Stat 5421 Lecture Notes: Proper Conjugate Priors for Exponential Families

Charles J. Geyer

November 23, 2022

Contents

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 $\mathcal{Q}$ if, when applying Bayes rule with the likelihood for $\mathcal{Q}$, whenever the prior is in $\mathcal{P}$, the posterior is also in $\mathcal{P}$.

There may be more than one conjugate prior family. The family of all distributions is always a (nonparametric) conjugate prior family. The question is whether there is a more tractable prior family.

The usual 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

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

(notes on exponential families). More generally, the likelihood for sample size n from this family is

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

(notes on exponential families) where now y is the canonical statistic for sample size n, what the notes on exponential families just cited 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 may be an unnormalized probability density function (UPDF) 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 in notes on exponential families.

We say a nonnegative function h is a UPDF provided it integrates to something that is nonzero and finite, in which case

$$f(x) = \frac{h(x)}{\int h(x)\, dx}$$

is a proper probability density function (PDF), meaning f is nonnegative and integrates to one. Any function like (3) that is never zero can never integrate to zero. So (3) is a UPDF if and only if it integrates to something finite. More on this below.

Even if (3) does not integrate to something finite and hence is not a UPDF, one can still think of it as an improper prior, which when combined with the likelihood using Bayes' rule (following section) may yield a proper posterior. The notes on Bayes and MCMC warn against using improper priors. But they are widely used.

One of the problems with using improper priors is that they may lead to improper posteriors. So we need conditions on when priors and posteriors are proper.

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

$$e^{\langle y, \theta \rangle - n c(\theta)} e^{\langle \eta, \theta \rangle - \nu c(\theta)} = e^{\langle y + \eta, \theta \rangle - (n + \nu) c(\theta)} \tag{5}$$

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. The Metropolis-Hastings-Green algorithm specifies target distributions by UPDF. It does not need proper PDF. See Sections 1.12.1, 1.12.4, 1.17.3, and 1.17.4 in Geyer (2011).

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 equations relating derivatives of cumulant functions to mean and variance of the canonical statistic (notes on exponential families). 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 finite amount of data, no matter how much. 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 and Posteriors

"Proper" in the section heading means an actual probability distribution, not an improper prior or (worse) an improper posterior, which is nonsense.

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

Theorem 1.1. *For a regular full exponential family, 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.*

Corollary 1.1. *For a regular full exponential family, the conjugate prior (3) determined by η and ν is proper if and only if $\nu > 0$, the canonical parameterization is identifiable, and η/ν is a possible value of the mean value parameter.*

Proof. This combines Theorem 1.1 and the characterization of the relative interior of the convex support in Theorem 9.2 in Barndorff-Nielsen (1978) and the characterization of identifiability in Theorem 1 in Geyer (2009). □

Of course, this theorem and corollary also apply to conjugate posterior distributions too. If the unnormalized posterior is (5), then this corresponds to a proper posterior if and only if $n + \nu > 0$ and $(y + \eta)/(n + \nu)$ lies in the interior of the convex support.

So what does convex support 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.

The corollary says for regular full exponential families with identifiable canonical parameterizations, the interior of the convex support is the set of all possible mean value parameter vectors.

1.5 Canonical Affine Submodels

Theorem 1.1 and Corollary 1.1 also apply to canonical affine submodels because they too are regular full exponential families if the saturated model is regular and full (notes on exponential families).

Let us recall that theory. The relation between submodel and saturated model canonical parameters is

$$\theta = a + M\beta$$

where a is the offset vector and M is the model matrix. Then the (submodel) log likelihood for β is

$$l(\beta) = \langle M^T y, \beta \rangle - c_{\text{sub}}(\beta)$$

where

$$c_{\text{sub}}(\beta) = c(a + M\beta) \tag{6}$$

and where y is the saturated model canonical statistic and c is the saturated model cumulant function. From this we see that

- $M^T y$ is the submodel canonical statistic,
- β is the submodel canonical parameter, and
- c_{sub} is the submodel cumulant function.

Also useful is the identity

$$\langle M^T y, \beta \rangle = \langle y, M\beta \rangle$$

So to apply the Diaconis-Ylvisaker theory to the canonical affine submodel we introduce a vector hyperparameter η and a scalar hyperparameter ν and the conjugate family has log unnormalized densities of the form

$$\langle \eta, \beta \rangle - \nu c_{\text{sub}}(\beta)$$

and Corollary 1.1 says that such a distribution is proper if and only if η/ν is a possible mean value of $M^T y$.

But that is difficult to determine, see the section about images and preimages in the vignette for R package `rcdd` Geyer *et al.* (2021).

Thus we offer a method that is much easier to use. This uses log unnormalized prior densities of the form

$$\langle M^T \eta, \beta \rangle - \nu c_{\text{sub}}(\beta) \tag{7}$$

and Corollary 1.1 says that such a distribution is proper if and only if $M^T\eta/\nu$ is a possible mean value of $M^T y$. But by linearity of expectation $E(M^T Y) = M^T E(Y)$ so such a distribution is proper if and only if η/ν is a possible mean value of y. And the latter is easy to determine.

Since $M^T\eta$ ranges over possible values of $M^T E(Y)$ as η ranges over possible values of $E(Y)$, this method entails no loss of generality. It gives the same family of proper priors as the other method. And we claim it is more natural, as the prior is made up by imagining hypothetical data for the response vector.

