models (normal, binomial, Poisson, etc.)
- Starting in the 1930s (but taking a long time to be fully appreciated), it became apparent that all of these parametric models have a common construction (the exponential family) and unified theorems can be obtained that apply to all of them
- In fact, as we will see today, this is not an accident – *only* exponential families enjoy certain properties of mathematical and computational simplicity

Exponential tilting

- Suppose we have the "standard" Poisson distribution ($\mu = 1$):

$$p_0(x) = e^{-1}/x!;$$

  how can we go about constructing a family of distributions, all using this as a starting point?
- Consider forming new distributions via *exponential tilting*:

$$\tilde{p}(x|\theta) = p_0(x)e^{\theta x}$$

- This isn't a proper distribution, hence the notation $\tilde{p}(x|\theta)$, but it would be if we determined the normalizing constant, which I will denote $\exp\{\psi(\theta)\}$, and divide:

$$p(x|\theta) = p_0(x)e^{\theta x - \psi(\theta)}$$

Poisson example

- Let's see how all this plays out for the Poisson distribution
- First, the normalizing constant:

$$\psi(\theta) = e^{\theta} - 1$$

- The family of distributions is therefore

$$p(x|\theta) = \exp\{x\theta - e^{\theta}\}/x!,$$

or in terms of the usual Poisson parameterization,

$$p(x|\theta) = \mu^x e^{-\mu}/x!,$$

where $\theta = \log \mu$

Remarks on tilting

- Here we "tilted" the reference distribution p_0 by $e^{\theta x}$, although note that the tilting parameter did not turn out to be the same as the "usual" parameter we would think of
- A similar phenomenon can happen for the observation x; some distributions are formed not by tilting with x itself but rather with a function $s(x)$
- Thus, in what follows, I will assume that we have tilted by $e^{s\theta}$, and for simplicity I will suppress the dependency of s on x in the notation as we could just as easily think of having observed s directly

Single parameter exponential family

A one-parameter exponential family therefore has the form

$$p(x|\theta) = \exp\{s\theta - \psi(\theta)\}p_0(x),$$

where

- θ is the *natural parameter*
- s is the *natural statistic*
- $\psi(\theta)$ is the *cumulant generating function*, for reasons that we will discuss shortly
- p_0 is the base or reference distribution, although it need not be a proper distribution; for example, our Poisson derivation would have been simpler if we had chosen $p_0(x) = 1/x!$

Cumulant generating functions

- The *cumulant generating function* is simply the log of the moment generating function
- Like moment generating functions, cumulant generating functions yield the moments of a distribution, but unlike MGFs, yield central moments:
  - Its derivative evaluated at zero is the mean
  - Second derivative evaluated at zero is the variance
  - Higher order derivatives yield quantities related to the skewness, kurtosis, etc.

ψ and cumulants

- Note that for a distribution in the exponential family, the moment generating function of the random variable $s(X)$ is

$$M(t) = \int e^{ts} e^{s\theta} p_0(x)\, dx / e^{\psi(\theta)}$$
$$= e^{\psi(t+\theta)} / e^{\psi(\theta)}$$

- Thus, its cumulant generating function is $\psi(t+\theta) - \psi(\theta)$, although for moment-finding purposes, we can simply treat ψ itself as the cumulant generating function (i.e., its derivatives still generate the desired cumulants)

Mean and variance

- In particular,

$$\mathbb{E}(S) = \dot{\psi}(\theta)$$
$$\mathbb{V}(S) = \ddot{\psi}(\theta)$$

- Note that these expressions provide the mean and variance of the natural statistic (which may or may not be the mean and variance of X)

Multi-parameter exponential families

- All of these concepts extend in a straightforward way to the d-parameter exponential family:

$$p(x|\boldsymbol{\theta}) = \exp\{\mathbf{s}^\top \boldsymbol{\theta} - \psi(\boldsymbol{\theta})\} p_0(x)$$

