L E C T U R E 18

Appendix. Dirichlet Distribution (continuation).

Let us prove the following statement:

Theorem 20.1. Let η_1 and η_2 be independent Gamma distributed random variables with parameters $(\alpha_1, 1)$, $(\alpha_2, 1)$ respectively, and

$$Y_1 = \frac{\eta_1}{\eta_1 + \eta_2}.$$

Then Y_1 has *Beta* distribution with parameters $B(\alpha_1, \alpha_2)$, that is density function of Y_1 has the following form:

$$f_{Y_1}(x) = \begin{cases} 0, & \text{if} \quad x \notin (0,1) \\ \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\cdot\Gamma(\alpha_2)}\, x^{\alpha_2-1} \cdot (1-x)^{\alpha_1-1}\cdot, & \text{if} \quad x \in (0,1) \end{cases}$$

Proof. By definition, the distribution function of Y_1 is

$$F_{Y_1}(x) = P\left(\frac{\eta_1}{\eta_1+\eta_2} \le x\right) =$$

and therefore, we get, because η_1 and η_2 are nonnegative:

$$= \begin{cases} 0, & \text{if} \quad x \le 0 \\ P(\eta_1 \le x\cdot\eta_1 + x\cdot\eta_2), & \text{if} \quad x > 0 \end{cases}$$

and

$$= \begin{cases} 0, & \text{if} \quad x \le 0 \\ P(\frac{\eta_1(1-x)}{x} \le \eta_2), & \text{if} \quad x > 0 \end{cases} = \begin{cases} 0, & \text{if} \quad x \le 0 \\ P(\frac{\eta_1(1-x)}{x} \le \eta_2), & \text{if} \quad 0 \le x \le 1 \\ 1, & \text{if} \quad x \ge 1 \end{cases}$$

since if $x \ge 1$ we have $1 - x \le 0$ and the above event is the certain event and thus probability equal to 1.

$$= \begin{cases} 0, & \text{if} \quad x \le 0 \\ 1 - P(\frac{1-x}{x}\eta_1 \ge \eta_2), & \text{if} \quad 0 \le x \le 1 \\ 1, & \text{if} \quad x \ge 1 \end{cases} = \begin{cases} 0, & \text{if} \quad x \le 0 \\ 1 - P(\frac{\eta_2}{\eta_1} \le \frac{1-x}{x}), & \text{if} \quad 0 \le x \le 1 \\ 1, & \text{if} \quad x \ge 1 \end{cases}$$

Further we get

$$F_{Y_1}(x) = \begin{cases} 0, & \text{if} \quad x \le 0 \\ 1 - F_{\frac{\eta_2}{\eta_1}}\left(\frac{1-x}{x}\right), & \text{if} \quad 0 \le x \le 1 \\ 1, & \text{if} \quad x \ge 1 \end{cases} \tag{20.1}$$

By differentiating both sides of (20.1) we get

$$f_{\frac{\eta_1}{\eta_1+\eta_2}}(x) = \begin{cases} 0, & \text{if} \quad x \notin (0,1) \\ -f_{\frac{\eta_2}{\eta_1}}\left(\frac{1-x}{x}\right)\cdot\left(\frac{1-x}{x}\right)', & \text{if} \quad x \in (0,1) \end{cases} \tag{20.2}$$

So, we need to use the following result from probability theory:

Lemma 20.3. If η_1 and η_2 are 2 independent random variables with $f_1(x)$ and $f_2(x)$ density functions. The density function of the variable $\dfrac{\eta_1}{\eta_2}$ has the following form:

$$f_{\eta_1/\eta_2}(x) = \int_0^{+\infty} y f_2(y) f_1(xy)\, dy - \int_{-\infty}^{0} y f_2(y) f_1(xy)\, dy.$$

In our case we have

$$f_i(y) = \frac{1}{\Gamma(\alpha_i)} y^{\alpha_i - 1} \cdot e^{-y}, \quad y > 0 \qquad i = 1, 2.$$

Using Lemma 20.3 and note that our random variables are positive, we obtain:

$$f_{\frac{\eta_2}{\eta_1}}(x) = \int_0^{\infty} y \cdot f_2(y) f_1(xy)\, dy = \frac{1}{\Gamma(\alpha_1) \cdot \Gamma(\alpha_2)} \int_0^{\infty} y^{\alpha_2 - 1} \cdot e^{-y} \cdot (x \cdot y)^{\alpha_1 - 1} \cdot e^{-xy}\, dy =$$

$$= \frac{x^{\alpha_1 - 1}}{\Gamma(\alpha_1) \cdot \Gamma(\alpha_2)} \int_0^{\infty} y^{\alpha_1 + \alpha_2 - 1} \cdot e^{-(1+x)y}\, dy =$$