We state all of the reasoning above as a corollary.

Corollary 1.2. *When log unnormalized priors for a canonical affine submodel have the form (7), the prior is proper if and only if η/ν is a possible value of the mean value parameter vector for the saturated model, $\nu > 0$, and the submodel has an identifiable canonical parameterization. Moreover, if the prior is proper, then so is the posterior.*

The last statement of the corollary is a triviality. It true of all Bayesian inference and has nothing to do with exponential families. One only needs to check that the posterior is proper when an improper prior has been used.

2 Examples

2.1 Poisson Sampling

The log likelihood for the saturated model for Poisson sampling is

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

where

$$c(\theta) = \sum_{i \in I} e^{\theta_i}$$

To get a conjugate prior we have to use the general form (7), where c_{sub} is given by (6), and the vector hyperparameter η and the scalar hyperparameter ν satisfy the conditions of Corollary 1.2.

R function `glm` always arranges for the submodel canonical parameterization to be identifiable by dropping some coefficients from the model, if necessary, dropped coefficients being indicated by having `NA` estimates.

2.2 Product Multinomial Sampling

Suppose we have a contingency table, assume product multinomial sampling, and have a canonical affine submodel.

Then one way to get a conjugate prior, which your humble author calls the "method of made-up data" is to add a positive quantity to each component of the response vector (the positive quantities do not have to be integer-valued).

Let's check that.

The log likelihood for the saturated model for product multinomial sampling is

$$l(\pi) = \sum_{i \in I} y_i \log(\pi_i)$$

The proposal we are checking is to say that we take as the log unnormalized posterior

$$h(\pi) = \sum_{i \in I} (y_i + \eta_i) \log(\pi_i)$$

The log unnormalized prior is the part that was added to the log likelihood, that is,

$$h(\pi) = \sum_{i \in I} \eta_i \log(\pi_i)$$

And this clearly has the form of the conjugate prior for the saturated model with vector hyperparameter η having components η_i and scalar hyperparameters $\nu = \sum_{i \in A} \eta_i$, $A \in \mathcal{A}$, where $\mathcal{A}$ is the partition for the product multinomial sampling scheme. This is because of how product multinomial works, the sample sizes for the likelihood are $n_A = \sum_{i \in A} y_i$, so we get the analogous quantities for the sample sizes for the conjugate prior (we don't have to plod through the math). And then η_A/ν_A is a possible mean value. And that finishes checking this example. Invoking Corollary 1.2 says if we are OK for the saturated model, then we are OK for the canonical affine submodel too (provided the submodel canonical parameterization is identifiable).

This example includes multinomial sampling and logistic regression and multinomial response regression, all of which are special cases of product multinomial sampling. In the logistic regression case, that means we add positive quantities to both successes and failures, not just to successes.

2.3 Volleyball

The volleyball example in the notes on Bayes and MCMC claims to use the method of made-up data but is a little bit different from the general product multinomial sampling example discussed above.

In those notes (on Bayes and MCMC) not only do we make up some data but also that made-up data involves made-up teams.

In the priors for the β_i, we "imagine that each real team had one win and one loss against this imaginary team whose ability parameter β is fixed at zero".

Thus we have to consider our data structure to be enlarged to include this imaginary team. That team also has wins and losses, but it has zero wins and losses in the real data. It only has wins and losses in the made-up data that induces the prior.

Moreover, our priors for the β_i do not involve the home court advantage parameter γ. So these imaginary games against the imaginary team must have occurred (in our imagination) at a neutral site.

Then in our prior for γ, we imagine two more imaginary teams, different from the one we already imagined. These are "two teams of equal ability (equal β's) played two games, and the home team won one and lost one".

So again we have to consider our data structure to be enlarged to include these two imaginary teams. Those teams also have wins and losses, but they have zero wins and losses in the real data. They only have wins and losses in the made-up data that induces the prior.

But if we accept this notion about made-up data for imaginary new cells in our contingency table, the theory of conjugate priors explained above still works. As long as ν and the components of η are positive we will have a proper prior.

In the volleyball example, we first used priors involving one made-up game for each team to give priors for the β_i and then two more made-up games to give the prior for γ. This is clearly a proper prior by the logic explained for general product multinomial sampling.

In the volleyball example, at the end we used a hierarchical prior.

This takes us out of the theory of conjugate priors, although we still used that theory in part of the argument. The total prior is factored into priors and hyperpriors. (That is what hierarchical Bayes does.) and the priors are conjugate priors as before. But the hyperpriors are not conjugate. That's just the way hierarchical Bayes works.

Bibliography

Barndorff-Nielsen, O. E. (1978) *Information and Exponential Families*. Chichester, England: Wiley.

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

Geyer, C. J. (2009) Likelihood inference in exponential families and directions of recession. *Electronic Journal of Statistics*, **3**, 259–289. DOI: 10.1214/08-EJS349.

Geyer, C. J. (2011) Introduction to MCMC. In *Handbook of Markov Chain Monte Carlo* (eds S. P. Brooks, A. E. Gelman, G. L. Jones, et al.), pp. 3–48. Boca Raton: Chapman & Hall/CRC. Available at: https://www.mcmchandbook.net/HandbookChapter1.pdf.

Geyer, C. J., Meeden, G. D. and Fukuda, K. (2021) *R Package* `rcdd`*: Computational Geometry, Version 1.4*. Available at: http://cran.r-project.org/package=rcdd.