- For example, the Gamma distribution is a 2-parameter exponential family:

$$p(x|\alpha, \beta) = \exp\{\alpha \log \beta - \log \Gamma(\alpha) + \alpha \log x - \beta x\}/x$$

or, in terms of $\boldsymbol{\theta} = [-\beta, \alpha], s = [x, \log x]$:

$$p(x|\boldsymbol{\theta}) = \exp\{\mathbf{s}^\top \boldsymbol{\theta} - [\log \Gamma(\theta_2) - \theta_2 \log(-\theta_1)]\}$$

Mean and variance

Analogous to the one-parameter case, we have

$$\mathbb{E}(\mathbf{s}) = \nabla\psi(\boldsymbol{\theta})$$
$$\mathbb{V}(\mathbf{s}) = \nabla^2\psi(\boldsymbol{\theta}),$$

where $\mathbb{E}(\mathbf{s})$ is a $d \times 1$ vector and $\mathbb{V}(\mathbf{s})$ is a $d \times d$ variance-covariance matrix

Repeated sampling

- Why are we interested in exponential tilting as opposed to some other way of generating new distributions from a base distribution?
- Let's consider what happens in the case of repeated sampling, where $x_1, \ldots, x_n \overset{\text{iid}}{\sim} p(x|\boldsymbol{\theta})$:

$$\begin{aligned} p(\mathbf{x}|\boldsymbol{\theta}) &= \prod_{i=1}^{n} \exp\{\mathbf{s}_i^\top \boldsymbol{\theta} - \psi(\boldsymbol{\theta})\} p_0(x_i) \\ &= \exp\{n[\bar{\mathbf{s}}^\top \boldsymbol{\theta} - \psi(\boldsymbol{\theta})]\} p_0(\mathbf{x}), \end{aligned}$$

where $\bar{\mathbf{s}} = \sum \mathbf{s}_i / n$

Sufficiency

- In other words, the joint distribution of the repeated sample is still in the same exponential family, just scaled up by a factor of n
- In particular, a quick look at the factorization theorem will show that $\mathbf{s}$ is a sufficient statistic for the exponential family
- Under repeated sampling, we easily obtain $\bar{\mathbf{s}}$ as a sufficient statistic
- Thus, no matter how large the sample, we can always reduce the information it contains down into a d-dimensional vector of means

Pitman-Darmois-Koopmans Theorem

- As it turns out, *only* exponential families have this property, in which the sufficient statistic remains of fixed dimension under repeated sampling
- This result was shown for one-dimensional exponential families by Fisher, who originally introduced the concepts of sufficiency and exponential tilting
- Later, a trio of authors working independently in different countries extended this result to multiparameter families; the result is known as the Pitman-Darmois-Koopmans theorem

Likelihood

- Furthermore, exponential families are particularly convenient in terms of their likelihood
- In particular, the log-likelihood of any exponential family is simply $n[\bar{\mathbf{s}}^\top \boldsymbol{\theta} - \psi(\boldsymbol{\theta})]$ plus a constant, so its gradient is

$$\nabla \ell(\boldsymbol{\theta}|\mathbf{x}) \propto \bar{\mathbf{s}} - \nabla\psi(\boldsymbol{\theta})$$

and

$$\hat{\boldsymbol{\theta}} = (\nabla\psi)^{-1}(\bar{\mathbf{s}})$$

Example: Poisson

- Returning to the Poisson distribution, where $s = x$ and $\psi(\theta) = e^\theta$, we have

$$\dot{\psi}(\theta) = e^\theta$$

and

$$\hat{\theta} = \log \bar{x}$$

- The inverse is not always so mathematically tractable, however: for example in the gamma distribution, $\nabla\psi(\boldsymbol{\theta})$ involves the digamma function, whose inverse is not available in closed form

Central limit theorem

- Furthermore, since the MLE is simply a function of the mean in exponential families, it is particularly easy to derive its limiting distribution
- Letting $\boldsymbol{\mu} = \mathbb{E}(\mathbf{s})$, the central limit theorem tells us that

$$\sqrt{n}(\bar{\mathbf{s}} - \boldsymbol{\mu}) \xrightarrow{\mathrm{d}} \mathrm{N}(\mathbf{0}, \mathbf{V}),$$

  where $\mathbf{V} = \nabla^2 \psi(\boldsymbol{\theta})$
- Thus, letting $\mathbf{g}$ denote the transformation $\boldsymbol{\theta} = \mathbf{g}(\boldsymbol{\mu})$, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{\mathrm{d}} \mathrm{N}(\mathbf{0}, \nabla \mathbf{g}(\boldsymbol{\mu})^\top \mathbf{V} \nabla \mathbf{g}(\boldsymbol{\mu}))$$

  by the delta method; keep in mind here that $\nabla \mathbf{g}$ and $\mathbf{V}$ are both $d \times d$ matrices