Make the change of variable $(1 + x) \cdot y = z$, we get:

$$= \frac{x^{\alpha_1 - 1}}{\Gamma(\alpha_1) \cdot \Gamma(\alpha_2)(1 + x)^{\alpha_1 + \alpha_2 - 1}} \int_0^{\infty} z^{\alpha_1 + \alpha_2 - 1} \cdot e^{-z} \frac{dz}{1 + x} =$$

$$= \frac{x^{\alpha_1 - 1}}{\Gamma(\alpha_1) \cdot \Gamma(\alpha_2)(1 + x)^{\alpha_1 + \alpha_2}} \int_0^{\infty} z^{\alpha_1 + \alpha_2 - 1} \cdot e^{-z}\, dz = \frac{x^{\alpha_1 - 1}\, \Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1) \cdot \Gamma(\alpha_2)(1 + x)^{\alpha_1 + \alpha_2}}$$

Substituting our result into (20.2) we obtain:

$$f_{\frac{\eta_1}{\eta_1 + \eta_2}}(x) = \begin{cases} 0, & \text{if} \quad x \notin (0, 1) \\ \frac{1}{x^2} \cdot f_{\frac{\eta_2}{\eta_1}}\left(\frac{1-x}{x}\right), & \text{if} \quad x \in (0, 1) \end{cases} =$$

$$= \begin{cases} 0, & \text{if} \quad x \notin (0, 1) \\ \frac{1}{x^2} \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1) \cdot \Gamma(\alpha_2)} \cdot \frac{(\frac{1-x}{x})^{\alpha_1 - 1}}{(1 + \frac{1-x}{x})^{\alpha_1 + \alpha_2}}, & \text{if} \quad x \in (0, 1) \end{cases} =$$

$$= \begin{cases} 0, & \text{if} \quad x \notin (0, 1) \\ \frac{1}{x^2} \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1) \cdot \Gamma(\alpha_2)} \cdot \frac{(1-x)^{\alpha_1 - 1} \cdot x^{\alpha_1 + \alpha_2}}{x^{\alpha_1 - 1}}, & \text{if} \quad x \in (0, 1) \end{cases} =$$

$$= \begin{cases} 0, & \text{if} \quad x \notin (0, 1) \\ \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1) \cdot \Gamma(\alpha_2)} x^{\alpha_2 - 1} \cdot (1 - x)^{\alpha_1 - 1} \cdot, & \text{if} \quad x \in (0, 1) \end{cases}$$

Therefore, Y_1 has Beta distribution with parameters α_1 and α_2. The proof is complete.

In general case we have the following result:

Theorem 20.2. Let η_1, η_2, ..., η_{k+1} be independent random variables Gamma distributed with parameters $(\alpha_1, 1)$, $(\alpha_2, 1)$,..., $(\alpha_{k+1}, 1)$ respectively, and

$$Y_i = \frac{\eta_i}{\eta_1 + \eta_2 + ... + \eta_{k+1}}, \qquad i = 1, 2, ..., k.$$

Then $(Y_1, Y_2, ..., Y_k)$ has k-dimensional Dirichlet $D(\alpha_1, \alpha_2, ..., \alpha_k; \alpha_{k+1})$ distribution.

Proof: Using Corollary 19.1, we obtain that

$$\sum_{i=1}^{k+1} \eta_i$$

has Gamma distribution with parameters $(\sum_{i=1}^{k+1} \alpha_i, 1)$, therefore,

$$P\left(\frac{\eta_1}{\eta_1 + \eta_2 + ... + \eta_{k+1}} \le x\right) = \begin{cases} 0 & \text{if} \quad x \le 0 \\ P(\eta_1 \le x\eta_1 + x\eta_2 + ... + x\eta_{k+1}) & \text{if} \quad x > 0 \end{cases} =$$

$$= \begin{cases} 0 & \text{if} \quad x \le 0 \\ P(\eta_1(1-x) \le x\eta_2 + ... + x\eta_{k+1}) & \text{if} \quad x > 0 \end{cases} =$$

$$= \begin{cases} 0 & \text{if} \quad x \le 0 \\ P\left(\eta_1 \dfrac{1-x}{x} \le \eta_2 + ... + \eta_{k+1}\right) & \text{if} \quad x > 0 \end{cases} =$$

since if $x \ge 1$ we have $1 - x \le 0$ and the above event is the certain event and thus probability equal to 1.

$$= \begin{cases} 0, & \text{if} \quad x \le 0 \\ 1 - P(\frac{1-x}{x}\eta_1 \ge \eta_2 + ... + \eta_{k+1}), & \text{if} \quad 0 \le x \le 1 \\ 1, & \text{if} \quad x \ge 1 \end{cases} =$$

$$= \begin{cases} 0, & \text{if} \quad x \le 0 \\ 1 - P(\frac{\eta_2 + ... + \eta_{k+1}}{\eta_1} \le \frac{1-x}{x}), & \text{if} \quad 0 \le x \le 1 \\ 1, & \text{if} \quad x \ge 1 \end{cases}$$

Thus, we obtain

$$F_{Y_1}(x) = \begin{cases} 0, & \text{if} \quad x \le 0 \\ 1 - F_{\frac{\eta_2 + ... + \eta_{k+1}}{\eta_1}}\left(\frac{1-x}{x}\right), & \text{if} \quad 0 \le x \le 1 \\ 1, & \text{if} \quad x \ge 1 \end{cases} \tag{20.3}$$

