Stat 5421 Lecture Notes

Proper Conjugate Priors for Exponential Families

Charles J. Geyer

March 28, 2016

1 Theory

This section explains the theory of conjugate priors for exponential families of distributions, which is due to Diaconis and Ylvisaker (1979).

1.1 Conjugate Priors

A family of distributions $\mathcal{P}$ is *conjugate* to another family of distributions $\mathcal{M}$ if, when applying Bayes rule with the likelihood for $\mathcal{M}$, whenever the prior is in $\mathcal{P}$, the posterior is also in $\mathcal{P}$.

The way one finds conjugate families is to make their PDF look like the likelihood.

For an exponential family with canonical statistic y, canonical parameter θ, and cumulant function c, the likelihood is

$$L(\theta) = e^{\langle y, \theta \rangle - c(\theta)}. \tag{1}$$

More generally, the likelihood for sample size n from this family (see Section 3 of Geyer, 2016a) is

$$L(\theta) = e^{\langle y, \theta \rangle - n c(\theta)}. \tag{2}$$

where now y is the canonical statistic for sample size n, what Section 3 of Geyer (2016a) writes as $\sum_{i=1}^n y_i$, and θ and c are the same as before, but now n appears.

Saying we want the prior to "look like" (2) means we want it to be a function that looks like it (considered as a function of θ). But, of course, a prior cannot depend on data. So we have to replace y by something known, and, while we are at it, we also replace n too.

Thus we say that

$$h_{\eta,\nu}(\theta) = e^{\langle \eta, \theta \rangle - \nu c(\theta)}, \qquad \theta \in \Theta, \tag{3}$$

defines a function $h_{\eta,\nu}$ that is an unnormalized conjugate prior probability density function (PDF) for θ, where η is a vector of the same dimension as the canonical statistic and canonical parameter, ν is a scalar, and Θ is the full canonical parameter space for the family given by equation (7) in Geyer (2016a).

The quantities η and ν are called *hyperparameters* of the prior, to distinguish them from θ which is the parameter we are making Bayesian inference about. They are treated as known constants, different choices of which determine which prior we are using. We don't treat η and ν as parameters the way Bayesians treat parameters (put priors on them, treat them as random). Instead we treat η and ν the way frequentists treat parameters (as non-random constants, different values of which determine different probability distributions).

1.2 Corresponding Posteriors

The "make the prior look like the likelihood" trick is not guaranteed to work. It depends on what the likelihood looks like. So we check that it does indeed work for exponential families.

Bayes rule can be expressed as

$$\text{unnormalized posterior} = \text{likelihood} \times \text{unnormalized prior}. \tag{4}$$

If we apply this with (2) as the likelihood and (3) as the unnormalized prior, we obtain

$$e^{\langle y, \theta \rangle - n c(\theta)} e^{\langle \eta, \theta \rangle - \nu c(\theta)} = e^{\langle y + \eta, \theta \rangle - (n + \nu) c(\theta)}$$

which is a distribution in the conjugate family with vector hyperparameter $y + \eta$ and scalar hyperparameter $n + \nu$. So it does work. This is indeed the conjugate family.

In general, there is no simple expression for the normalized PDF of the conjugate family, so we still have to use Markov chain Monte Carlo (MCMC) to do calculations about the posterior.

1.3 The Philosophy of Conjugate Priors

What is the point of conjugate priors? Suppose we started with a flat prior, which is the special case of the conjugate family with $\eta = 0$ (the zero vector) and $\nu = 0$ (the zero scalar). Then no matter how much data we ever collect, our posterior will be some distribution in the conjugate family. (If we start in the conjugate family, we stay in the conjugate family.)

The Bayesian learning paradigm says that the posterior distribution incorporating past data serves as the prior distribution for future data. So this suggests that any prior based on data should be a conjugate prior.

But, of course, the whole argument depends on starting with a flat prior. If we don't start with a prior in the conjugate family, then we don't (in general) get a posterior distribution in the conjugate family.

But this argument does suggest that conjugate priors are one kind of prior that is reasonable for the model. For example, they have the kind of tail behavior that could have come from observation of data from the model.