Application to the Poisson case

- In the Poisson case, $\ddot{\psi}(\theta) = e^{\theta} = \mu$ and $g(\mu) = \log \mu$, so $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathrm{d}} \mathrm{N}(0, e^{-\theta})$
- Thus, $\hat{\theta} \pm 1.96\sqrt{e^{-\hat{\theta}}/n}$ is an approximate 95% confidence interval for θ, which we could transform to get a confidence interval for μ

Remarks

- The maximum likelihood estimator is asymptotically normal not only in exponential families, but in a much wider class of models
- Specifically, we require only that the likelihood is a "smooth" function of θ, in a sense that we will discuss later
- We'll go into more details regarding likelihood-based inference, confidence intervals, tests, etc., soon

Introduction

- Until now, we have assumed that the dimension of $\boldsymbol{\theta}$ and $\mathbf{s}$ was the same as the number of unknown parameters
- However, it can also be the case that the parameter space $\boldsymbol{\Theta}$ is constrained somehow; for example if $\boldsymbol{\theta}$ is a function of $\boldsymbol{\beta}$, with $\dim(\boldsymbol{\beta}) = k < d$
- In such cases the exponential family is no longer said to be "full" or "full rank"

Curved vs flat exponential families

- How large an impact this makes on likelihood-based inference depends on whether the function $\boldsymbol{\theta}(\boldsymbol{\beta})$ is linear ("flat") or not ("curved")
- If there is a matrix $\mathbf{M}$ such that $\boldsymbol{\theta} = \mathbf{M}\boldsymbol{\beta}$, then

$$\begin{aligned}\exp\{\mathbf{s}^\top\boldsymbol{\theta} - \psi(\boldsymbol{\theta})\} &= \exp\{\mathbf{s}^\top\mathbf{M}\boldsymbol{\beta} - \psi(\mathbf{M}\boldsymbol{\beta})\} \\ &= \exp\{\tilde{\mathbf{s}}^\top\boldsymbol{\beta} - \tilde{\psi}(\boldsymbol{\beta})\}\end{aligned}$$

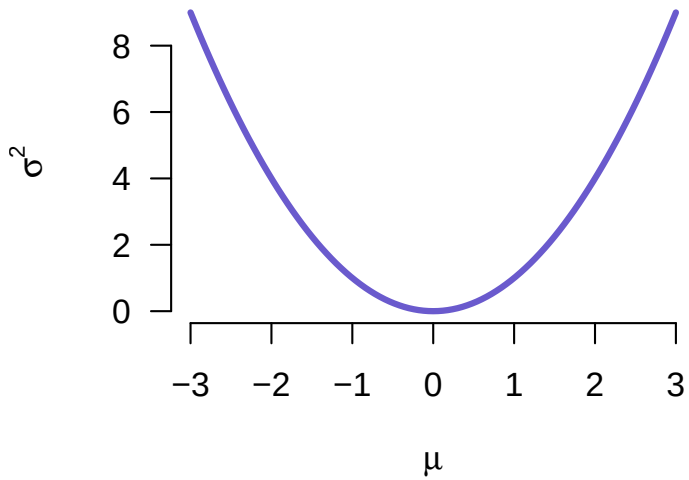
  in other words, we still have a regular exponential family, albeit with reduced rank $k < d$, new summary statistics $\tilde{\mathbf{s}}$, and a new normalizing function $\tilde{\psi}$
- If $\boldsymbol{\theta}(\boldsymbol{\beta})$ is a nonlinear function, however, things can be much more complicated

