# A Generalization of the Dirichlet Distribution

**Robin K. S. Hankin**
Auckland University of Technology

**Abstract**

This vignette is based on Hankin (2010). It discusses a generalization of the Dirichlet distribution, the 'hyperdirichlet', in which various types of incomplete observations may be incorporated. It is conjugate to the multinomial distribution when some observations are censored or grouped. The **hyperdirichlet** R package is introduced and examples given. A number of statistical tests are performed on the example datasets, which are drawn from diverse disciplines including sports statistics, the sociology of climate change, and psephology.

For reasons of performance, some of the more computationally expensive results are pre-loaded. To calculate them from scratch, change "`calc_from_scratch <- TRUE`" to "`calc_from_scratch <- FALSE`" in chunk `time_saver`.

*Keywords*: Dirichlet distribution, combinatorics, R, multinomial distribution, constrained optimization.

## 1. Introduction

The Dirichlet distribution is conjugate to the multinomial distribution in the following sense. If random variables $\mathbf{p} = (p_1, \ldots, p_k)$ satisfy $\sum_{i=1}^{k} p_i = 1$ and are Dirichlet, that is, they have a prior distribution

$$f(\mathbf{p}) = p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} \cdots p_k^{\alpha_k - 1} \cdot \frac{\Gamma\left(\sum_{i=1}^{n} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma\left(\alpha_i\right)} \tag{1}$$

then if the $p_i$ are interpreted as the parameters of a multinomial distribution from which $a_1, \ldots, a_k$ independent observations of types $1, \ldots, k$ are made, then the posterior PDF for $\mathbf{p}$ will be

$$p_1^{a_1 + \alpha_1 - 1} p_2^{a_2 + \alpha_2 - 1} \cdots p_k^{a_k + \alpha_k - 1} \cdot \frac{\Gamma\left(\sum_{i=1}^{n} a_i + \alpha_i\right)}{\prod_{i=1}^{n} \Gamma\left(a_i + \alpha_i\right)}, \tag{2}$$

thus belonging to the same family as the prior, the Dirichlet.

It is convenient to denote the distribution of Equation 1 as $D(\boldsymbol{\alpha})$ where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ is the $k$-tuple of Dirichlet parameters; thus that of Equation 2 would be $D(\boldsymbol{\alpha} + \mathbf{a})$, where $\mathbf{a} = (a_1, \ldots, a_k)$ is the $k$-tuple of observation counts.

In this paradigm, an observation is informative because it increases the Dirichlet parameter of its category by one. However, an observation may be informative even if it does not belong unambiguously to a single category: Consider making $r = r_{123} + r_{456}$ censored observations whose exact classes are not observed but $r_{123}$ are known to be one of categories 1, 2, or 3, and $r_{456}$ are known to be one of categories 4,5, or 6. The posterior would satisfy

$$f\left(\mathbf{p}\right) \propto p_1^{\alpha_1-1} p_2^{\alpha_2-1} \cdots p_k^{\alpha_k-1} \cdot \left(p_1+p_2+p_3\right)^{r_{123}} \cdot \left(p_4+p_5+p_6\right)^{r_{456}} \tag{3}$$

and is not Dirichlet. Consider now the case where observations are made from a conditional multinomial. Suppose $s_{123}$ observations are made whose class is known *a priori* to be one of 1,2, and 3, and there are $s_i$ of class $i$ where $i = 1, 2, 3$, then the posterior would be

$$f(\mathbf{p}) \propto p_1^{\alpha_1-1} p_2^{\alpha_2-1} \cdots p_k^{\alpha_k-1} \cdot \frac{p_1^{s_1} p_2^{s_2} p_3^{s_3}}{\left(p_1+p_2+p_3\right)^{s_{123}}}, \tag{4}$$

also not Dirichlet. These types of observation occur frequently in a wide range of contexts and naturally lead one to consider the following generalization of the Dirichlet distribution:

$$f(\mathbf{p}) \propto \left(\prod_{i=1}^{k} p_i\right)^{-1} \prod_{G \in \wp(K)} \left(\sum_{i \in G} p_i\right)^{\mathcal{F}(G)} \tag{5}$$

where $K$ is the set of positive integers not exceeding $k$, $\wp(K)$ is its power set, and $\mathcal{F}$ is a function that maps $\wp(K)$ to the real numbers[1]. Here, $p_i \geqslant 0$ for $1 \leqslant i \leqslant k$ and $\sum_{i=1}^{k} p_i = 1$. We call this the *hyperdirichlet distribution* and denote it by $H(\mathcal{F})$.

The first term is there so that defining

$$\mathcal{F}(G) = \begin{cases} \alpha_i & \text{if } G == \{i\} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

results in $H(\mathcal{F})$ being identical to $D(\boldsymbol{\alpha})$.

The distribution appears to have $|\wp(K)| = 2^k$ real parameters but the effective number of degrees of freedom is actually $2^k - 2$: the first and last parameter correspond to the empty set and the complete set respectively, and so do not affect the PDF.

*Normalizing constant*

The normalizing factor of the PDF given in equation 5 is given by

$$B(\mathcal{F}) = \int_{\mathbf{p} \geqslant \mathbf{0}, \sum_{i=1}^{k-1} p_i \leqslant 1} \left(\prod_{i=1}^{k} p_i\right)^{-1} \prod_{G \in \wp(K)} \left(\sum_{i \in G} p_i\right)^{\mathcal{F}(G)} d\left(p_1, \ldots, p_{k-1}\right) \tag{7}$$

where $p_k = 1 - \sum_{i=1}^{k-1} p_i$. This is given by function `B()` in the package. If the distribution is Dirichlet or Generalized Dirichlet, the closed form expression for the normalizing constant is used. If not, numerical methods are used[2].

