Exponential Families and Conjugate Priors

Aleandre Bouchard-Côté

March 14, 2007

1 Exponential Families

Inference with continuous distributions present an additional challenge compared to inference with discrete distributions: how to represent these continuous objects within finite-memory computers? A common solution to this problem is to use a (much smaller) subset (or family) of distributions instead of all possible distributions and to adopt a *parametrization* that identifies elements of this family with elements of $\mathbb{R}^d$. For instance, normal distributions can be characterized with their mean and variance; Poisson distributions can be characterized by their mean, etc.

An important family of distribution that has special properties with respect to statistical inference is the *exponential family*, introduced by Pitman (father), Darmois and Koopman. As a preview, here are some important properties of the exponential family that explain their central role in statistics:

- Suppose $X_1, X_2, \ldots$ are iid with a distribution known to be in some family of probability distribution whose support does not vary. Only if that family is an exponential family is there a sufficient statistic $\vec{T}(X_1, ..., X_n)$ whose number of scalar components does not increase as the sample size n increases. This is a very attractive result for those interested in efficient statistical inference.
- The exponential family arises naturally as the the answer to the following question: what is the most "uninformative" distribution with given constraints on expected values.
- Exponential families always have *conjugate priors.*

This document will concentrate on the last property, but first, we will review the definition of exponential families.

Definition 1 *Fix a reference measure, say the Lebesgue measure on $\mathbb{R}^m$ for concreteness, and let S be a family of continuous distribution. It is an exponential family if there are:*

1. *an integer $k \in \mathbb{Z}$ (the dimensionality of the parametrization),*

2. a function $T : \mathbb{R}^m \to \mathbb{R}^k$ (the sufficient statistics),

3. a function $A : \mathbb{R}^k \to \mathbb{R}$ (the log normalization) and

4. a function $h : \mathbb{R}^m \to \mathbb{R}$,[1]

such that for all $F \in S$, there is an $\eta \in \mathbb{R}^k$ such that F can be expressed with respect to the Lebesgue measure as the density:

$$f(\vec{x}) = h(\vec{x}) \exp\Big(\eta^T T(\vec{x}) - A(\eta)\Big).$$

2 Conjugate priors for exponential families

Recall that a family of prior probability distributions $p(\theta)$ is said to be conjugate to a family of likelihood functions $p(x|\theta)$ if the resulting posterior distributions $p(\theta|x)$ are in the same family as $p(\theta)$. In the case where the likelihood functions happen to be an exponential family, there is a general recipe for finding a conjugate prior.

Warning: in what follows, do not rely on which greek letter or symbol I use to represent sufficient statistics and natural parameters. The reason is that two different exponential families will be considered, one for the likelihood and one for the prior, and the sufficient statistics of the likelihood will be used to construct the natural parameters of the exponential family for the prior. The notation is the only difficulty here.

Theorem 1 *Let X be the $\mathbb{R}^m$-valued data and Θ, the $\mathbb{R}^k$-valued parameters. Assume that the conditional distribution of X given θ is in an exponential family with sufficient statistics $T : \mathbb{R}^m \to \mathbb{R}^k$, log-normalization $A : \mathbb{R}^k \to \mathbb{R}$ and has natural parameters $\psi(\theta) \in \mathbb{R}^k$. Then, the $(k+1)$-parameter exponential family with sufficient statistics $(\psi_1(\theta), \psi_2(\theta), \ldots, \psi_k(\theta), A(\theta))$ and the same base measure as the likelihood's, is a conjugate prior, where $\psi_i(\theta)$ is the i-th component of the natural parametrization of the likelihood model's exponential family. Note that this fully specifies an exponential family, as the normalization for this new exponential family can be obtained by integration.*

An important observation to make is that this recipe does not always yields a conjugate prior that is computationally tractable (in particular, there is no guarantee that it can be written only with elementary functions). Indeed, computing posterior parameters requires evaluation of the sufficient statistics $(\psi_1(\theta), \psi_2(\theta), \ldots, \psi_k(\theta), A(\theta))$ of the parameters. Note that this requires in turn evaluation of the normalization of the likelihood model, which is often hard in practice. For example, in the Ising model this problem is in #P, but can be approximated using probabilistic algorithms such as MCMC.

Note finally that families of conjugate prior are not unique. For instance, the set of all probability distribution is always a conjugate prior, as well as mixtures

[1] Which could be absorbed in the reference measure actually, since it is not allowed to depend on the parameters.

of conjugate priors. However, the one we constructed is minimal (in terms of number of dimension of them minimal smooth parametrization).

3 References

- Wikipedia, http://en.wikipedia.org/wiki/Exponential_family
- Mathematical Statistics, P. Bickel, K. Doksum