Example: Regression

- Flat exponential families come up quite often in regression models, especially generalized linear models
- For example, we might observe $Y_i \overset{\perp\!\!\!\perp}{\sim} \text{Pois}(\theta_i)$, but impose a model $g(\theta_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$, which restricts $\boldsymbol{\Theta}$ to a lower-dimensional subspace of $\mathbb{R}^n$
- If the systematic component of our model is $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$ (i.e., we assume a linear model with respect to the natural parameters), then our exponential family is not curved
- In the GLM literature, this is known as the *canonical* link

Example: Normal, known coefficient of variation

- As a simple example of a curved exponential family, suppose $x \sim \mathrm{N}(\mu, c^2\mu^2)$, where c, the coefficient of variation, is known
- The natural parameter and statistic are 2-dimensional, but there is only one unknown parameter
- The parameter space forms a one-dimensional line curving through $\mathbb{R}^2$:



Example: Empirical Bayes

- Curved exponential families also arise commonly in *empirical Bayes* procedures
- Here, we again observe something like $Y_i \overset{\perp\!\!\!\perp}{\sim} \text{Pois}(\theta_i)$, but instead of a fixed linear model, impose a distribution on the natural parameters $\theta_i \overset{\perp\!\!\!\perp}{\sim} g(\boldsymbol{\beta})$; this is typically done when estimation of θ_i using only Y_i is noisy or unstable, and we wish to "borrow information" from other observations
- We won't delve deeply into the theory of curved exponential families in this course, but will note that they enjoy many, but not all, of the theoretical properties of full-rank exponential families

Definition

- A variation on exponential tilting, and one that is often very useful in statistical modeling, is to introduce a *dispersion parameter* and tilt by $\exp\{\mathbf{s}^\top \boldsymbol{\theta}/\phi\}$
- The resulting model is then of the form

$$p(x|\boldsymbol{\theta}, \phi) = \exp\left\{\frac{\mathbf{s}^\top \boldsymbol{\theta} - \psi(\boldsymbol{\theta})}{\phi}\right\} p_0(x, \phi)$$

- Note that the normalizing constant is now $\exp\{\psi(\boldsymbol{\theta})/\phi\}$

Mean and variance

- The primary motivation for doing this is to allow the variance to be parameterized separately from the mean
- Specifically,

$$\mathbb{E}(\mathbf{s}) = \nabla\psi(\boldsymbol{\theta}) = \boldsymbol{\mu}$$
$$\mathbb{V}(\mathbf{s}) = \phi\nabla^2\psi(\boldsymbol{\theta}) = \phi\mathbf{V}(\boldsymbol{\mu});$$

you will derive these results in the next homework assignment

Example: Poisson distribution

- In practice, the normalizing quantity $p_0(x, \phi)$ is often left unspecified (or rather, implicitly specified)
- For example, by introducing a dispersion parameter into the Poisson model, we now have the useful result that $\mathbb{V}(X) = \phi\mu$; instead of requiring that the variance equals the mean, we can instead allow the model to accommodate over- or under-dispersion
- However, $p_0(x, \phi)$ is the function that satisfies

$$\sum_{x=0}^{\infty} \exp\left\{\frac{x\theta - e^{\theta}}{\phi}\right\} p_0(x, \phi) = 1;$$

not so trivial to find

Estimation

- Note that this does not actually affect estimation of $\boldsymbol{\theta}$, since we still have $\hat{\boldsymbol{\theta}} = (\nabla\psi)^{-1}(\bar{\mathbf{s}})$
- However, it does have two meaningful implications for modeling:
  - We cannot find the MLE of ϕ
  - We cannot compute likelihood ratios
- In practice, one typically uses some other estimation strategy, such as method of moments, to obtain $\hat{\phi}$

Inference

- Its impact on likelihood-based inference, however, is not so trivial to remedy
- In practice, what is often done is to simply replace ϕ with $\hat{\phi}$ in the likelihood and treat the likelihood as though $\hat{\phi}$ were a known constant rather than an unknown parameter
- This approach, which goes by a variety of names (approximate likelihood, estimated likelihood, pseudo-likelihood) often works reasonably well, but it must be noted that by treating an unknown quantity as a known one, we are biasing our inference towards being overconfident (confidence intervals too narrow)