Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.437 Inference and Information
Spring 2015

18 Conjugate Priors

Our recent developments have introduced the concepts of model capacity and the least informative prior. We have already discussed the computational challenges inherent in solving for the least informative prior. But even when we can derive it, the least informative prior may be computationally inconvenient to apply when updating beliefs as more data is acquired.

In these notes, we take advantage of the relative insensitivity of inference performance to the details of the prior to develop alternatives that are particularly amenable to both design and implementation. Specifically, we develop the concept of *conjugate priors* for this purpose.

The basic idea behind conjugate priors is conceptually simple. Given a model $p_{\mathsf{y}|\mathsf{x}}(\cdot|x)$, we look for a family of distributions such that if we use any element of the family as a prior distribution $p_{\mathsf{x}}(\cdot)$, the induced posterior distribution $p_{\mathsf{x}|\mathsf{y}}(\cdot|y)$ is also in this family. The importance of this choice is perhaps most apparent for successive belief update, i.e., updating the distribution for x given a sequence of observations, $y_1, y_2, \ldots$. Specifically, the conditional distribution for x given the first N of these observations is also in this family and serves as the new "prior" for the incorporation of the next observation. Using conjugate priors also obviates the need to compute the least informative prior for every data set size N.

Without specifying additional constraints on the conjugate family, the concept of conjugacy is not particularly useful. In particular, we could always consider the set of all probability distributions as a candidate family, but that would not result in more efficient computations. Moreover, we seek a family that is conjugate for sets of measurements of arbitrary size. As we will see in this section, if the dimensionality of the sufficient statistic does not change with the size of the data set, we can achieve this nice structure that leads to efficient belief updates.

We will find it convenient to introduce some new notation in this section. Given a model $p_{\mathsf{y}|\mathsf{x}}(\cdot|x)$, we use $p^N_{\mathbf{\mathsf{y}}|\mathsf{x}}(\cdot|x)$ to denote the joint probability distribution of a set of N i.i.d. random variables $\mathbf{\mathsf{y}} = [\mathsf{y}_1, \ldots, \mathsf{y}_N]^{\mathrm{T}}$ generated by the model with the same value of the parameter x:

$$p^N_{\mathbf{\mathsf{y}}|\mathsf{x}}(\mathbf{y}|x) \triangleq p_{\mathbf{\mathsf{y}}|\mathsf{x}}(y_1, \ldots, y_N|x) = \prod_{n=1}^{N} p_{\mathsf{y}_n|\mathsf{x}}(y_n|x). \tag{1}$$

Example 1. Let $\mathbf{\mathsf{y}} = [\mathsf{y}_1, \ldots, \mathsf{y}_N]^{\mathrm{T}}$ be a vector of i.i.d. Bernoulli random variables with parameter x, i.e.,

$$p_{\mathsf{y}_n|\mathsf{x}}(y|x) = x^y(1-x)^{1-y}, \quad y \in \{0,1\},\ x \in (0,1),\ n = 1, \ldots, N. \tag{2}$$

Then from (1) we obtain

$$p^N_{\mathbf{y}|\mathsf{x}}(\mathbf{y}|x) = x^{t_N(\mathbf{y})}(1-x)^{N-t_N(\mathbf{y})}, \tag{3}$$

with

$$t_N(\mathbf{y}) = \sum_{n=1}^{N} y_n,$$

which is clearly a sufficient statistic for the model $p_{\mathbf{y}|\mathsf{x}}(\cdot|x)$. Now let x be distributed according to a *beta distribution* with parameters α and β:

$$p_{\mathsf{x}}(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{Z(\alpha,\beta)}, \quad x \in (0,1), \tag{4}$$

where $Z(\alpha,\beta)$ is a normalization constant. It is easy to show that

$$p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y}) \propto x^{(\alpha+t_N(\mathbf{y}))-1}(1-x)^{(\beta+N-t_N(\mathbf{y}))-1}, \tag{5}$$

which for any N is again a member of the family of beta distributions with new parameters

$$\alpha' = \alpha + t_N(\mathbf{y}), \quad \beta' = \beta + N - t_N(\mathbf{y}). \tag{6}$$

Hence, the beta distribution is a conjugate prior for the Bernoulli distribution.

Example 2. Let $\mathsf{y}_1, \dots, \mathsf{y}_N$ be i.i.d. Gaussian random variables with mean x and known variance σ^2, i.e.,

$$p^N_{\mathbf{y}|\mathsf{x}}(\mathbf{y}|x) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - x)^2\right\}. \tag{7}$$

Then a conjugate prior for the mean parameter is also Gaussian. Specifically, if

$$\mathsf{x} \sim \mathtt{N}\left(\mu, \sigma_0^2\right) \tag{8}$$

for some parameters μ and σ_0^2, then as we recall from our earlier discussion of jointly Gaussian random variables (or, equivalently, as we obtain from some minor algebra), it follows that the posterior distribution of x given the observations $\mathbf{y}$ is also Gaussian regardless of N, i.e.,

$$\mathsf{x}|\mathbf{y} \sim \mathtt{N}\left(\frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\left(\frac{1}{N}\sum_{n=1}^{N} y_n\right), \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\sigma_0^2\right). \tag{9}$$

This procedure for updating the mean and the variance of the conditional distribution for x is precisely the computation for Bayes Least Squares estimation for jointly Gaussian variables, as we developed earlier.