This is something that just using, for example, normal prior distributions does not do. The normal distribution has tails that decrease more rapidly than any other widely used distribution, and a lot faster than conjugate priors for some discrete exponential families. Consider a family with bounded support of the canonical statistic, for example, logistic regression. The log of (3) has derivative

$$\nabla \log h_{\eta,\nu}(\theta) = \eta - \nu \nabla c(\theta) = \eta - \nu E_\theta(y)$$

using equation (5) in Geyer (2016a). The point is that since y is bounded, so is $E_\theta(y)$, bounded considered as a function of θ, that is. And that means the logarithmic derivative of the conjugate prior is bounded. And that means the conjugate prior PDF has tails that decrease no more than exponentially fast. But the normal distribution has tails that decrease superexponentially fast (like $\exp(-\|\beta\|^2)$, where $\|\cdot\|$ denotes the Euclidean norm). So normal priors are more informative than conjugate priors (have lighter tails, much lighter when far out in the tails). They express "uncertainty" about the parameter that has more "certainty" than could reflect what is learned from any amount of data. Something wrong there (philosophically).

In summary, when you use conjugate priors, they are guaranteed to be something that could reflect uncertainty that comes from actual data. Other priors pulled out of nowhere do not have this property.

1.4 Proper Priors

The fundamental theorem about conjugate priors for exponential families is Theorem 1 in Diaconis and Ylvisaker (1979), which we repeat here.

Theorem 1 (Diaconis-Ylvisaker)**.** *The conjugate prior* (3) *determined by* η *and* ν *is proper if and only if* $\nu > 0$ *and* η/ν *lies in the interior of the convex support of the exponential family, which is the smallest convex set that contains the canonical statistic of the family with probability one.*

So what does that last bit mean? A set S in a vector space is *convex* if it contains all convex combinations of its points, which in turn are defined to be points of the form $\sum_{i=1}^{k} p_i x_i$, where x_i are points and p_i are scalars that are nonnegative and sum to one.

The "nonnegative and sum to one" should ring a bell. It is the condition for probability mass functions. So another way to explain convexity is to say that S is convex if the mean of every probability distribution concentrated on a finite set of points of S is contained in S.

Corollary 2. *If a conjugate prior for the canonical parameter of exponential family is proper, then the parameterization is identifiable.*

Proof. By Theorem 1 in Geyer (2009), which also is Theorem 1 in Geyer (2016a), the canonical parameterization is identifiable if and only if the convex support has the same dimension as the canonical statistic and parameter vectors. So if the parameterization is not identifiable, then the convex support lies in a lower-dimensional subspace and has empty interior, but by Theorem 1 above, that implies the prior is not proper. □

The converse to this corollary is that, if we want a proper prior, then we had better have an identifiable parameterization.

1.4.1 Saturated Models

To get a more concrete picture, let us calculate the convex supports for a few simple probability models.

Bernoulli Regression First consider the saturated model for Bernoulli regression. The the components y_i of the response vector y are Bernoulli. Each y_i has support $\{0, 1\}$. Hence y has support $\{0, 1\}^n$, where n is the dimension of y. The convex support must fill in all the points in between, so it is $[0, 1]^n$. The interior of the convex support is the points inside, not on the surface, so that is $(0, 1)^n$. Here we are using the convention that square brackets for an interval indicate that the end points are included, and round brackets indicate that the end points are excluded. To be absolutely clear the interior of the convex support is

$$(0, 1)^n = \{\, y \in \mathbb{R}^n : 0 < y_i < 1 \text{ for all } i \,\}. \tag{5}$$

Poisson Regression Poisson regression is similar. The components y_i of the response vector y are Poisson. Each y_i has support $\mathbb{N}$ (the set of nonnegative integers). Hence y has support $\mathbb{N}^n$, where n is the dimension of y. The convex support is thus $[0, \infty)^n$, and the interior of the convex support is

$$(0, \infty)^n = \{\, y \in \mathbb{R}^n : 0 < y_i \text{ for all } i \,\}.$$

Multinomial, Try III The multinomial distribution is a bit different. If y is a multinomial random vector of dimension k for sample size n, then the components of y are nonnegative integers that sum to n. Unlike the case for Bernoulli or Poisson regression, the components of y are not independent and the support is not a Cartesian product. But we just said what it was (components are nonnegative integers that sum to n). In math notation that is

$$\Bigl\{\, y \in \mathbb{N}^k : \textstyle\sum_{i=1}^k y_i = n \,\Bigr\},$$

and the convex support must fill in points in between

$$\Bigl\{\, y \in [0, \infty)^k : \textstyle\sum_{i=1}^k y_i = n \,\Bigr\},$$

and the interior of the convex support is empty because the convex support is contained in a lower-dimensional subspace and hence is not a neighborhood of any point.