[1] Here, a letter in a calligraphic font always denotes a function from $\wp(K)$ to the real numbers; it is a generalization of the vector of parameters $\boldsymbol{\alpha}$ used in the Dirichlet distribution; bold letters (such as $\mathbf{p}$) always denote $k$-tuples. Taking $\mathcal{F}$ as an example, the hyperdirichlet distribution itself is denoted $H(\mathcal{F})$ and its PDF would be $f\left(\mathbf{p}; \mathcal{F}\right)$.

[2] Certain special cases of the hyperdirichlet may be manipulated using multivariate polynomials so that closed-form expressions for the normalization constant become available (Altham 2009, page 88). However, further work would be required to translate Altham's insight into workable R idiom (Hankin 2008).

Numerical evaluation of Equation 7 is computationally expensive, especially when $k$ becomes large—Evans and Swartz (2000) and others refer to "the curse of dimensionality" when discussing the difficulty of integrating over spaces of large dimension.

The determination of p-values often requires evaluating integrals of this type, in addition to more computationally demanding integrals such as evaluated in section 3.1 on page 9. The **hyperdirichlet** package provides functionality to calculate p-values for a wide range of natural hypotheses, but many such calculations are prohibitively time consuming.

An alternative to p-values is furnished by the Method of Support (Edwards 1992), which requires no integration for its calculation; examples of this are provided in Section 3 below. It is anticipated that practitioners using the package will be able to choose between computationally expensive p-value calculation and the much faster assessment provided by the Method of Support.

*Moments*

Moments—that is, $\mathsf{E}\left(p_1^{m_1}\cdots p_k^{m_k}\right)$—are given by $B(\mathcal{F}+\mathcal{M})/B(\mathcal{F})$, where

$$\mathcal{M}(G)=\begin{cases} m_i & \text{if } G==\{i\}\,(1\leqslant i\leqslant k)\\ 0 & \text{otherwise.}\end{cases} \tag{8}$$

Updating a prior $H(\mathcal{F})$ in the light of observations is straightforward. If an observation $i$, drawn from a multinomial distribution, is made, then the posterior is $H\left(\mathcal{F}+\mathcal{S}_{\{i\}}\right)$, where

$$\mathcal{S}_X(G)=\begin{cases} 1 & \text{if } G==X\\ 0 & \text{otherwise.}\end{cases} \tag{9}$$

If the observation is restricted *a priori* to be in $G\subseteq K$, and subsequently specified to be amongst $C\subseteq G$, then the posterior is $H\left(\mathcal{F}+\mathcal{S}_C-\mathcal{S}_G\right)$.

*Restrictions*

Not every $\mathcal{F}$ corresponds to a normalizable $H(\mathcal{F})$, that is, a distribution with a finite integral. A sufficient condition is that for all nonempty $G\subseteq K, \sum_{H\subseteq G}\mathcal{F}(H)>0$. For example, for $k=4$,

$$\begin{aligned}
\alpha_1 &> 0\\
\alpha_2 &> 0\\
\alpha_3 &> 0\\
\alpha_4 &> 0\\
\alpha_1+\alpha_2+\alpha_{12} &> 0\\
\alpha_1+\alpha_3+\alpha_{13} &> 0\\
\alpha_1+\alpha_4+\alpha_{14} &> 0\\
\alpha_2+\alpha_3+\alpha_{23} &> 0\\
\alpha_2+\alpha_4+\alpha_{24} &> 0\\
\alpha_3+\alpha_4+\alpha_{34} &> 0\\
\alpha_1+\alpha_2+\alpha_3+\alpha_{12}+\alpha_{13}+\alpha_{23}+\alpha_{123} &> 0\\
\alpha_1+\alpha_2+\alpha_4+\alpha_{12}+\alpha_{14}+\alpha_{24}+\alpha_{124} &> 0\\
\alpha_1+\alpha_3+\alpha_4+\alpha_{13}+\alpha_{14}+\alpha_{34}+\alpha_{134} &> 0\\
\alpha_2+\alpha_3+\alpha_4+\alpha_{23}+\alpha_{24}+\alpha_{34}+\alpha_{234} &> 0
\end{aligned} \tag{10}$$

[function `is.proper()` in the package tests for normalizability].

If $\mathcal{F}(G) = 0$ whenever $|G| > 1$ then $H(\mathcal{F})$ reduces to a Dirichlet; likewise Equation 10 reduces to the standard Dirichlet restriction $\alpha_i > 0$ for $1 \leqslant i \leqslant k$.

In this paper I discuss this natural generalization of the Dirichlet distribution and introduce an R (R Development Core Team 2009) package, **hyperdirichlet**, that provides some numerical functionality.

*Generalizations of the Dirichlet distribution*

Previous generalizations of the Dirichlet distribution include the work of Bradley and Terry (1952), who considered rank analysis of incomplete designs. In the case of pairs, ranking is equivalent to choosing a winner from two items, their likelihood function would correspond to

$$\prod_{i<j} \frac{p_i^{n_{ij}}\, p_j^{n_{ji}}}{(p_i + p_j)^{n_{ij}+n_{ji}}} \tag{11}$$

in current notation (here there are a total of $n_{ij} + n_{ji}$ Bernoulli trials beween player $i$ and player $j > i$ of which $n_{ij}$ are won by player $i$). This is a special case of Equation 5.

Censored observations, in which the class of an object is specified to be one of a subset of $\{1, \ldots, k\}$, lead naturally to a likelihood function that is a generalization of Dirichlet's; a survey is given by Paulino (1991). Paulino and de Bragança Pereira (1995) present a comprehensive Bayesian methodology for censored observations and a simplified analysis of their sample dataset is provided *exempli gratia* in the package, documented under `paulino`.

A different generalization was presented by Connor and Mosimann (1969), who observed that the Dirichlet distribution was neutral[3] and proved that

$$f(\mathbf{p}) = \prod_{i=1}^{k-1} \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\,\Gamma(b_i)} \cdot p_k^{b_{k-1}-1} \prod_{i=1}^{k-1} \left[ p_i^{a_i-1} \left( \sum_{j=i}^{k} p_j \right)^{b_{i-1}-(a_i+b_i)} \right] \tag{12}$$

[function `gd()` in the package] is the most general form of a random variable with neutrality. Wong (1998) extended this work and showed that the generalized Dirichlet distribution was conjugate to a particular type of sampling experiment.

