Integration in the hyper2 package
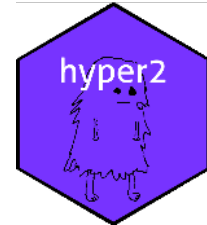
Robin K. S. Hankin

University of Stirling

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.

1. Introduction

To cite this work in publications, please use Hankin (2017). The **hyper2** package 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:

```
> chess
```

```
log(Anand^36 * (Anand + Karpov)^-35 * (Anand +
Topalov)^-35 * Karpov^22 * (Karpov + Topalov)^-18 *
Topalov^30)
```

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

We can calculate the normalizing constant:

```
B(chess)
```

```
[1] 1.442828e-28
```

comparing well with the value given by the **hyperdirichlet** package of 1.47×10^{-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:

```
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:

```
T.lt.A <- function(p){p[1]<p[2]}
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:

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

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 $\alpha = (1, 2, 3, 4)$:

```
> x <- c(a=1, b=2, c=3, d=4)  # needs a named vector
> ans1 <- B(dirichlet(alpha = x),tol=0.1)
> ans2 <- prod(gamma(x))/gamma(sum(x))
> c(numerical=ans1,theoretical=ans2)   # should agree

    numerical    theoretical
3.30092705e-05 3.30687831e-05
```

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 $\frac{1}{2}$:

```
> 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] 0.49730037
```

(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 $\frac{1}{6}$:

```
> g <- function(p){(p[1]<p[2]) & (p[2]<p[3])}
> 1-probability(H,disallowed=g,tol=0.1)

[1] 0.186621983
```

(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.

```
> icons

log(L^24 * (L + NB + OA + THC)^-20 * (L + NB + OA +
WAIS)^-9 * (L + NB + THC + 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 + THC + WAIS)^-18 *
OA^14 * PB^30 * THC^24 * WAIS^9)


> maxp(icons)

          NB            L           PB          THC
0.2523041150 0.1736443259 0.2245818764 0.1701128100
          OA         WAIS
0.1106860420 0.0686708306
```

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 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:

- $H_1 \colon p_1 \geqslant \frac{1}{6}$

- $H_2 \colon p_1 \geqslant \max\{p_2, \ldots p_6\}$

- $H_3 \colon p_5 + p_6 \geqslant \frac{1}{3}$

- $H_4 \colon \max\{p_5, p_6\} \geqslant \min\{p_1, p_2, p_3, p_4\}$

```
> 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 $H_1 \colon p_1 \leqslant \frac{1}{6}$ 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**:

```
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, $P(H_1) \simeq 0.985$. The likelihood ratio reported was about 2.608, which would correspond to a *p*-value of about

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

```
[1] 0.0223799725
```

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?

Reviewer 1	Reviewer 2			
	Yes	No	Missing	Total
Yes	26	1	2	29
No	5	18	9	32
Missing	4	4	0	8
Total	35	23	11	69

Table 3: Two surgeon reviews of malpractice claims data

Reviewer 1	Reviewer 2			
	Yes	No	Missing	Total
Yes	y_{11}	y_{10}	z_{1+}	$y_{1+} + z_{1+}$
No	y_{01}	y_{00}	z_{0+}	$y_{0+} + z_{0+}$
Missing	u_{+1}	u_{+0}	0	u_{++}
Total	$y_{+1} + u_{+1}$	$y_{+0} + u_{+0}$	z_{++}	n

Table 4: Notation for the data

We may implement an appropriate likelihood function as follows:

```
> 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 <- balance(H)
> H

log(t00^18 * (t00 + t01)^9 * (t00 + t01 + t10 +
t11)^-69 * (t00 + t10)^4 * t01^5 * (t01 + t11)^4 *
t10 * (t10 + t11)^2 * t11^26)
```

(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:

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

        t00          t01          t10          t11
0.4195489413 0.1112755384 0.0179871939 0.4511883264

> free$value

[1] -64.1453767
```

Then the constrained optimization:

```
> 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 $H_A\colon p_2 < p_3$ using the method of support.

```
> 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

$barrier.value
[1] 0.0001060435

$likes
 [1] -82.48451 -66.73119 -82.48454 -66.14478 -67.33553 -67.11853 -66.14607
 [8] -66.16411 -66.27162 -66.67668
```

Thus the support for H_A is about $66.14478 - 64.14538 = 1.9999$, or almost exactly 2 units of support.

References

Altham PME, Hankin RKS (2010). "Using recently developed software on a 2x2 table of matched pairs with incompletely classified data." *Journal of the Royal Statistical Society, series C*, **59**(2), 377–379.

Hankin RKS (2010). "A Generalization of the Dirichlet Distribution." *Journal of Statistical Software*, **33**(11), 1–18. doi:10.18637/jss.v033.i11.

Hankin RKS (2017). "Partial rank data with the **hyper2** package: likelihood functions for generalized Bradley-Terry models." *The R Journal*, **9**(2), 429–439.

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

Affiliation:

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