Estimating a Dirichlet distribution

Thomas P. Minka

November 15, 2000

Abstract

This note covers maximum-likelihood and leave-one-out estimation of a Dirichlet distribution from probability vectors and from counts. Estimation from counts is equivalent to estimating a compound multinomial distribution. In each case, a fixed-point algorithm and a Newton-Raphson algorithm is provided. Newton-Raphson is faster but more complex to implement since it requires stepsize control.

1 Estimation from probability vectors

The Dirichlet density is

$$\begin{aligned} p(\mathbf{p}) &\sim \mathcal{D}(\alpha_1, ..., \alpha_K) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k p_k^{\alpha_k - 1} && (1) \\ \text{where } p_k &> 0 && (2) \\ \sum_k p_k &= 1 && (3) \end{aligned}$$

If we observe a training set of probability vectors $D = \{\mathbf{p}_1, ..., \mathbf{p}_N\}$, then the ML estimate of α would maximize $p(D|\alpha) = \prod_i p(\mathbf{p}_i|\alpha)$. The log-likelihood is

$$\begin{aligned} \log p(D|\alpha) &= N \log \Gamma(\sum_k \alpha_k) - N \sum_k \log \Gamma(\alpha_k) + N \sum_k (\alpha_k - 1) \log \bar{p}_k && (4) \\ \text{where } \log \bar{p}_k &= \frac{1}{N} \sum_i \log p_{ik} && (5) \end{aligned}$$

This objective is convex in α since Dirichlet is in the exponential family. A direct convexity proof has also been given by Ronning (1989). The gradient of the log-likelihood is

$$g_k = \frac{d \log p(D|\alpha)}{d\alpha_k} = N\Psi(\sum_k \alpha_k) - N\Psi(\alpha_k) + N \log \bar{p}_k \qquad (6)$$

where Ψ is the digamma function. As always with the exponential family, the MLE will set the expected sufficient statistics equal to the observed sufficient statistics. In this case, the expected sufficient statistics are

$$E[\log p_k] = \Psi(\alpha_k) - \Psi(\sum_k \alpha_k) \qquad (7)$$

and the observed sufficient statistics are $\log \bar{p}_k$. The MLE can be computed by iterating the fixed-point equation

$$\Psi(\alpha_k^{new}) = \Psi(\sum_k \alpha_k^{old}) + \log \bar{p}_k \tag{8}$$

Each iteration provably increases likelihood, because it maximizes a lower bound; see appendix A. This algorithm requires inverting the Ψ function; see appendix C.

Another approach is Newton iteration. The Hessian of the log-likelihood is

$$\begin{aligned}
\frac{d \log p(D|\alpha)}{d\alpha_k^2} &= N\Psi'(\sum_k \alpha_k) - N\Psi'(\alpha_k) && (9)\\
\frac{d \log p(D|\alpha)}{d\alpha_k d\alpha_j} &= N\Psi'(\sum_k \alpha_k) \qquad (k \neq j) && (10)
\end{aligned}$$

where Ψ' is the trigamma function. The Hessian can be written in matrix form as

$$\begin{aligned}
\mathbf{H} &= \mathbf{Q} + \mathbf{1}\mathbf{1}^{\mathrm{T}} z && (11)\\
q_{jk} &= -N\Psi'(\alpha_k)\delta(j-k) && (12)\\
z &= N\Psi'(\sum_k \alpha_k) && (13)
\end{aligned}$$

One Newton step is therefore

$$\begin{aligned}
\alpha^{\text{new}} &= \alpha^{\text{old}} - \mathbf{H}^{-1}\mathbf{g} && (14)\\
\mathbf{H}^{-1} &= \mathbf{Q}^{-1} - \frac{\mathbf{Q}^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathbf{Q}^{-1}}{1/z + \mathbf{1}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{1}} && (15)\\
(\mathbf{H}^{-1}\mathbf{g})_k &= \frac{g_k - b}{q_{kk}} && (16)\\
\text{where } b &= \frac{\mathbf{1}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{g}}{1/z + \mathbf{1}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{1}} = \frac{\sum_j g_j/q_{jj}}{1/z + \sum_j 1/q_{jj}} && (17)
\end{aligned}$$

The same Newton algorithm was given by Ronning (1989) and Naryanan (1991). Naryanan also derives a stopping rule for the iteration.

