A symmetric generalized Dirichlet distribution

Bret Larget

5/14/2017

The Dirichlet Distribution

The Dirichlet distribution is a very commonly used probability distribution on sets of positive random variables constrained to sum to one. The random variables $X_1, \ldots, X_k$ are said to have a Dirichlet distribution when they have the joint density

$$f(x_1, \ldots, x_k) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k x_i^{\alpha_i - 1}, \qquad \text{where } x_i > 0 \text{ for all } i \text{ and } \sum_{i=1}^k x_i = 1$$

where the parameters $\alpha_i > 0$ for $i = 1, \ldots, k$.

Each random variable X_i has a marginal $\text{Beta}(\alpha_i, \theta - \alpha_i)$ distribution where $\theta = \sum_{i=1}^k \alpha_i$. It follows that X_i has mean $\mathsf{E}(X_i) = \alpha_i/\theta$ and variance $\mathsf{Var}(X_i) = \alpha_i(\theta - \alpha_i)/(\theta^2(\theta + 1))$. A consequence is that when attempting to select a Dirichlet distribution to match the distribution of a given set of random variables constrained to equal one, while it is possible to select the parameters $\{\alpha_i\}$ to match the marginal means by letting α_i be proportional to the desired marginal mean, there remains only a single scale factor which determines all of the marginal variances. We seek a generalization which a larger parameterization that is flexible enough to match, at least approximately, the means and variances of each marginal distribution.

We know of another generalization of the Dirichlet distribution (described HERE) that is different than what we propose here in that it is asymmetric in the indices of the random variables and has the property that some correlations may be positive.

Generation of random variables

To generate random variables $X_1, \ldots, X_k \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$, one simply generates independent random variables $Y_i \sim \text{Gamma}(\alpha_i, \lambda)$ for $i = 1, \ldots, k$ and any arbitrary $\lambda > 0$ (typically $\lambda = 1$) and letting $X_i = Y_i / \sum_{j=1}^k Y_j$. This suggests that by allowing the value of λ to vary with i that we may be able to create a distribution on positive random variables constrained to sum to one with the desired flexibility in the first and second moments.

A Generalized Dirichlet Distribution

Define the symmetric generalized Dirichlet distribution on $X_1, \ldots, X_n$ to be the distribution of $(X_1, \ldots, X_k)$ where $X_i = Y_i / \sum_{j=1}^k Y_j$ for $i = 1, \ldots, k$ where the random variables $\{Y_i\}$ are mutually independent and $Y_i \sim \text{Gamma}(\alpha_i, \lambda_i)$. As the distribution of the $\{X_i\}$ would be the same if all $\{Y_i\}$ were multiplied by a common constant, we add the constraint that $\sum_{i=1}^k \lambda_i = k$ so that the average values of the $\{\lambda_i\}$ parameters is one. (CHECK IF SETTING MEAN OF $1/\lambda_i$ TO BE ONE IS ANY MORE CONVENIENT).

It is known (REFERENCES) that the distribution of the sum $S = \sum_{i=1}^k Y_i$ may be written as an infinite mixture of Gamma densities. However, the joint density of $X_1, \ldots, X_k$) has a closed form solution.

$$f(x_1, \ldots, x_k) = \frac{\Gamma\big(\sum_{i=1}^k \alpha_i\big)\big(\prod_{i=1}^k \lambda_i^{\alpha_i}\big)}{\prod_{i=1}^k \Gamma(\alpha_i)} \times \frac{\prod_{i=1}^k x_i^{\alpha_i - 1}}{\big(\sum_{i=1}^k \lambda_i x_i\big)^{\sum_{i=1}^k \alpha_i}}, \qquad \text{where } x_i > 0 \text{ for all } i \text{ and } \sum_{i=1}^k x_i = 1$$

The derivation is shown in the appendix.

I have not been able to derive closed form solutions for the marginal means and variances, but the means are close (if not exactly equal to) $(\alpha_i/\lambda_i)/\sum_{j=1}^{k}(\alpha_j/\lambda_j)$.

Parameter Estimation

Suppose that a probability density g on the k-dimensional simplex has marginal mean $\{\mu_i\}$ and marginal variances $\{v_i\}$. We do the following.

$$\alpha_i = \frac{\mu_i^2(1-\mu_i)}{v_i}$$

$$\lambda_i = \frac{\mu_i(1-\mu_i)}{v_i} \Bigg/ \sum_{j=1}^{k} \frac{\mu_j(1-\mu_j)}{kv_j}$$

By construction, the mean of the $\{\lambda_i\}$ is one.

I need to provide some more theoretical evidence that these parameter estimates work.