## 2. Prior information and the hyperdirichlet distribution

The Bayesian paradigm allows one to use prior information in the form of a prior distribution on the parameters. There are many types of prior information that are expressed in a natural way using the hyperdirichlet distribution and some examples are discussed here.

Consider four tennis players $P_1$ through $P_4$. When $P_i$ plays $P_j$ with $i \neq j$, the result is a single observation from a Bernoulli distribution with parameters $\left(\frac{p_i}{p_i+p_j}, \frac{p_j}{p_i+p_j}\right)$ (Zermelo 1929), where the $p_i$ are the unknown probabilities of victory; we require $\sum p_i = 1$.

[3] Consider a random vector $\mathbf{V} = (P_1, \ldots, P_k)$. Element $i, 1 \leqslant i < k$ is *neutral* if $P_i$ and $P_j / \left(1 - \sum_{k=1}^{i} P_k\right)$ are independent for $j > i$ (Connor and Mosimann 1969). A *completely neutral vector* is one all of whose elements are neutral. Note that the ordering of the vector is relevant: Thus neutrality of $\mathbf{V}$ does not imply neutrality of $\mathbf{V}' = (P_2, P_1, P_3, \ldots, P_k)$. If $\mathbf{V}$ is Dirichlet, then any permutation of $\mathbf{V}$ is neutral.

A Dirichlet prior would be proportional to $\prod_{i=1}^{4} p_i^{\alpha_i - 1}$ where $\alpha_i > 0$, but suppose our prior information is that $P_1$ and $P_2$ are considerably stronger than $P_3$ and $P_4$ (perhaps we know $P_1$ and $P_2$ to be strong squash players, and $P_3$ and $P_4$ weak badminton players—surely informative about the $p_i$) but remain ignorant of $P_1$'s strength relative to $P_2$, and of $P_3$'s strength relative to $P_4$.

Then an appropriate prior might be $\propto (p_1 + p_2)^{\gamma_{12}}$ where the magnitude of $\gamma_{12}$ reflects the strength of our prior beliefs. If $\gamma_{12}$ is large, then the probability density is small everywhere except near points with $p_1 + p_2 = 1$.

The best one could do with a standard Dirichlet prior would be to assign large values for $\alpha_1$ and $\alpha_2$ and small values for $\alpha_3$ and $\alpha_4$. But this would have the disadvantage that one would have to have firm beliefs about the relative strengths of $P_1$ and $P_2$, and in particular that a match between $P_1$ and $P_2$ would be a Bernoulli trial with unknown probability $p$, where $p$ is itself drawn from a beta distribution with parameters $(\alpha_1, \alpha_2)$. Thus $\mathsf{E}(p) = \alpha_1/(\alpha_1 + \alpha_2)$ and $\mathsf{VAR}(p) = \alpha_1\alpha_2 \big/ \big((\alpha_1 + \alpha_2)^2(\alpha_1 + \alpha_2 + 1)\big)$ [ie small if $\alpha_1, \alpha_2$ are large]; and one might not have sufficient information to make such an assertion—compare this with a prior $\propto (p_1 + p_2)^{\gamma_{12}}$ in which the density is uniform along lines of constant $p_1 + p_2$.

Situations where one has prior information that is not representable with a Dirichlet distribution arise frequently, especially when the identities of the various players are not known. The special case of $k = 3$ is readily visualized because the system possesses two degress of freedom and the PDF may be plotted on a triangular plot. In the context of the sports estimation problem above, an example of prior information might be that a knowledgeable person observed the players and noted that two were very much stronger than the third; he in fact reported that "the guy with a red shirt got hammered" (West 2008). But whether it was player 2 or player 3 who wore the red shirt is not known; and no information about the relative strengths of the two non-red wearing players is available. Figure 1 shows an example of how observations affect prior information in this case.

```
> null <- dev.off()
```

# 3. Examples

This section presents the **hyperdirichlet** package in use. Examples drawn from diverse disciplines are given.

## 3.1. Chess

Many attributes of the hyperdirichlet distribution are evident in the simplest non-trivial case, that of $k = 3$. This case is also facilitated by the fact that, having two degrees of freedom, the distribution may be readily visualized. In addition, the normalization factor is easily evaluated, the integrand having arity two.

Consider Table 1 in which matches between three chess players are tabulated; this dataset has been used by West and Hankin (2008).

The likelihood function is

$$C\frac{p_1^{30}p_2^{36}p_3^{22}}{(p_1+p_2)^{35}\,(p_2+p_3)^{35}\,(p_1+p_3)^{18}}$$