An approximate MLE, useful for initialization, is given by finding the density which matches the moments of the data. The first two moments of the density are

$$\begin{aligned}
E[p_k] &= \frac{\alpha_k}{\sum_k \alpha_k} && (18)\\
E[p_k^2] &= E[p_k]\frac{1+\alpha_k}{1+\sum_k \alpha_k} && (19)\\
\sum_k \alpha_k &= \frac{E[p_1] - E[p_1^2]}{E[p_1^2] - E[p_1]^2} && (20)
\end{aligned}$$

Multiplying (20) and (18) gives a formula for α_k in terms of moments. Equation (20) uses p_1, but any other p_k could also be used to estimate $\sum_k \alpha_k$. Ronning (1989) suggests instead using all of the p_k's via

$$
\begin{aligned}
\mathrm{var}(p_k) &= \frac{E[p_k](1-E[p_k])}{1+\sum_k \alpha_k} && (21)\\
\log \sum_k \alpha_k &= \frac{1}{K-1}\sum_{k=1}^{K-1} \log\left(\frac{E[p_k](1-E[p_k])}{\mathrm{var}(p_k)} - 1\right) && (22)
\end{aligned}
$$

Another approximate MLE, specifically for the case $K = 2$, is given by Johnson & Kotz (1970):

$$
\begin{aligned}
\alpha_1 &= \frac{1}{2}\frac{1-\bar{p}_2}{1-\bar{p}_1-\bar{p}_2} && (23)\\
\alpha_2 &= \frac{1}{2}\frac{1-\bar{p}_1}{1-\bar{p}_1-\bar{p}_2} && (24)
\end{aligned}
$$

2 Estimation from counts

Sometimes we only observe a multinomial sample $\mathbf{x}_i$, of length n_i, from each $\mathbf{p}_i$. The marginal probability of a sample $\mathbf{x}$ is

$$
\begin{aligned}
p(\mathbf{x}|\alpha) &= \int_{\mathbf{p}} p(\mathbf{x}|\mathbf{p})p(\mathbf{p}|\alpha) && (25)\\
&= \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(n+\sum_k \alpha_k)} \prod_k \frac{\Gamma(n_k+\alpha_k)}{\Gamma(\alpha_k)} && (26)\\
n_k &= \sum_j \delta(x_j - k) && (27)
\end{aligned}
$$

The ML estimate of α given $D = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ now maximizes

$$
\begin{aligned}
p(D|\alpha) &= \prod_i p(\mathbf{x}_i|\alpha) && (28)\\
&= \prod_i \left(\frac{\Gamma(\sum_k \alpha_k)}{\Gamma(n_i+\sum_k \alpha_k)} \prod_k \frac{\Gamma(n_{ik}+\alpha_k)}{\Gamma(\alpha_k)}\right) && (29)
\end{aligned}
$$

This is useful for estimating a compound multinomial distribution (Mosimann) from data. It also arises in "empirical Bayes" or "type II maximum likelihood" (Good) inference, where we wish to conclude something about the $\mathbf{p}_i$ but don't want to commit to any particular Dirichlet prior. The gradient of the log-likelihood is

$$
g_k = \frac{d\log p(D|\alpha)}{d\alpha_k} = \sum_i \Psi(\sum_k \alpha_k) - \Psi(n_i + \sum_k \alpha_k) + \Psi(n_{ik}+\alpha_k) - \Psi(\alpha_k) \qquad (30)
$$

The maximum can be computed via the fixed-point iteration

$$\alpha_k^{new} = \alpha_k \frac{\sum_i \Psi(n_{ik} + \alpha_k) - \Psi(\alpha_k)}{\sum_i \Psi(n_i + \sum_k \alpha_k) - \Psi(\sum_k \alpha_k)} \tag{31}$$

(see appendix B). Alternatively, the Hessian of the log-likelihood is

$$\begin{aligned}
\frac{d \log p(D|\alpha)}{d\alpha_k^2} &= \sum_i \Psi'(\sum_k \alpha_k) - \Psi'(n_i + \sum_k \alpha_k) + \Psi'(n_{ik} + \alpha_k) - \Psi'(\alpha_k) && (32)\\
\frac{d \log p(D|\alpha)}{d\alpha_k d\alpha_j} &= \sum_i \Psi'(\sum_k \alpha_k) - \Psi'(n_i + \sum_k \alpha_k) \qquad (k \neq j) && (33)
\end{aligned}$$

The Hessian can be written in matrix form as

$$\begin{aligned}
\mathbf{H} &= \mathbf{Q} + \mathbf{1}\mathbf{1}^{\mathrm{T}} z && (34)\\
q_{jk} &= \delta(j-k) \sum_i \Psi'(n_{ik} + \alpha_k) - \Psi'(\alpha_k) && (35)\\
z &= \sum_i \Psi'(\sum_k \alpha_k) - \Psi'(n_i + \sum_k \alpha_k) && (36)
\end{aligned}$$

from which a Newton step can be computed as before. The search can be initialized with the moment matching estimate where p_{ik} is approximated by n_{ik}/n_i.

Another approach is to reduce this problem to the previous one via EM; see appendix D.

