Notes on exponential family distributions and generalized linear models

Andreas Vlachos

May 3, 2010

1 Exponential family distributions

1.1 Formulations

Exponential family is a class of distributions that all share the following form:

$$p(y|\eta) = h(y) \exp\{\eta^T T(y) - A(\eta)\} \tag{1}$$

- η is the *natural parameter*, (a.k.a. *exponential parameter*).[1] For a given distribution (e.g. Bernoulli), η specifies all the parameters needed for that distribution.
- $T(y)$ is the *sufficent statistic* of the data (in many cases $T(y) = y$, in which case the distribution is said to be in *canonical form* and η is referred to as the *canonical parameter*).
- $A(\eta)$ is the *log-partition function* (a.k.a. *normalization factor*, *cumulant generating function*) which ensures that $p(y|\eta)$ remains a probability distribution.
- $h(y)$ is the non-negative *base measure* (in many cases it is equal to 1).

Note that since η contains all the parameters needed for a particular distribution in its original form, we can express it with respect to the mean parameter θ:

$$p(y|\theta) = h(y) \exp\{\eta(\theta)^T T(y) - A(\eta(\theta))\} \tag{2}$$

Examples of distributions that are exponential families: Gaussian, multionmial, exponential, Dirichlet, Poisson, Gamma...

Examples of distributions that are *not* exponential families: Cauchy, uniform...

Let's see how the Bernoulli distribution can be converted into the exponential family form:

[1] There is also a more general version of the exponential family form in which the natural parameter is defined as a function over η (Dobson book, Clark and Thayer primer).

$$\begin{aligned} p(y|\theta) &= \theta^y (1-\theta)^{1-y} \\ &= \exp\{y \log \theta + (1-y) \log(1-\theta)\} \\ &= \exp\{y \log \frac{\theta}{1-\theta} + \log(1-\theta)\} \end{aligned} \tag{3}$$

Then we define:

- natural parameter $\eta(\theta) = log \frac{\theta}{1-\theta}$
- log-partition function $A(\eta(\theta)) = \log(1 + exp(\eta(\theta))) = \log(1 + \frac{\theta}{1-\theta}) = -log(1-\theta)$
- sufficient statistic $T(y) = y$
- base measure $h(y) = 1$

It is important to note that the exponential family form for a given distribution is not unique. If there are linear or affine dependencies between the elements of the sufficient statistic then some components of η are redundant and the representation is called *over-complete*. Otherwise, it is called *minimal*. For example, a K-dimensional multinomial distribution can be represented by a K-dimensional parameter vector where one parameter is redundant. Many useful properties hold for the minimal representation only, but the over-complete one can be more elegant notationally.

1.2 Properties

- The dimensionality of the sufficient statistic $(T(x))$ is equal to the number of parameters (dimensionality of vector η). For exponential family distributions exists a sufficient statistic whose dimension is independent of the size of the sample.
- Products of exponential family distributions are exponential family distributions, but unnormalized.
- Moments:

  First moment: $E(T(y)) = \nabla_\eta A(\eta) = \theta$ (θ is also called *moment parameter*)

  Second moment: $Var(T(y)) = \nabla^2_\eta A(\eta)$
- The set of values of η for which the function $A(\eta) < +\infty$ is called the *natural parameter space*.
- The log-partition function $A(\eta)$ and its first derivative are convex, since the second derivative is a variance and therefore always positive (useful for MLE estimation).
- Every exponential family distribution has a conjugate prior (useful for Bayesian estimation) w: This is because the conjugate prior when multiplied by the likelihood yields a posterior that is in the same family, and the

likelihood is an exponential family distribution. The form of the conjugate prior for the distribution in Eq. 2 is:

$$p(\theta|\tau, \tau_0) = \frac{1}{Z(\tau, \tau_0)} \exp\{\tau^T \eta(\theta) - \tau_0 A(\eta(\theta))\} \tag{4}$$

2 Generalized linear models

2.1 Formulation

Generalized linear models (GLMs, not to be confused with *General Linear Models*) is a generalization of linear regression to response types other than Gaussian, as long as the distribution of that response is a member of the exponential family.

