Integration in the hyper2 package

Robin K. S. Hankin
Auckland University of Technology

Abstract

The hyper2 package presented a new formulation of the hyperdirichlet package, offering speed advantages and the ability to deal with higher-dimensional datasets. However, hyper2 was based on likelihood methods and as originally uploaded did not have the ability to integrate over the unit-sum simplex. This functionality has now been incorporated into the package which is documented here, by reproducing earlier analysis.

Keywords: Dirichlet distribution, hyperdirichlet, hyper2, combinatorics, R, multinomial distribution, constrained optimization, integration, simplex, unit-sum constraint.

1. Introduction

The hyper2 package (Hankin 2017) presented a new formulation of the hyperdirichlet distribution (Hankin 2010) which offered speed advantages over the original hyperdirichlet package, and the ability to deal with higher-dimensional datasets. However, hyper2 was based on likelihood methods and as originally uploaded did not have the ability to integrate over the unit-sum simplex. This functionality has now been incorporated into the package which is documented here, by reproducing earlier analysis.

2. Chess

Consider Table 1 in which matches between three chess players are tabulated; this dataset was analysed by Hankin (2010).

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

(the symbol ‘C’ consistently stands for an undetermined constant). This likelihood function is provided in the hyper2 package as the chess dataset:

```r
> chess
log(Anand^36 * (Anand + Karpov)^35 * (Anand + Topalov)^35 * Karpov^22 * (Karpov + Topalov)^18 * Topalov^30)
```

We can calculate the normalizing constant:
2 Integration in the hyper2 package

<table>
<thead>
<tr>
<th>Topalov</th>
<th>Anand</th>
<th>Karpov</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>13</td>
<td>-</td>
<td>35</td>
</tr>
<tr>
<td>-</td>
<td>23</td>
<td>12</td>
<td>35</td>
</tr>
<tr>
<td>8</td>
<td>-</td>
<td>10</td>
<td>18</td>
</tr>
<tr>
<td>30</td>
<td>36</td>
<td>22</td>
<td>88</td>
</tr>
</tbody>
</table>

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.

$$B(\text{chess})$$

[1] 1.442828e-28

comparing well with the value given by the hyperdirichlet package of $1.47 \times 10^{-28}$. Hankin (2010) went on to calculate the $p$-value for $H_0: p = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ as 0.395, a calculation which may be performed in the hyper2 package as follows:

```r
f <- function(p){loglik(indep(p),chess) > loglik(c(1,1)/3,chess)}
probability(chess, disallowed=f,tol=0.1)
```

[1] 0.4099

Again comparing well with the older result (smaller values of tol give closer agreement at the expense of increased computation time). Finally, we can calculate the probability that Topalov is a better player than Anand:

```r
probability(chess, disallowed=T.lt.A,tol=0.01)
```

[1] 0.7123

again showing reasonable agreement with the 2010 value of 0.701.

3. Verification

In a breathtaking display of arrogance and/or incompetence, Hankin (2010) did not actually provide any evidence that the integration suite of hyperdirichlet was accurate. Here I compensate for that inexcusable lapse by comparing numerical results with analytical formulae. Consider the standard Dirichlet distribution:

$$\frac{p_1^{\alpha_1-1} \cdots p_k^{\alpha_k-1}}{B(\alpha_1, \ldots, \alpha_k)}$$

where it is understood that the $p_i > 0$ and $\sum p_i = 1$; here $B = \frac{\Gamma(\sum \alpha_i)}{\prod \Gamma(\alpha_i)}$ is the normalization constant. We can verify that hyper2::B() is operating as expected for the case $\alpha = (1, 2, 3, 4)$:
Further, consider a Dirichlet distribution with $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 3$. Then, by symmetry, the probability that $p_1 < p_2$ should be exactly $\frac{1}{2}$:

```r
f <- function(p){p[1]<p[2]}
H <- dirichlet(alpha=c(a=3,b=3,c=3,d=3))
probability(H,f, tol=0.1)
```

[1] 0.4973

(compare exact value of 0.5; note the loose tolerance of 0.1, needed to keep computational time short—the integrand has a severe discontinuity which is computationally expensive to integrate across). Further, $P(p_1 < p_2 < p_3)$ should be exactly $\frac{1}{6}$:

```r
1-probability(H,disallowed=g, tol=0.1)
```

[1] 0.1866

(compare exact value of 0.1666).

4. More results: icons dataset

Consider the icons dataset, shown in table 2, and the following hypotheses, again following Hankin (2010), and reproduced here for convenience.

```r
icons
```

```
log(L^-24 * (L + NB + OA + THC)^-20 * (L + NB + OA + WAIS)^-9 * (L + NB + TH + WAIS)^-15 * (L + OA + PB + THC)^-11 * (L + OA + PB + WAIS)^-18 * (L + PB + THC + WAIS)^-16 * NB^-32 * (NB + OA + PB + THC)^-18 * (NB + OA + PB + WAIS)^-8 * (NB + PB + TH + WAIS)^-18 * OA^-14 * PB^-30 * THC^-24 * WAIS^-9)
```

```r
maxp(icons)
```

```
NB  L  PB  THC  OA  WAIS
0.25230 0.17364 0.22458 0.17011 0.11069 0.06867
```
Integration in the hyper2 package

<table>
<thead>
<tr>
<th></th>
<th>NB</th>
<th>L</th>
<th>PB</th>
<th>THC</th>
<th>OA</th>
<th>WAIS</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>3</td>
<td>9</td>
<td></td>
<td>1</td>
<td>1</td>
<td>15</td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td></td>
<td>1</td>
<td>16</td>
<td>18</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td></td>
<td>2</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>3</td>
<td>4</td>
<td></td>
<td>3</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>-</td>
<td>3</td>
<td></td>
<td>3</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td></td>
<td>3</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>9</td>
<td>7</td>
<td>-</td>
<td>2</td>
<td>0</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>24</td>
<td>30</td>
<td>24</td>
<td>14</td>
<td>9</td>
<td>124</td>
<td></td>
</tr>
</tbody>
</table>

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 9, column 6: this option was available to the 18 respondents of that row, but none of them actually chose WAIS.