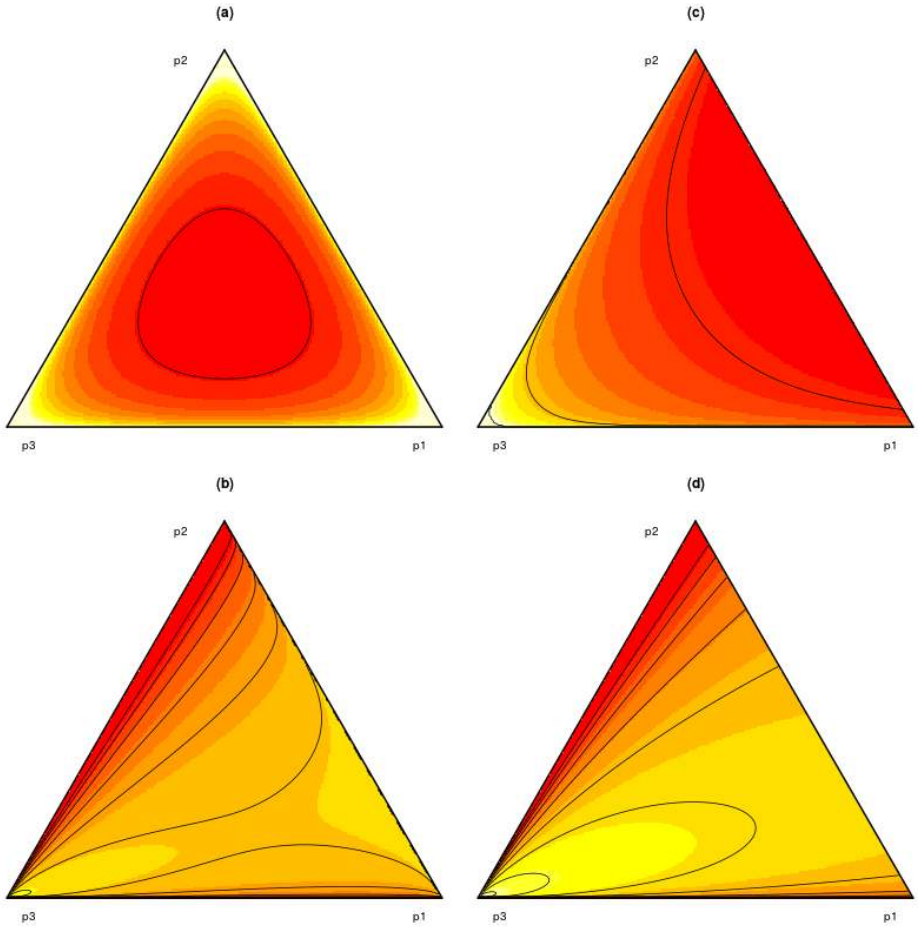


Figure 1: Density plots for the three-way hyperdirichlet distribution corresponding to different information sets. (a), prior PDF $\propto \left[\frac{p_1p_2p_3}{(p_1+p_2)(p_1+p_3)(p_2+p_3)}\right]^\alpha$ with $\alpha = 0.1$ corresponding to one player being known to be weaker than the other two; see how the high-probability region adheres to the edges of the triangle, thus implying that at least one player is weak. (b), posterior PDF following the observation that $p_1$ beat $p_2$ 7 times out of 10 (note the induced asymmetry between $p_1$ and $p_2$). (c), prior PDF $\propto \left[\frac{p_1p_2}{(p_1+p_2)^2}\right]^\alpha$, again with $\alpha = 0.1$, corresponding to $p_3$ being good and one (but not both) of $p_1$ or $p_2$ being good. (d), posterior, again following $p_1$ beating $p_2$ 7 times out of 10

| Topalov | Anand | Karpov | total |
|---|---|---|---|
| 22 | 13 | - | 35 |
| - | 23 | 12 | 35 |
| 8 | - | 10 | 18 |
| 30 | 36 | 22 | 88 |

Table 1: Results of 88 chess matches (dataset `chess` in the **aylmer** package) between three Grandmasters; entries show number of games won up to 2001 (draws are discarded). Topalov beats Anand 22-13; Anand beats Karpov 23-12; and Karpov beats Topalov 10-8

(the symbol '$C$' consistently stands for an undetermined constant), and this corresponds to a hyperdirichlet distribution, say $H(\mathcal{W})$

This dataset is included in the **aylmer** package; it may be loaded and coerced to an S4 object of class `hyperdirichlet`:

```
> data("chess")
> (w <- as.hyperdirichlet(chess))

    Topalov Anand Karpov params powers
[1]       0     0      0      0      0
[2]       0     0      1     23     22
[3]       0     1      0     37     36
[4]       0     1      1    -35    -35
[5]       1     0      0     31     30
[6]       1     0      1    -18    -18
[7]       1     1      0    -35    -35
[8]       1     1      1      0      0

Normalizing constant not known
```

thus R object `w` corresponds to $H(\mathcal{W})$. This simple example shows how a matrix, each row of which corresponds to repeated multinomial trials (here restricted to two outcomes), may be coerced to a `hyperdirichlet` object. Each output line of the print method corresponds to a subset of $\{p_1, p_2, p_3\}$; in columns 1-3, `0` means "not included" and `1` means "included"; thus, for example, the second line shows that Karpov won 22 (=10+12) games overall; and the fourth line shows that Anand and Karpov played 35 games.

The final two columns show the parameters and the powers of the $2^k = 8$ subsets respectively. Although these two columns give identical information, having both displayed simultaneously avoids much confusion in practice.

```
null device
          1
```

The normalizing constant $B$ is as yet unknown; it is unevaluated by default as its calculation is numerically expensive, especially when $k$ becomes large. The R idiom to calculate $B$ is

```
> (w <- as.hyperdirichlet(w , calculate_NC = TRUE))

    Topalov Anand Karpov params powers
[1]       0     0      0      0      0
[2]       0     0      1     23     22
[3]       0     1      0     37     36
[4]       0     1      1    -35    -35
[5]       1     0      0     31     30
[6]       1     0      1    -18    -18
[7]       1     1      0    -35    -35
```