Hence there can be no proper conjugate prior if we take y to be the canonical statistic of the family, what Geyer (2016a, Section 2.4.3) calls the "Try III" parameterization of the multinomial.

Multinomial, Try II If instead we use the "Try II" parameterization (Geyer, 2016a, Section 2.4.2) in which the canonical statistic has components y_1, y_2, $\ldots$, y_{k-1} (omitting the count for the last category), then the support of the canonical statistic vector is

$$\Bigl\{\, y \in \mathbb{N}^{k-1} : \textstyle\sum_{i=1}^{k-1} y_i \le n \,\Bigr\},$$

(less than or equal to n because we have dropped y_k from the sum), and the interior of convex support is

$$\Bigl\{\, y \in (0, \infty)^{k-1} : \textstyle\sum_{i=1}^{k-1} y_i < n \,\Bigr\}.$$

1.4.2 Canonical Affine Models

Theorem 1 also applies to canonical affine submodels because they too are full exponential families (Geyer, 2016a, Section 4.6). The only difference is that the canonical statistic vector has the form $M^T y$ where M is the model matrix and y is the canonical statistic vector of the saturated model.

It is usually hard to calculate the convex support of $M^T y$ even if the convex support of y is known. Thus it is usually hard to know the set of all hyperparameters that yield proper conjugate priors. Hence the following theorem, which, at least, allows us to identify some proper conjugate priors.

Theorem 3. *If η / ν is a vector in the relative interior of the convex support of a full exponential family and M is the model matrix for a canonical affine submodel that has identifiable canonical parameterization, then $M^T \eta$ and ν are hyperparameters for a proper conjugate prior for the submodel canonical parameter.*

To understand this we need to know what "relative interior" is. It is the interior relative to the smallest affine subspace containing the set (Rockafellar, 1970, Chapter 6).

Again, we can understand this concept by examples. If the convex support of the saturated model has the same dimension as the canonical statistic, as we saw was the case for Bernoulli regression, Poisson regression, and the multinomial distribution with "Try II" parameterization, then the relative interior is the same as the interior.

Only with the multinomial distribution with "Try III" parameterization did we have a lower-dimensional convex support which yields an empty interior. That is where the notion of relative interior is useful. It is what we might think is the interior if we weren't careful. For the multinomial it is what we get if we replace $\le$ by $<$ in the formula for the convex support obtaining

$$\left\{ y \in (0, \infty)^k : \textstyle\sum_{i=1}^k y_i = n \right\}.$$

Proof of Theorem 3. By Theorem 6.6 in Rockafellar (1970) if η / ν is in the relative interior of the convex support saturated model, then $M^T \eta / \nu$ is in the relative interior of convex support the submodel. By Theorem 1 in Geyer (2009) the relative interior of the convex support of the submodel is equal to the interior if and only if the submodel canonical parameterization is identifiable. □

2 Examples

2.1 Logistic Regression

2.1.1 Saturated Model

This example follows an example that is just R code (Geyer, 2016b) where is it said that adding 1/2 to the count in each cell of a product binomial (logistic regression) model has the same effect as using a proper prior.

The log likelihood for the saturated model in terms of ordinary parameters is

$$l(p) = \sum_{i=1}^{n} \bigl[y_i \log(p_i) + (n_i - y_i) \log(1 - p_i) \bigr] \tag{6}$$

(we have a reason for starting with ordinary parameters rather than canonical parameters — we will get to them).

By adding 1/2 to each cell of the table, what was meant was to add 1/2 to each of the y_i and to each of the $n_i - y_i$, which is clearly what the cited computer code does. Let us generalize writing ε rather than 1/2 and allowing it to be any positive real number.