For reference, the other hypotheses were:

- \( H_1: p_1 \geq \frac{1}{6} \)
- \( H_2: p_1 \geq \max \{p_2, \ldots, p_6\} \)
- \( H_3: p_5 + p_6 \geq \frac{1}{3} \)
- \( H_4: \max \{p_5, p_6\} \geq \min \{p_1, p_2, p_3, p_4\} \)

```r
> f1 <- function(p){p[1] > 1/6}
> f2 <- function(p){p[1] > max(fillup(p)[-1])}
> f3 <- function(p){sum(fillup(p)[5:6]) > 1/3}
> f4 <- function(p){max(fillup(p)[1:2]) > min(fillup(p)[3:6])}
```

Here I will analyse just the first hypothesis, that is \( H_1: p_1 \leq \frac{1}{6} \) using the integration facilities of the hyper2 package, and compare with previous results. Here we perform a Bayesian analysis, made possible by the efficient coding of hyper2:

```r
probability/icons, disallowed=function(p){p[1] > 1/6}, tol=0.1)
```

\[1\] 0.01502

See how the disallowed region is the expected bit of the parameter space. Thus the probability that the \( p_i \) are unexpected (that is, \( p_1 < 1/6 \)) is about 1.5% or conversely, \( P(H_1) \approx 0.985 \). The likelihood ratio reported was about 2.608, which would correspond to a \( p \)-value of about

```r
> pchisq(2*2.608, df=1, lower.tail=FALSE)
```

\[1\] 0.01502
or just over 2% under an asymptotic distribution; thus this frequentist technique gives comparable strength of evidence for $H_1$ to the Bayesian approach.

5. Incomplete survey data

This section performs the analysis originally presented in Altham and Hankin (2010). The data, given here in table 4 arises from 69 medical malpractice claims, and are the two surgeons’ answers to the question: was there a communication breakdown in the hand-off between physicians caring for the patient?

<table>
<thead>
<tr>
<th>Reviewer 1</th>
<th>Reviewer 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>26</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>18</td>
</tr>
<tr>
<td>Missing</td>
<td>Missing</td>
<td>4</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>Total</strong></td>
<td><strong>69</strong></td>
</tr>
</tbody>
</table>

Table 3: Two surgeon reviews of malpractice claims data

<table>
<thead>
<tr>
<th>Reviewer 1</th>
<th>Reviewer 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>$y_{11}$</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>$y_{01}$</td>
</tr>
<tr>
<td>Missing</td>
<td>Yes</td>
<td>$u_{+1}$</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>Total</strong></td>
<td>$y_{+1}$</td>
</tr>
</tbody>
</table>

Table 4: Notation for the data

We may implement an appropriate likelihood function as follows:

```r
> H <- hyper2()
> H["t00"] <- 18
> H["t10"] <- 01
> H["t01"] <- 05
> H["t11"] <- 26
> H[c("t11", "t10")]) <- 2
> H[c("t01", "t00")]) <- 9
> H[c("t11", "t01")]) <- 4
> H[c("t10", "t00")]) <- 4
> H

log(t00^18 * (t00 + t01)^9 * (t00 + t10)^4 * t01^5 * (t01 + t11)^4 * t10 * (t10 + t11)^2 * t11^26)
```
Integration in the hyper2 package

(object H is provided as handover in the package). Then we may estimate the probability that reviewer 2 is more likely to give a ‘yes’ than reviewer 1 as follows:

```r
> free <- maxp(H, give=TRUE)
> m <- fillup(free$par)
> names(m) <- pnames(H)
> m
```

```
t00  t01  t10  t11
0.41955 0.11128 0.01799 0.45119
```

```r
> free$value
```

```
[1] -64.15
```

Then the constrained optimization:

```r
> obj <- function(p){-loglik(p,H)} # objective func
> gr <- function(p){-gradient(H,p)} # gradient, needed for speed
> UI <- rbind(diag(3),-1) # UI and CI specify constraints
> CI <- c(rep(0,3),-1) # p_i >= 0 and sum p_i <= 1
```

We will test \( H_A: p_2 < p_3 \) using the method of support.

```r
> constrained <- maxp(H, give=TRUE, fcm = rbind(c(0,-1,1)), fcv=0, maxtry=1e5)
> constrained
```

```
$par
[1] 0.42735779 0.06018069 0.06018069

$value
[1] -66.14478

$counts
function gradient
  318  43

$convergence
[1] 0

$message
NULL

$outer.iterations
[1] 2
```
Thus the support for $H_A$ is about $66.14478 - 64.14538 = 1.9999$, or almost exactly 2 units of support.

References


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