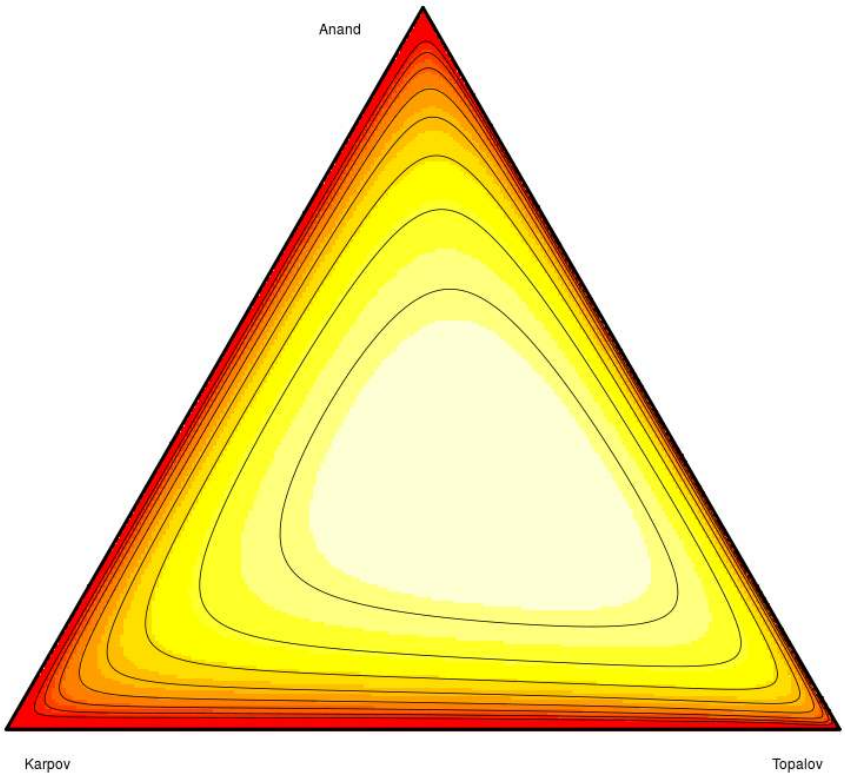


Figure 2: Support function for the three chess players of Table 1. Each player has an associated $p$, and we demand $p_1 + p_2 + p_3 = 1$. When player $i$ plays player $j \neq i$, the outcome is a Bernoulli trial with parameter $p_i/(p_i + p_j)$. Each labelled corner corresponds to a canonical basis vector; the top corner, for example, is point $(0, 1, 0)$: Anand wins all games (this point has zero likelihood as the dataset includes games in which Anand lost). Note that the support is unimodal

```
[8]       1     1       1       0       0

Normalizing constant not known
```

Thus object `w` now includes the normalizing constant.

This allows one to test various hypotheses using the standard methodology. For example, consider $H_0 : \mathbf{p} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$. The p-value for such a test is the integrated probability density, the integration proceeding over regions 'more extreme' (that is, regions with lower likelihood) than $H_0$. The R idiom would be

```
> f <- function(p){dhyperdirichlet(p, w) > dhyperdirichlet(rep(1/3, 3), w)}
> calculate_B(w, disallowed=f) / B(w)

[1] 0.3951652
```

Here, function `calculate_B()` integrates over the domain of the distribution, but excluding regions where `f()` returns `TRUE`. In this case, the integration proceeds over regions of the simplex that are more extreme than $H_0$, where a point is held to be 'more extreme' if its likelihood is lower than that of $H_0$. The test has a p-value of about 0.395, indicating that there is insufficient evidence to reject $H_0$ at the 5% level (in practice one would use function `probability()` which achieves the same result more compactly).

This functionality can be applied in a slightly different context. If `w` is interpreted as a probability density function with domain $\mathbf{p} = (p_1, p_2, p_3)$ where $\sum p_i = 1$, it is straightforward to use the Bayesian paradigm (taking a uniform prior for simplicity) to estimate the probability that $\mathbf{p}$ lies within any specified region. For example, the probability that Topalov is indeed a better player than Anand is merely the probability that $\mathbf{p} \in \{\mathbf{p} | p_1 \geqslant p_2\}$. This is given by

$$\frac{\displaystyle\int_{\mathbf{p}\geqslant\mathbf{0},\ p_1+p_2\leqslant 1,\ p_1\geqslant p_2} \left(\prod_{i=1}^{3} p_i\right)^{-1} \prod_{G\in\wp(\{1,2,3\})} \left(\sum_{i\in G} p_i\right)^{\mathcal{W}(G)} d\,(p_1, p_2)}{\displaystyle\int_{\mathbf{p}\geqslant\mathbf{0},\ p_1+p_2\leqslant 1} \left(\prod_{i=1}^{3} p_i\right)^{-1} \prod_{G\in\wp(\{1,2,3\})} \left(\sum_{i\in G} p_i\right)^{\mathcal{W}(G)} d\,(p_1, p_2)} \tag{13}$$

which may be evaluated with function `probability()`:

```
> T.lt.A <- function(p){p[1] < p[2]}
> probability(w , disallowed = T.lt.A)

[1] 0.7011418
```

Note that this is *not* the probability that Topalov would beat Anand in a game. The figure is the posterior probability that the Bernoulli parameter for such a game would exceed 0.5 (recall that uncertain probabilities are held to be random variables in the Bayesian paradigm).

Examples are given below which illustrate inferential techniques that do not require the value of the normalizing constant (or indeed any integral) to be evaluated.

| icon | | | | | | |
|---|---|---|---|---|---|---|
| NB | L | PB | THC | OA | WAIS | total |
| 5 | 3 | - | 4 | - | 3 | 15 |
| 3 | - | 5 | 8 | - | 2 | 18 |
| - | 4 | 9 | 2 | - | 1 | 16 |
| 1 | 3 | - | 3 | 4 | - | 11 |
| 4 | - | 5 | 6 | 3 | - | 18 |
| - | 4 | 3 | 1 | 3 | - | 11 |
| 5 | 1 | - | - | 1 | 2 | 9 |
| 5 | - | 1 | - | 1 | 1 | 8 |
| - | 9 | 7 | - | 2 | 0 | 18 |
| 23 | 24 | 30 | 24 | 14 | 9 | 124 |

Table 2: Experimental results from O'Neill (2007) (dataset `icons` in the package): Respondents' choice of 'most concerning' icon of those presented. Thus the first row shows results from respondents presented with icons NB, L, THC, and WAIS; of the 15 respondents, 5 chose NB as the most concerning (see text for a key to the acronyms). Note the "0" in row 6, column 9: This option was available to the 18 respondents of that row, but none of them actually chose WAIS

### 3.2. Public perception of climate change

Lay perception of climate change is a complex and interesting process (Moser and Dilling 2007); the issue of immediate practical import is the engagement of non-experts by the use of "icons"[4] that illustrate different impacts of climate change.