The examples above illustrate the mechanics of using the conjugate priors. We now formalize the notion of conjugate priors and state a sufficient condition for their existence.

Definition 1. *A family of distributions $q(\cdot;\boldsymbol{\theta})$ with finite-dimensional parameter vector $\boldsymbol{\theta}$ is* conjugate *to a model $p_{y|\mathsf{x}}(\cdot|\mathbf{x})$ if*

(1) for any sample size N and any set of samples $\mathbf{y} = [y_1, \ldots, y_N]^{\mathrm{T}}$ generated i.i.d. by the model $p_{y|\mathsf{x}}(\cdot|\mathbf{x})$, there exists a value of the parameter $\boldsymbol{\theta}$ such that the joint probability of the sample, when viewed as a function of $\mathbf{x}$, is proportional to $q(\mathbf{x};\boldsymbol{\theta})$:

$$p_{\mathsf{y}|\mathsf{x}}(y_1, \cdots, y_N|\mathbf{x}) \propto q(\mathbf{x};\boldsymbol{\theta}); \tag{10}$$

and

(2) the family is closed under multiplication, i.e., for any pair of parameter vector values $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, there exists a third parameter vector $\boldsymbol{\theta}_3$ such that

$$q(\mathbf{x};\boldsymbol{\theta}_1)\, q(\mathbf{x},\boldsymbol{\theta}_2) \propto q(\mathbf{x};\boldsymbol{\theta}_3). \tag{11}$$

The parameter $\boldsymbol{\theta}$ is often referred to as a hyper-parameter or meta-parameter.

Given a candidate family, we can check if it satisfies the two conditions in the definition. Furthermore, the definition intuitively gives us a hint on how to construct conjugate priors. We need to manipulate the model as a function of the parameter $\mathbf{x}$ and see if the data $\mathbf{y}$ participates only as a finite dimensional summary, which would be the meta-parameter $\boldsymbol{\theta}$. The theorem below formalizes this intuition.

Theorem 1. *Let $p_{y|\mathsf{x}}(\cdot|\mathbf{x})$ be a model. If for each sample set size N, the corresponding joint distribution $p^N_{\mathsf{y}|\mathsf{x}}(\cdot|\mathbf{x})$ has a smooth sufficient statistic whose dimensionality does not depend on N, then a conjugate prior family exists for this model.*

Proof. Let $\mathbf{t}_N(\mathbf{y})$ be the sufficient statistic for the joint model $p^N_{\mathsf{y}|\mathsf{x}}(\cdot|\mathbf{x})$. From the Neyman Factorization Theorem, there exist functions $a_N(\cdot)$ and $b_N(\cdot)$ such that

$$p_{\mathsf{y}|\mathsf{x}}(\mathbf{y}|\mathbf{x}) = a_N(\mathbf{t}_N(\mathbf{y});\mathbf{x})\, b_N(\mathbf{y}) \qquad \text{for all } \mathbf{x}, \mathbf{y}. \tag{12}$$

We define $\boldsymbol{\theta} = [\mathbf{t}\ N]^{\mathrm{T}}$ where N is any positive integer, and for each value of N the range of possible values of $\mathbf{t}$ is that for which there exists a value of $\mathbf{y}$ such that $\mathbf{t} = \mathbf{t}_N(\mathbf{y})$. We further define

$$q(\mathbf{x};\boldsymbol{\theta}) \propto a_N(\mathbf{t};\mathbf{x}), \tag{13}$$

where the right-hand side of this equation needs to be normalized to create a proper probability distribution.

Eq. (12) implies that $q(\cdot;\boldsymbol{\theta})$ satisfies the first condition for conjugacy. Now we use the fact that the sufficient statistic is of finite dimensionality that does not depend on the number of samples N to prove that the family $q(\cdot;\boldsymbol{\theta})$ is closed under multiplication. Specifically, let $\boldsymbol{\theta}_1 = [\mathbf{t}_1\ N_1]^{\mathrm{T}}$ and $\boldsymbol{\theta}_2 = [\mathbf{t}_2\ N_2]^{\mathrm{T}}$ be two valid values of the parameter $\boldsymbol{\theta}$. By construction, this implies that there exist two sets of observations, $\mathbf{y}^1 = [y^1_1, \ldots, y^1_{N_1}]^{\mathrm{T}}$ and $\mathbf{y}^2 = [y^2_1, \ldots, y^2_{N_2}]^{\mathrm{T}}$, such that

$$\mathbf{t}_1 = \mathbf{t}_{N_1}(\mathbf{y}^1), \quad \mathbf{t}_2 = \mathbf{t}_{N_2}(\mathbf{y}^2). \tag{14}$$