Example

The data set 024 does not work well with Bistro. A major issue is that the Q matrix parameters $\pi = \{\pi_i\}$ and $s = \{s_i\}$ have a Bayesian posterior density determined by MCMC simulation that is not well fit by a Dirichlet distribution, but we attempt to propose values for π and s from Dirichlet distributions nonetheless.

For this data set, here are the empirical means, standard deviations, and variances, of the marginal distributions of π and s.

```
## # A tibble: 10 × 6
##    parameter     n       mean          sd          var     scale
##        <chr> <int>      <dbl>       <dbl>        <dbl>     <dbl>
## 1        pi1 10000 0.24273965 0.002651711 7.031571e-06 26141.684
## 2        pi2 10000 0.25767324 0.002708244 7.334586e-06 26078.872
## 3        pi3 10000 0.18633048 0.002861878 8.190347e-06 18510.991
## 4        pi4 10000 0.31325663 0.003199113 1.023432e-05 21020.138
## 5         s1 10000 0.36681672 0.004417770 1.951669e-05 11900.696
## 6         s2 10000 0.15875919 0.003009095 9.054652e-06 14749.844
## 7         s3 10000 0.10927809 0.003373955 1.138357e-05  8550.602
## 8         s4 10000 0.03653933 0.001948581 3.796967e-06  9271.664
## 9         s5 10000 0.30430322 0.003855970 1.486851e-05 14238.333
## 10        s6 10000 0.02430345 0.001225131 1.500945e-06 15798.577
```

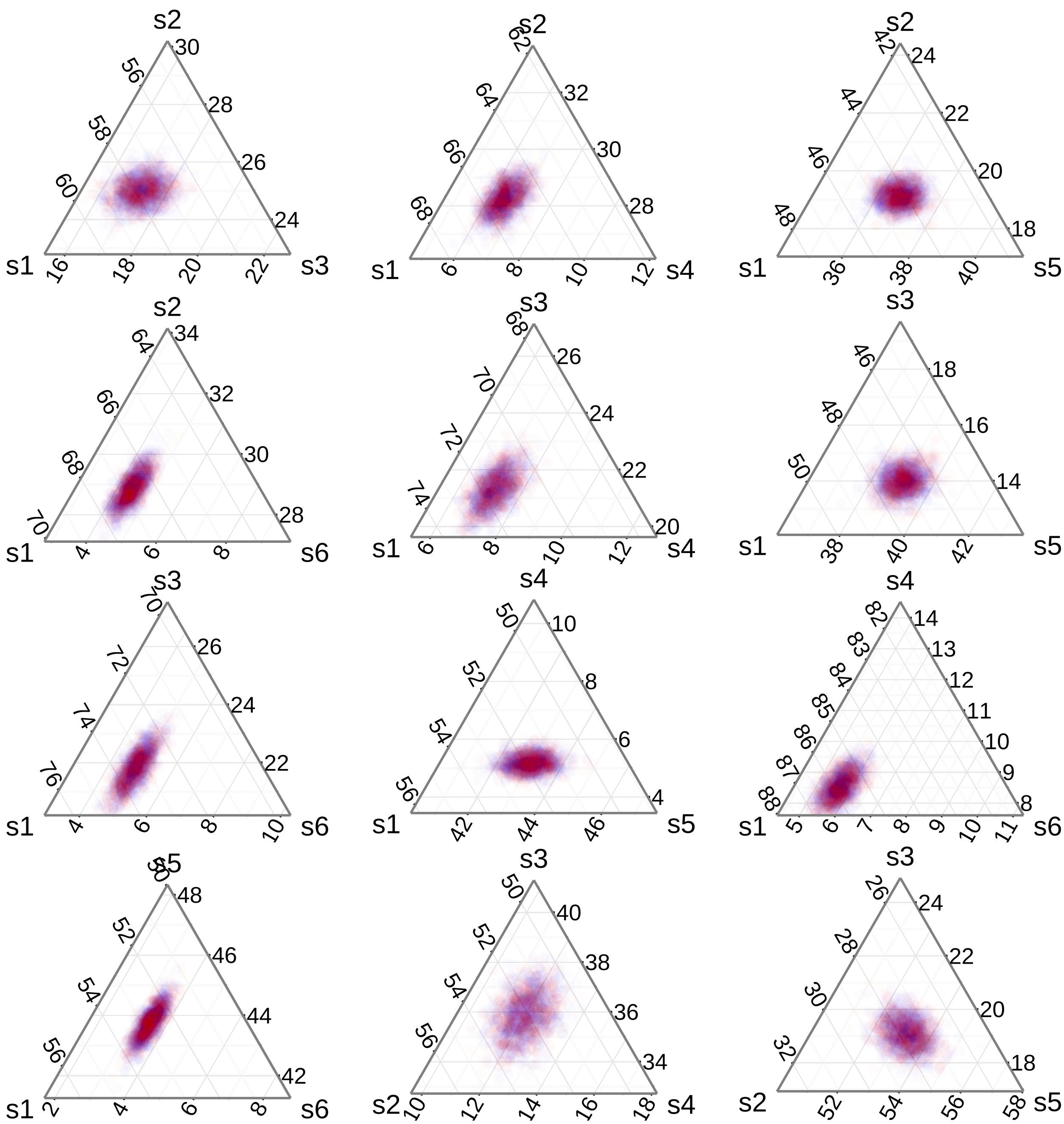
Using the formulas above, here are the estimated values of α_i and λ_i for the s parameters.