In one study (O'Neill 2007), subjects are presented with a set of icons of climate change and asked to identify which of them they find most concerning. Six icons were used: PB [polar bears, which face extinction through loss of ice floe hunting grounds], NB [the Norfolk Broads, which flood due to intense rainfall events], L [London flooding, as a result of sea level rise], THC [the thermo-haline circulation, which may slow or stop as a result of anthropogenic modification of the water cycle], OA [oceanic acidification as a result of anthropogenic emissions of $CO_2$], and WAIS [the West Antarctic Ice Sheet, which is rapidly calving as a result of climate change].

Methodological constraints dictated that each respondent could be presented with a maximum of four icons. Table 2 (dataset `icons` in the package) shows the experimental results.

One natural hypothesis $H_F$ is that there exist $\mathbf{p} = (p_1, \ldots, p_6)$ with $\sum p_i = 1$ such that the probability of choosing icon $i$ is proportional to $p_i$; the subscript '$F$' indicates here and elsewhere that the $p_i$ may be freely chosen subject to their sum. The Aylmer test (West and Hankin 2008) shows that there is insufficient evidence to reject this hypothesis and we proceed on the assumption that such a $\mathbf{p}$ does in fact exist: This is the object of inference.

This paper follows Esty (1992), who gives an example drawn from the field of psephology. In his voting model, $k$ choices are evaluated by voters; the object of inference is the set $\mathbf{p} = (p_1, \ldots, p_k)$, where $\sum_{i=1}^{k} p_i = 1$. If the voter has evaluated nominee $j$, then nominee $j$ is selected with probability $p_j / \sum p_i$, where the summation is over all evaluated nominees.

[4] This word is standard in this context. An icon is a "representative symbol".

The maximum likelihood estimate for $\mathbf{p}$ is obtained straightforwardly in the package using `maximum_likelihood()` function; numerical techniques must be used because analytical solutions are not generally available[5].

```
> data("icons")
> ic <- as.hyperdirichlet(icons)
> (m.free <- maximum_likelihood(ic))

$MLE
     NB       L      PB     THC      OA    WAIS
0.25234 0.17362 0.22459 0.17009 0.11071 0.06865

$likelihood
[1] 9.990315e-77

$support
[1] -174.9974
```

Observe how the first element, NB—corresponding to the Norfolk Broads—is the largest of the six; this is consistent with the sociological arguments presented by O'Neill in which "local" issues dominate more distant concerns (the test took place in Norwich).

One natural line of enquiry is to test whether the finding that the point estimate of NB is the largest of the six is statistically significant.

There seem to be a number of closely related alternative hypotheses. Firstly, one may consider the hypothesis $H_F : \sum p_i = 1$, and compare with $H_1 : \sum p_i = 1, p_1 \leqslant \frac{1}{6}$. Recalling that the normalizing factor is difficult to calculate, it is possible to use the Method of Support (Edwards 1992).

The maximum support for $H_1$ is given by the following R idiom; the `disallowed` argument to `maximum_likelihood()` prevents the optimization routine searching outside the domain of $H_1$.

```
>  f1 <- function(p){p[1] > 1/6}
>  (m.f1 <- maximum_likelihood(ic , disallowed=f1))

$MLE
     NB       L      PB     THC      OA    WAIS
0.16667 0.19406 0.25617 0.18303 0.12336 0.07670

$likelihood
[1] 7.359708e-78

$support
[1] -177.6056
```

[5]Even the very simplest nontrivial cases have complicated expressions for the maximum likelihood estimate: three dimensional hyperdirichlet distributions such as the `chess` dataset do possess an analytical expression for the MLE, but Maple's tightest simplification for it occupies over 23 sides of A4.

Observe that the MLE subject to $H_1$ is on the boundary of admissibility as (to within numerical accuracy) $p_1 = \frac{1}{6}$. The relevant statistic is thus

```
> m.free$support - m.f1$support

[1] 2.608181
```

indicating that the support at *any* point admissible under $H_1$ may be increased by 2.6 by the expedient of allowing the optimization to proceed freely over the domain of $H_F$. Edwards's criterion of 2 units of support per degree of freedom is thus met and $H_1$ may be rejected.

Secondly, one might consider $H_2 : \sum p_i = 1, p_1 > \max(p_2, \ldots, p_6)$; thus $p_1$ is held to be greater than all the others.

```
>  f2 <- function(p){p[1] > max(p[-1])}
>  (m.f2 <- maximum_likelihood(ic , disallowed=f2))

$MLE
     NB       L      PB     THC      OA    WAIS
0.23757 0.17396 0.23762 0.16989 0.11138 0.06958

$likelihood
[1] 9.173346e-77

$support
[1] -175.0828
```

Again observe that the MLE lies on the boundary of its restricted hypothesis $[p_1 == p_3]$. We have

```
> m.free$support - m.f2$support

[1] 0.08531408
```

indicating that there is insufficient evidence to reject $H_2$: There are points within the region of admissibility of $H_2$ whence one can gain only a small amount of support (viz. 0.0853) by optimizing over the whole of $H_F$.

*Low frequency responses*

O'Neill argues that the fifth and sixth icons are both considered by her respondents to be "remote" (cf the first, which is definitely local). Thus one might consider $H_3 : \sum p_i = 1, p_5 + p_6 \geqslant \frac{1}{3}$:

```
>  f3 <- function(p){sum(p[5:6]) > 1/3}
>  m.f3 <- maximum_likelihood(ic , disallowed=f3)
>  m.free$support - m.f3$support
```

| $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ | $p_6$ | $p_7$ | $p_8$ | $p_9$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 | NA | 1 | 0 | 0 | NA | 1 | NA |
| NA | NA | 1 | 1 | 0 | 1 | 0 | 0 | NA |
| NA | NA | 1 | 1 | 0 | NA | 1 | 0 | NA |
| NA | 1 | 1 | 0 | 0 | NA | 1 | 1 | NA |
| 1 | 1 | 1 | 0 | 0 | 0 | NA | NA | NA |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

Table 3: First five results from a sports league comprising five players, $p_1$ to $p_9$; dataset `volleyball` in the package. On any given line, a '1' denotes that that player was on the winning side, a '0' that he was on the losing side, and `NA` that he did not take part for that game

```
[1] 7.711396
```

Thus indicating that the observed low frequencies of respondents choosing `OA` and `WAIS` are unlikely to be due to chance, consistent with O'Neill's sociological analysis.