We construct a new set of observations

$$\mathbf{y}^3 = [y^1_1, \ldots, y^1_{N_1} y^2_1, \ldots, y^2_{N_2}]^{\mathrm{T}} \tag{15}$$

and the corresponding parameter value $\boldsymbol{\theta}_3 = [\mathbf{t}_{N_1+N_2}(\mathbf{y}^3)\ N_1 + N_2]^{\mathrm{T}}$. Since all observations are generated i.i.d. by the model $p_{y|\mathbf{x}}(\cdot|\mathbf{x})$, we can show that

$$\begin{aligned} q(\mathbf{x};\boldsymbol{\theta}_1)\, q(\mathbf{x};\boldsymbol{\theta}_2) &\propto p_{\mathbf{y}_1}(\mathbf{y}^1|\mathbf{x})\, p_{\mathbf{y}_2}(\mathbf{y}^2|\mathbf{x}) = p_{\mathbf{y}_3}(\mathbf{y}^3|\mathbf{x}) & (16)\\ &\propto a_{N_1+N_2}(\mathbf{t}_{N_1+N_2}(\mathbf{y}^3);\mathbf{x}) \propto q(\mathbf{x};\boldsymbol{\theta}_3), & (17) \end{aligned}$$

where the proportionality is with respect to $\mathbf{x}$. This proves that $q(\cdot;\boldsymbol{\theta})$ is the conjugate family for the model $p_{y|\mathbf{x}}(\cdot|\mathbf{x})$. □

This theorem not only proves the existence of the conjugate priors for a broad set of distributions, but also provides a way to construct the conjugate prior family from the original model, as we illustrate next.

18.1 Conjugate Priors for Exponential Families

In this section, we consider the case of a sequence of i.i.d. vector observations $\mathbf{y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]^{\mathrm{T}}$, each of which is generated by the model

$$p_{\mathbf{y}_n|\mathbf{x}}(\mathbf{y}|\mathbf{x}) = \exp\left\{\boldsymbol{\lambda}(\mathbf{x})^{\mathrm{T}}\mathbf{t}(\mathbf{y}) - \alpha(\mathbf{x}) + \beta(\mathbf{y})\right\}. \tag{18}$$

Hence,

$$p^N_{\mathbf{y}|\mathbf{x}}(\mathbf{y}_1, \ldots, \mathbf{y}_N|\mathbf{x}) = \exp\left\{\boldsymbol{\lambda}(\mathbf{x})^T \sum_{n=1}^N \mathbf{t}(\mathbf{y}_n) - N\alpha(\mathbf{x}) + \sum_{n=1}^N \beta(\mathbf{y}_n)\right\}. \tag{19}$$

Following the steps in the proof of Theorem 1, we construct the Neyman factorization

$$p^N_{\mathbf{y}|\mathbf{x}}(\mathbf{y}_1, \ldots, \mathbf{y}_N|\mathbf{x}) = \exp\left\{\boldsymbol{\lambda}(\mathbf{x})^T \sum_{n=1}^N \mathbf{t}(\mathbf{y}_n) - N\alpha(\mathbf{x})\right\} \exp\left\{\sum_{n=1}^N \beta(\mathbf{y}_n)\right\} \tag{20}$$

and the conjugate prior

$$q(\mathbf{x};\mathbf{t},N) \propto \exp\left\{\mathbf{t}^T\boldsymbol{\lambda}(\mathbf{x}) - N\alpha(\mathbf{x})\right\}. \tag{21}$$

It is easy to verify that if we use an element of this family as a prior on x, i.e.,

$$p_{\mathsf{x}}(\mathbf{x}) = \frac{1}{Z} q(\mathbf{x}; \mathbf{t}_0, N_0), \tag{22}$$

then the posterior distribution for x given the observations $\mathbf{y}_1, \dots, \mathbf{y}_N$ is also in this family, with parameters

$$\mathbf{t}' = \mathbf{t}_0 + \sum_{n=1}^{N} \mathbf{t}(\mathbf{y}_n), \quad N' = N_0 + N. \tag{23}$$

Noting that (21) is itself a (linear) exponential family with natural statistic

$$\begin{bmatrix} \boldsymbol{\lambda}(\mathbf{x}) & -\alpha(\mathbf{x}) \end{bmatrix}^{\mathrm{T}},$$

we conclude that a conjugate prior for an exponential family is also an exponential family.

18.2 Summary

As we saw in several examples of this section, performing Bayesian belief update is simple if we use conjugate priors. We also saw that conjugate priors exist for a broad set of distributions, namely, the exponential families. Moreover, the conjugate families are often quite rich, allowing one to approximate other densities to suitable accuracy. For example, in a nonBayesian setting, they can be used to approximate the least informative priors in pursuing inference performance close to model capacity. Or in a Bayesian setting involving inference or estimation where the relevant prior is known but difficult to work with, a suitably chosen conjugate prior can be used as an approximation.