By differentiating both sides of (20.3) we get

$$f_{Y_1}(x) = \begin{cases} 0, & \text{if} \quad x \notin (0, 1) \\ -f_{\frac{\eta_2 + ... + \eta_{k+1}}{\eta_1}}\left(\frac{1-x}{x}\right) \cdot \left(\frac{1-x}{x}\right)', & \text{if} \quad x \in (0, 1) \end{cases} \tag{20.4}$$

Using Lemma 20.3 and that the distribution of $\eta_2 + ... + \eta_{k+1}$ and η_1 are Gamma with parameters $\alpha_2 + ... + \alpha_{k+1}$ and α_1 respectively, arguing as in the theorem 20.1 we obtain the result. Theorem 20.2 is proved.

§21. THE CHOICE OF A PRIOR.

Obviously, a critical feature of any Bayesian analysis is the choice of a prior. The key here is that when the data have sufficient signal, even a bad prior will still not greatly influence the posterior. In a sense, this is an asymptotic property of Bayesian analysis in that all but pathological priors will be overcome by sufficient amounts of data. As mentioned above, one can check the impact of the prior by seeing how stable to posterior distribution is to different choices of priors. If the posterior is highly dependent on the prior, then the data (the likelihood function) may not contain sufficient information. However, if the posterior is relatively stable over a choice of priors, then the data indeed contain significant information. The location of a parameter (mean or mode) and its precision (the reciprocal of the variance) of the prior is usually more critical than its actual shape in terms of conveying prior information. The shape (family) of the prior distribution is often chosen to facilitate calculation of the prior, especially through the use of conjugate priors that, for a given likelihood function, return a posterior in the same distribution family as the prior (i.e., a gamma prior returning a gamma posterior when the likelihood is Poisson). We will return to conjugate priors shortly, but we first discuss other standard approaches for construction of priors.

§21.1. Diffuse Priors.

One of the most common priors is the flat, or diffuse (often called uninformative) prior which is simply a constant,

$$p(\theta) = \begin{cases} k = \dfrac{1}{b-a} & \text{for} \qquad a \le \theta \le b \\ 0 & \text{otherwise} \end{cases} \tag{21.1}$$

This conveys that we have no a priori reason to favor any particular parameter value over another. With a flat prior, the posterior just a constant times the likelihood,

$$p(\theta|x_1, x_2, ..., x_n) = C\, L(x_1, x_2, ..., x_n|\theta) \tag{21.2}$$

and we typically write that

$$p(\theta|x_1, x_2, ..., x_n) \propto L(x_1, x_2, ..., x_n|\theta).$$

In many cases, classical expressions from frequentist statistics are obtained by Bayesian analysis assuming a flat prior.

If the variable (i.e. parameter) of interest ranges over $(0, +\infty)$ or $(-\infty, +\infty)$, then strictly speaking a flat prior does not exist, as if the constant takes on any nonzero value, the integral does not exist. In such cases a flat prior (i.e., assuming $p(\theta|x_1, x_2, ..., x_n) \propto L(x_1, x_2, ..., x_n|\theta)$) is referred to as an improper prior.

§21.2 Sufficient Statistics and Data-Transformed Likelihoods.

In the lessons on point estimation, we derived estimators of various parameters using two methods, namely, the method of maximum likelihood and the method of moments. The estimators resulting from these two methods are typically intuitive estimators. It makes sense, for example, that we would want to use the sample mean

$$\overline{X}$$

and sample variance S^2 to estimate the mean μ and variance σ^2 of a normal population.

In the process of estimating such a parameter, we summarize, or reduce, the information in a sample $(X_1, X_2, ..., X_n)$, of size n to a single number, such as the sample mean

$$\overline{X}.$$

The actual sample values are no longer important to us. That is, if we use a sample mean of 3 to estimate the population mean μ, it doesn't matter if the original data values were $(1, 3, 5)$ or $(2, 3, 4)$. Has this process of reducing the n data points to a single number retained all of the information about μ that was contained in the original n data points? Or has some information about the parameter been lost through the process of summarizing the data?

How to find statistics that summarize all of the information in a sample about the desired parameter? Such statistics are called sufficient statistics.

Suppose that $X_1, ..., X_n \sim p(x, \theta)$. T is sufficient for θ if the conditional distribution

$$p(X_1, ..., X_n|T = t, \theta) = p(X_1, ..., X_n|T = t).$$

Intuitively, this means that you can replace X_1, ...,X_n with $T = t(X_1, ..., X_n)$ without losing information, that is if $T = t(X_1, ..., X_n)$ is sufficient, then T contains all information you need from the data to compute the likelihood function.

Theorem 21.1. $T = t(X_1, ..., X_n)$ is sufficient for θ if the joint pdf (or pmf) of $X_1, ..., X_n)$ can be factored as

$$p(X_1, ..., X_n; \theta) = h(X_1, ..., X_n)\, g(t(X_1, ..., X_n); \theta).$$