Let's assume that we are trying to predict response Y (labels, counts, real values) from a set of covariates X (features). In a linear model with parameters β, we assume that:

$$E(Y(X)) = X^T \beta \tag{5}$$

The generalization is obtained by assuming that $E(Y(X))$ is not identical to the linear combination $X^T\beta$, but it is connected to it through a function that is chosen according to the nature of Y. More formally, GLMs consist of 3 components:

- The *random* component (a.k.a. *response variable*), which is the exponential family distribution with canoncical parameter η that determines the form of the response, e.g. Poisson for counts. Note that we need to be able to write the exponential family distributions in its canonical form ($T(y) = y$ in Equation 1). For most of the exponential family distributions this is possible (a.k.a. *natural exponential family*) but there are cases like the LogNormal distribution which while it belongs to the exponential family it cannot be writen in the canonical form.

- The *systematic* component which specifies that the covariates X enter the model via linear combination $X^T\beta$ and since we are in the natural exponential family of distributions they define the natural parameter η.

- The monotone and differentiable function g which connects the systematic component with the mean parameter (θ):

$$g(\theta) = X^T \beta \tag{6}$$

$$E(Y) = \theta = g^{-1}(X^T \beta) \tag{7}$$

  g is called the *link* function and its inverse the *response* function. Since $X^T\beta = \eta$, the link function is the same as the mapping function between the natural and the mean parameter $\eta(\theta)$ and response function is the same as $\nabla_\eta A(\eta)$.

Using the above the form for logistic regression is obtained by we assuming the Bernoulli distribution for the response variable, which has the link function:

$$g(\theta) = \eta(\theta) = \log \frac{\theta}{1 - \theta} \tag{8}$$

and response function:

$$g^{-1}(\eta) = \frac{1}{1+\exp(\eta)} = \frac{1}{1+\exp(X^T\beta)} \tag{9}$$

Note that one can use exponential family distributions that cannot be written in canonical form (i.e., $T(y) \neq y$), thus resulting in *non-canonical link functions* (Peter John MacCullagh, J. A. Nelder book), however their use is quite rare.

2.2 Estimation

The standard way of estimating the parameters of GLMs with maximum likelihood estimation, in particular using the Newton-Raphson algorithm. Assuming a dataset $D = \{(x_1, y_1), (x_N, y_N)\}$ we want to find parameters β that maximize the log-likelihood ℓ:

$$\begin{align}
\ell(\beta|D) &= \log \prod_{i=1}^{N} h(y_i) \exp\{\eta(\beta)_i^T y_i - A(\eta_i)\} \tag{10}\\
&= \sum_{i=1}^{N} \log h(y_i) + \sum_{i=1}^{N} \{(\beta^T x_i) y_i - A(\eta_i)\} \tag{11}\\
&= \sum_{i=1}^{N} \log h(y_i) + \beta^T \sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} A(\eta_i) \tag{12}
\end{align}$$

Convexity of the log-partition function $A(\eta)$ guarantees global maximum. Newton-Raphson is a general method for maximizing (minimizing) a real-valued function $\ell(\beta)$ by iterating:

$$\beta_{(t+1)} \leftarrow \beta_{(t)} - [H(\ell(\beta))]^{-1} \nabla_\beta \ell(\beta) \tag{13}$$

If the function $\ell(\beta)$ is a quadratic function, then the maximum can be found in one step, as is the case with the normal distribution. The Hessian is needed and can be expensive in cases of non-canonical link functions. Then one can use Fisher's scoring method which used the expected Hessian instead. Also, it is worth mentioning that in some extensions of GLMs MCMC methods are used, e.g. (Hannah, Blei, Powell).

3 Other uses of exponential family distributions

Fuzzy clustering/Dimensionality reduction with hamiltonian monte carlo (Heller, Mohamed, Ghahramani)

Semi-supervised classification with variational inference (Liang)

Graphical models can be represented and solved as exponential families (Jordan and Wainwright)