As a final example, consider $H_4 : \sum p_i = 1, \max\{p_5, p_6\} \geqslant \min\{p_1, p_2, p_3, p_4\}$. This corresponds to an assertion that the maximum of the two distant icons is less than any local icon. The support for this hypothesis is about 3.16, indicating that one may reject $H_4$.

The same techniques can be applied to any dataset in which repeated conditional multinomial observations are made; observe that a numerical value for the normalizing constant is not necessary for this type of inference.

### 3.3. Team sports

Table 3 shows the result of a sports league in which up to $n = 9$ players compete. A 'game' is a disjoint pair of subsets of $K = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ together with an identification of one of these subsets as the winning side.

Thus the likelihood function for the first two games would be

$$C \cdot \frac{p_1 + p_4 + p_8}{p_1 + p_2 + p_4 + p_5 + p_6 + p_8} \cdot \frac{p_3 + p_4 + p_6}{p_3 + p_4 + p_5 + p_6 + p_7 + p_8},$$

on the assumption of independence. The dataset of results provided with the package corresponds to a very flat likelihood curve; unrealistically large datasets of this type are apparently necessary to reject alternative hypotheses of practical interest. The analysis below is based on a synthetic dataset of 4000 games in which the players' strengths are proportional to $\left(1, \frac{1}{2}, \frac{1}{3}, \cdots, \frac{1}{9}\right)$: Zipf's law (1949).

The first step is to estimate the strengths of the players:

```
> data("volleyball")
> v.HF <- maximum_likelihood(vb_synthetic))
> v.HF$MLE
```

```
    p1     p2     p3     p4     p5     p6     p7     p8     p9
0.3044 0.1772 0.1005 0.0929 0.0733 0.0841 0.0589 0.0419 0.0668
```

Given that the actual strengths follow Zipf's law, the error in the estimate is given by:

```
> zipf(9) - v.HF$MLE
```

```
     p1      p2      p3      p4      p5      p6      p7      p8      p9
 0.0490 -0.0004  0.0173 -0.0046 -0.0026 -0.0251 -0.0084  0.0023 -0.0275
```

showing that the estimate is quite accurate: Esty (1992) points out that numerical means will find the maximum likelihood estimate easily if the data is irreducible, as here.

One topic frequently of interest in this context is the ranking of the players. On the basis of this point estimate, one might assert that $p_1 \geqslant p_2 \geqslant p_3 \geqslant p_4$; observe that the ranks of the MLE are not correct beyond the fifth, even with the large amount of data used. How strong is the evidence for this ranking?

```
> o <- function(p){all(order(p[1:4])==1:4)}
> v.HA <- maximum_likelihood(vb_synthetic, disallowed=o, start_p=1:9)
```

(the `start_p` argument specifies a non-disallowed start point for the optimization routine). Then

```
> v.HF$support - v.HA$support
```

```
[1] 1.576043
```

shows that there is no strong statistical evidence to support the assertion that the players are ranked as in the MLE: There exist regions of parameter space with a different ranking for which less than two units of support are lost.

*Tennis*

The above analysis assumed that the strength of a team is proportional to the sum of the strengths of the players.

However, many team sports appear to include an element of team cohesion; Carron, Bray, and Eys (2002) suggest that there is a 'strong relationship' between cohesion and team success.

In the current context, the simplest team is a pair. Doubles tennis appears to be a particularly favourable example: "if the two partners coordinate... well, they force their opponents to execute increasingly difficult shots" (Cayer 2004). Note that Cayer's assertion is independent of the individual players' strengths.

The hyperdirichlet distribution affords a direct way of assessing and quantifying such claims, using the likelihood function induced by teams' scorelines directly. Consider Table 4, in which results from repeated doubles tennis matches are shown. The likelihood function is

$$\begin{aligned}\mathcal{L}(p_1,p_2,p_3,p_4) = \\ C\cdot(p_1+p_2)^9(p_3+p_4)^2\cdot(p_1+p_3)^4(p_2+p_4)^4\cdot(p_1+p_4)^6(p_2+p_3)^7\cdot \\ \frac{p_1^{10}p_3^{14}}{(p_1+p_3)^{24}}\cdot\frac{p_2^{12}p_3^{14}}{(p_2+p_3)^{26}}\cdot\frac{p_1^{10}p_4^{14}}{(p_1+p_4)^{24}}\cdot\frac{p_2^{11}p_4^{10}}{(p_2+p_4)^{21}}\cdot\frac{p_3^{13}p_4^{13}}{(p_3+p_4)^{26}}\end{aligned}$$

| match | score |
|---|---|
| $\{P_1, P_2\}$ vs $\{P_3, P_4\}$ | 9-2 |
| $\{P_1, P_3\}$ vs $\{P_2, P_4\}$ | 4-4 |
| $\{P_1, P_4\}$ vs $\{P_2, P_3\}$ | 6-7 |
| $\{P_1\}$ vs $\{P_3\}$ | 10-14 |
| $\{P_2\}$ vs $\{P_3\}$ | 12-14 |
| $\{P_1\}$ vs $\{P_4\}$ | 10-14 |
| $\{P_2\}$ vs $\{P_4\}$ | 11-10 |
| $\{P_3\}$ vs $\{P_4\}$ | 13-13 |

Table 4: Results from singles (lines 4-8) and doubles (lines 1-3) tennis matches among four players, $P_1$ to $P_4$; dataset `doubles` in the package. Note how $P_1$ and $P_2$ dominate the other players when they play together (winning 9 games out of 11) but are otherwise undistinguished

where $\sum p_i = 1$ is understood. Players $P_1$ and $P_2$ are known to play together frequently and one might expect them to win more often when they play together than by chance. Indeed, each matching has a scoreline of roughly 50-50, except $\{P_1, P_2\}$ vs $\{P_3, P_4\}$, which results in a win for $\{P_1, P_2\}$ 9 times out of 11. Is this likely to have arisen if team cohesion is in fact absent?