Then the unnormalized log posterior is

$$l(p) = \sum_{i=1}^{n} \bigl[(y_i + \varepsilon) \log(p_i) + (n_i - y_i + \varepsilon) \log(1 - p_i) \bigr]. \tag{7}$$

From (4) we get

$$\text{log unnormalized posterior} = \text{log likelihood} + \text{log unnormalized prior}$$

from which we see that the log unnormalized prior is (7) minus (6), which is

$$\sum_{i=1}^{n} \bigl[\varepsilon \log(p_i) + \varepsilon \log(1 - p_i) \bigr]$$

is the log unnormalized prior we are using. Now we introduce canonical parameters

$$p_i = \frac{e^{\theta_i}}{1 + e^{\theta_i}}$$

so the log unnormalized prior becomes

$$\begin{aligned} h(\theta) &= \sum_{i=1}^{n} \left[\varepsilon \log \left(\frac{e^{\theta_i}}{1 + e^{\theta_i}} \right) + \varepsilon \log \left(\frac{1}{1 + e^{\theta_i}} \right) \right] \\ &= \sum_{i=1}^{n} \left[\varepsilon \theta_i - 2 \varepsilon \log \left(1 + e^{\theta_i} \right) \right] \end{aligned} \tag{8}$$

and we have now put the prior in exponential family form. The vector hyperparameter η is the vector having all components equal to ε and $\nu = 2\varepsilon$ (compare with the form of the log likelihood for logistic regression in Section 5.1 of Geyer, 2016a).

So does this prior satisfy the conditions to be proper given in Theorem 1? Or, to be more precise, would it be proper if we were using the saturated model?

Since we said $\varepsilon > 0$, we clearly have $\nu = 2\varepsilon > 0$. That's one part. Then η / ν is the vector having all components equal to $1/2$. And this is indeed a point in (5), so this does give a proper prior.

2.1.2 Canonical Affine Submodel

When we go to canonical affine submodels, Corollary 2 assures us that adding 1/2 to each cell of the table gives a proper conjugate prior for the submodel canonical parameter.

2.2 Poisson Regression

Now we turn to our other example, which is homework problem 3-2. In the statement of the problem, it is claimed that adding 1/2 to the count for each cell of the contingency again results in a proper conjugate prior. As we shall see, this claim is wrong.

Now the saturated model log likelihood is

$$l(\theta) = \langle y, \theta \rangle - c(\theta),$$

where

$$c(\theta) = \sum_{i=1}^{n} e^{\theta_i}$$

(Geyer, 2016a, equation (40)).

It is now clear from Theorem 1 that a proper conjugate prior has the form (3) with η having components $\eta_i > 0$ and also with $\nu > 0$.

And the latter requirement is what adding $1/2$ to each component of y leaves out. That gives an unnormalized log posterior of

$$\langle y + 1/2, \theta \rangle - c(\theta) = \langle y + \eta, \theta \rangle - (1 + \nu) c(\theta)$$

with $\eta_i = 1/2$ for all i and $\nu = 0$.

We see that our unnormalized prior for the saturated model is just

$$g(\theta) = e^{\langle \eta, \theta \rangle}$$

and applying this to the canonical affine submodel gives

$$g(\beta) = e^{\langle M^T \eta, \beta \rangle}$$

and this cannot possibly integrate to something finite because it goes to infinity as β goes to infinity in some direction (which depends on what M is).

If we wanted a proper prior, its log unnormalized density would have to have the form

$$h(\beta) = \langle M^T \eta, \beta \rangle - \nu c(a + M\beta)$$

with all components of η strictly positive and $\nu > 0$.

References

Geyer, C. J. (2009). Likelihood inference in exponential families and directions of recession. *Electronic Journal of Statistics*, **3**, 259–289.

Geyer, C. J. (2016a). Exponential families, part I. Lecture notes for Stat 5421 (categorical data analysis). `http://www.stat.umn.edu/geyer/5421/notes/expfam.pdf`.

Geyer, C. J. (2016b). Untitled R script for Stat 5421 (categorical data analysis). `http://www.stat.umn.edu/geyer/5421/examp/mcmc-too.Rout`.

Diaconis, P., and Ylvisaker, D. (1979). Conjugate priors for exponential families. *Annals of Statistics*, **7**, 269–281.

Rockafellar, R. T. 1970. *Convex Analysis.* Princeton University Press, Princeton, NJ.