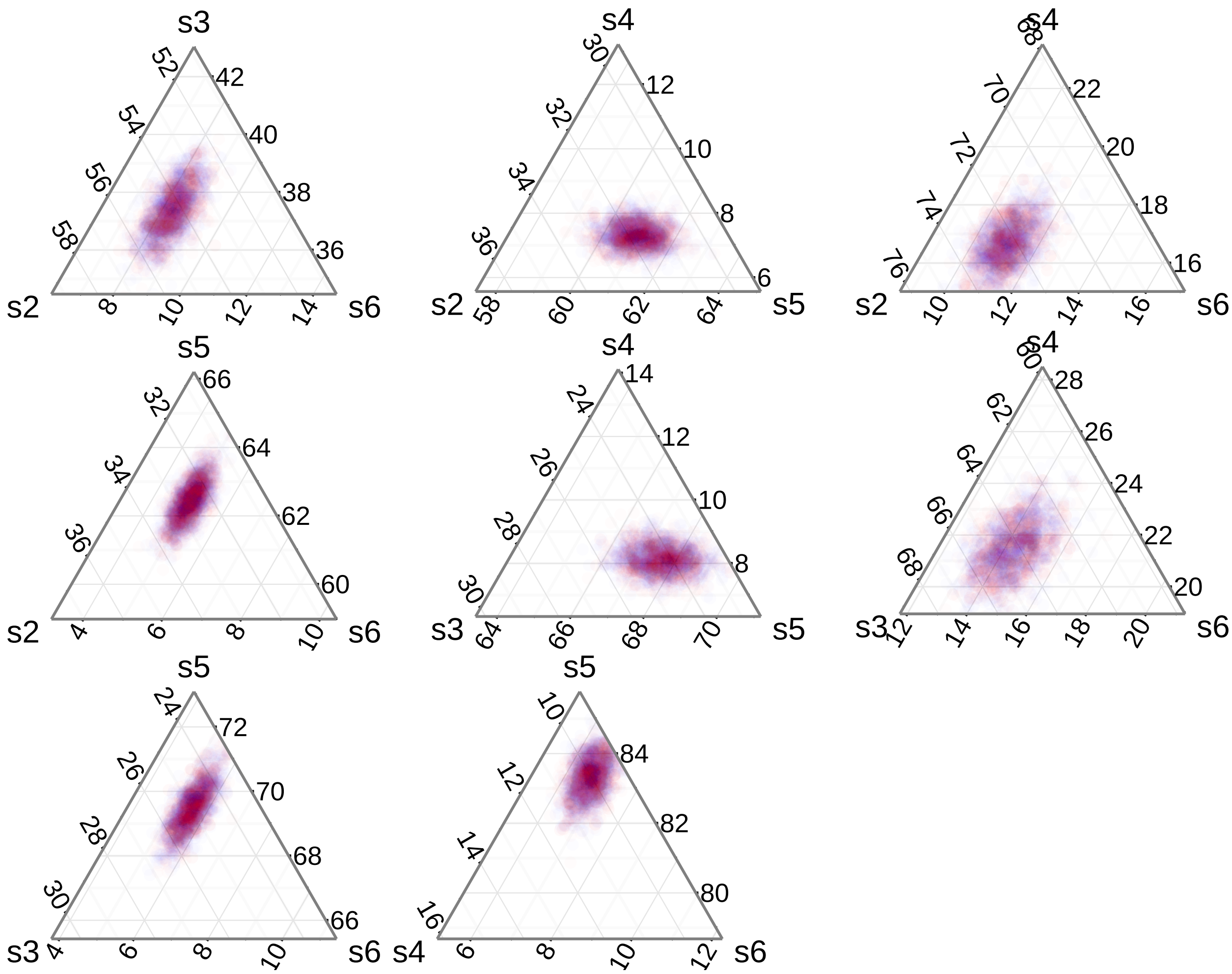
```
## # A tibble: 6 × 3
##   parameter     alpha    lambda
##       <chr>     <dbl>     <dbl>
## 1        s1 4365.3742 0.9583203
## 2        s2 2341.6733 1.1877520
## 3        s3  934.3934 0.6885493
## 4        s4  338.7804 0.7466139
## 5        s5 4332.7706 1.1465618
## 6        s6  383.9600 1.2722027
```