Consider the following likelihood function:

$$\mathcal{L}\left(p_{\mathfrak{g}}; p_1, p_2, p_3, p_4\right) = C \cdot \left(p_1 + p_2 + p_{\mathfrak{g}}\right)^9 \left(p_3 + p_4\right)^2 \cdot \frac{\left(p_1 + p_3\right)^4 \left(p_2 + p_4\right)^4}{\left(p_1 + p_2 + p_3 + p_4\right)^8} \cdot \ldots \tag{14}$$

which formalizes the effectiveness of team cohesion in terms of a 'ghost' player with skill $p_{\mathfrak{g}}$ who accounts for the additional skill arising when $P_1$ and $P_2$ play together; the null is then simply $p_{\mathfrak{g}} = 0$.

It is straightforward to apply the method of support. Function `maximum_likelihood()` takes a `zero` argument that specifies which components of the $p_i$ are to be constrained at zero; here we specify that $p_{\mathfrak{g}} = 0$:

```
> data("doubles")
> maximum_likelihood(doubles)$support - maximum_likelihood(doubles,zero=5)$support

[1] 2.773369
```

thus one may reject the hypothesis the ghost player has zero strength. The inference is that $P_1$ and $P_2$ when playing together are stronger than one would expect on the basis of their performance either in singles matches, or doubles partnering with other players: The scoreline provides strong objective evidence that team cohesion is operating.

This technique may be applied to any of the datasets considered in this paper, and in the context of scorelines the ghost may be any factor whose existence is in doubt. Negative factors (for example, a member of the audience whose presence adversely affects one competitor's

performance) may be assessed by recasting the negative effect as a helpful ghost whose skill is added to the opposition's.

## 4. Conclusions

The Dirichlet distribution is conjugate to the multinomial distribution. This paper presents a generalization of the Dirichlet distribution which is conjugate to a more general class of observations that arise naturally in a variety of contexts. The distribution is dubbed 'hyper-dirichlet' as it is clearly the most general form of its type.

The **hyperdirichlet** package of R routines for analysis of the distribution is introduced and examples of the package in use are given.

One difficulty in using the distribution is that there does not appear to be a closed-form analytical expression for the normalizing constant; numerical methods must be used. The normalizing constant is difficult to calculate numerically, especially for distributions of large dimension.

The normalizing constant is needed for conventional statistical tests; but its evaluation is not necessary for the Method of Support, which is used to test a wide variety of plausible and interesting hypotheses using datasets drawn from a range of disciplines.

## Acknowledgement

I would like to acknowledge the many stimulating and helpful comments made by the R-help list while preparing this software.

## References

Altham PME (2009). Worksheet 20, URL http://www.statslab.cam.ac.uk/~pat/misc.ps.

Bradley RA, Terry ME (1952). "The Rank Analysis of Incomplete Block Designs I. The Method of Paired Comparisons." *Biometrika*, **39**, 324–345.

Carron AV, Bray SR, Eys MA (2002). "Team Cohesion and Team Success in Sport." *Journal of Sports Sciences*, **20**, 119–126.

Cayer L (2004). *Doubles Tennis Tactics*. Human Kinetics.

Connor RJ, Mosimann JE (1969). "Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution." *Journal of the American Statistical Association*, **64**(325), 194–206.

Edwards AWF (1992). *Likelihood (Expanded Edition)*. John Hopkins.

Esty WW (1992). "Votes or Competitions Which Determine a Winner by Estimating Expected Plurality." *Journal of the American Statistical Association*, **87**(418), 373–375.

Evans M, Swartz T (2000). *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford University Press.

Hankin RKS (2008). "Programmers' Niche: Multivariate polynomials in R." *R News*, **8**(1), 41–45. URL http://CRAN.R-project.org/doc/Rnews/.

Hankin RKS (2010). "A Generalization of the Dirichlet Distribution." *Journal of Statistical Software*, **33**(11), 1–18. URL http://www.jstatsoft.org/v33/i11/.

Moser SC, Dilling L (eds.) (2007). *Creating a Climate for Change: Communicating Climate Change and Facilitating Social Change*. Cambridge University Press.

O'Neill S (2007). *An Iconic Approach to Representing Climate Change*. Ph.D. thesis, School of Environmental Science, University of East Anglia.

Paulino CDM (1991). "Analysis of Incomplete Categorical Data: A Survey of the Conditional Maximum Likelihood and Weighted Least Squares Approaches." *Brazilian Journal of Probability and Statistics*, **5**, 1–42.

Paulino CDM, de Bragança Pereira CA (1995). "Bayesian Methods for Categorical Data Under Informative General Censoring." *Biometrika*, **82**(2), 439–446.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

West LJ (2008). Personal Communication. Verbal report from Bournemouth Tennis Centre, Dorset, UK.

West LJ, Hankin RKS (2008). "Exact Tests for Two-Way Contingency Tables with Structural Zeros." *Journal of Statistical Software*, **28**(11). URL http://www.jstatsoft.org/v28/i11/.

Wong TT (1998). "Generalized Dirichlet Distribution in Bayesian Analysis." *Applied Mathematics and Computation*, **97**, 165–181.

Zermelo E (1929). "Die Berechnung der Turnier-Ergebnisse als ein Maximum-problem der Wahrscheinlichkeitsrechnung." *Mathematische Zeitschrift*, **29**, 436–460.

Zipf G (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley.

## Affiliation:

Robin K. S. Hankin
Auckland University of Technology
E-mail: hankin.robin@gmail.com