Using these values, I generated 1000 sets of s and compare the means and variances with the empirical values from s.

```
## # A tibble: 6 × 5
##   parameter     n       mean          sd          var
##       <chr> <int>      <dbl>       <dbl>        <dbl>
## 1        s1  1000 0.36686937 0.004419572 1.953262e-05
## 2        s2  1000 0.15873109 0.003030395 9.183297e-06
## 3        s3  1000 0.10931068 0.003333917 1.111500e-05
## 4        s4  1000 0.03658887 0.001907006 3.636671e-06
## 5        s5  1000 0.30422188 0.003956965 1.565757e-05
## 6        s6  1000 0.02427811 0.001182733 1.398858e-06
```

Plots of the simulated s values agree very well with the original sampled values.





Appendix

Here is a derivation of the density of the symmetric generalized Dirichlet distribution.

The joint density of $(Y_1, \ldots, Y_k)$ where $Y_i \sim \text{Gamma}(\alpha_i, \lambda_i)$ and are mutually independent is

$$f(y_1, \ldots, y_k) = \prod_{i=1}^{k} \left(\frac{\lambda_i^{\alpha_i}}{\Gamma(\alpha_i)} y_i^{\alpha_i - 1} \mathrm{e}^{-\lambda_i y_i} \right)$$

Let $S = \sum_{i=1}^{k} Y_i$ and $X_i = Y_i / S$ for $i = 1, \ldots, k$. Note that $Y_i = S X_i$ for $i = 1, \ldots, k-1$ and $Y_k = S(1 - \sum_{i=1}^{k-1} X_i)$. We find the joint density of $(S, X_1, \ldots, X_{k-1})$. The Jacobian matrix $J = \partial(y_1, \ldots, y_k) / \partial(x_1, \ldots, x_{k-1}, s)$ satisfies

$$J_{ij} = \begin{cases} s & \text{if } i = j,\ i < k \\ 0 & \text{if } i \neq j,\ i < k \\ x_j & \text{if } i = k,\ j < k \\ -s & \text{if } j = k,\ i < k \\ 1 - \sum_{i=1}^{k-1} x_i & \text{if } i = j = k \end{cases}$$

To determine the determinant, replace the kth row by itself minus x_i/s times the ith row for $i = 1, \ldots, k-1$, which does not affect the value of the determinant. The resulting kth row has values 0 in columns $j = 1, \ldots, k-1$

and value 1 in column k and is a diagonal matrix with diagonal elements s in the first $k-1$ rows and 1 in the last row. Thus $|\det J| = s^{k-1}$. It follows that the joint density of $(X_1, \ldots, X_{k-1}, S)$ is

$$f(x_1, \ldots, x_{k-1}, s) = s^{k-1} \prod_{i=1}^{k} \left(\frac{\lambda_i^{\alpha_i}}{\Gamma(\alpha_i)} (s x_i)^{\alpha_i - 1} \mathrm{e}^{-\lambda_i s x_i} \right)$$

where $x_k = 1 - \sum_{i=1}^{k-1} x_i$. Rewriting, the joint density is as follows.

$$f(x_1, \ldots, x_{k-1}, s) = \prod_{i=1}^{k} \left(\frac{\lambda_i^{\alpha_i} x_i^{\alpha_i - 1}}{\Gamma(\alpha_i)} \right) s^{\sum_{i=1}^{k} \alpha_i - 1} \mathrm{e}^{-\left(\sum_{i=1}^{k} \lambda_i x_i \right) s}$$

Holding all of the x_i constant, we recognize the gamma density in s up to constants and can thus integrate out s to find the joint density of the x_i.

$$f(x_1, \ldots, x_{k-1}) = \prod_{i=1}^{k} \left(\frac{\lambda_i^{\alpha_i} x_i^{\alpha_i - 1}}{\Gamma(\alpha_i)} \right) \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i \right)}{\left(\sum_{i=1}^{k} \lambda_i x_i \right)^{\sum_{i=1}^{k} \alpha_i}}$$

where $x_k = 1 - \sum_{i=1}^{k-1} x_i$. Reorganization yields the equation at the bottom of page 1.