A different method is to maximize the leave-one-out (LOO) likelihood instead of the true likelihood. The LOO likelihood is the product of the probability of each sample given the remaining data and the parameters. The LOO log-likelihood is

$$f(\alpha) = \sum_{ik} n_{ik} \log \left(\frac{n_{ik} - 1 + \alpha_k}{n_i - 1 + \sum_k \alpha_k} \right) = \sum_{ik} n_{ik} \log(n_{ik} - 1 + \alpha_k) - \sum_i n_i \log(n_i - 1 + \sum_k \alpha_k) \tag{37}$$

Note that it doesn't involve any special functions. The derivatives are

$$\begin{aligned}
\frac{df(\alpha)}{d\alpha_k} &= \sum_i \frac{n_{ik}}{n_{ik} - 1 + \alpha_k} - \frac{n_i}{n_i - 1 + \sum_k \alpha_k} && (38)\\
\frac{df(\alpha)}{d\alpha_k^2} &= \sum_i -\frac{n_{ik}}{(n_{ik} - 1 + \alpha_k)^2} + \frac{n_i}{(n_i - 1 + \sum_k \alpha_k)^2} && (39)\\
\frac{df(\alpha)}{d\alpha_k \alpha_j} &= \sum_i \frac{n_i}{(n_i - 1 + \sum_k \alpha_k)^2} && (40)
\end{aligned}$$

A convergent fixed-point iteration is

$$\alpha_k^{new} = \alpha_k \frac{\sum_i \frac{n_{ik}}{n_{ik} - 1 + \alpha_k}}{\sum_i \frac{n_i}{n_i - 1 + \sum_k \alpha_k}} \tag{41}$$

Proof Use the bounds

$$\log(n+x) \geq q\log x + (1-q)\log n - q\log q - (1-q)\log(1-q) \quad (42)$$

$$q = \frac{\hat{x}}{n+\hat{x}} \quad (43)$$

$$\log(x) \leq ax - 1 + \log\hat{x} \quad (44)$$

$$a = 1/\hat{x} \quad (45)$$

to get

$$f(\alpha) \geq \sum_i n_{ik}q_{ik}\log\alpha_k - n_i a_i \sum_k \alpha_k + \text{(const.)} \quad (46)$$

leading to (41).

The LOO likelihood can be interpreted as the approximation

$$\frac{\Gamma(x+n)}{\Gamma(x)} \approx (x+n-1)^n \quad (47)$$

A better approximation would be

$$\frac{\Gamma(x+n)}{\Gamma(x)} \approx (x+(n-1)/2)^n \quad (48)$$

or better still

$$\frac{\Gamma(x+n)}{\Gamma(x)} = \frac{\Gamma(x+1)\Gamma(x+1+n-1)}{\Gamma(x)\Gamma(x+1)} \approx x(x+n/2)^{n-1} \quad (49)$$

3 Estimating variance from probability vectors

Since the mean and variance of the Dirichlet are roughly decoupled in the ML objective, we can get simplifications and speedups by optimizing them alternately. Reparameterize the distribution with $(s, \mathbf{m})$ where

$$\alpha_k = sm_k \quad (50)$$

$$\sum_k m_k = 1 \quad (51)$$

The likelihood is

$$p(D|s) \propto \left(\frac{\Gamma(s)\exp(s\sum_k m_k \log\bar{p}_k)}{\prod_k \Gamma(sm_k)}\right)^N \quad (52)$$

whose derivatives are

$$\frac{d\log p(D|s)}{ds} = N\Psi(s) - N\sum_k m_k\Psi(sm_k) + N\sum_k m_k\log\bar{p}_k \quad (53)$$

$$\frac{d^2\log p(D|s)}{ds^2} = N\Psi'(s) - N\sum_k m_k^2\Psi'(sm_k) \quad (54)$$

A convergent fixed-point iteration is

$$1/s^{new} = 1/s - \Psi(s) + \sum_k m_k \Psi(s m_k) - \sum_k m_k \log \bar{p}_k \tag{55}$$

Proof Use the bound

$$\begin{aligned} \frac{\Gamma(s)}{\prod_k \Gamma(s m_k)} &\geq \exp(sb + \log(s) + c) &(56)\\ b &= \Psi(\hat{s}) - \sum_k m_k \Psi(\hat{s} m_k) - 1/\hat{s} &(57)\end{aligned}$$

to get

$$\log p(D|s) \geq s \sum_k m_k \log \bar{p}_k + sb + \log(s) + (\text{const.}) \tag{58}$$

from which (55) follows.

This iteration is only first-order convergent because the bound only matches the first derivative of the likelihood. We can derive a second-order method by matching the first two derivatives:

$$\begin{aligned} \frac{\Gamma(s)}{\prod_k \Gamma(s m_k)} &\approx \exp(sb + a\log(s) + c) &(59)\\ a &= -\hat{s}^2(\Psi'(\hat{s}) - \sum_k m_k^2 \Psi'(\hat{s} m_k)) &(60)\\ b &= \Psi(\hat{s}) - \sum_k m_k \Psi(\hat{s} m_k) - a/\hat{s} &(61)\end{aligned}$$

which leads to the update

$$\frac{1}{s^{new}} = \frac{1}{s} + \frac{1}{s^2}\left(\frac{d^2 \log p(D|s)}{ds^2}\right)^{-1}\left(\frac{d \log p(D|s)}{ds}\right) \tag{62}$$

This update resembles Newton-Raphson, but converges faster. See Minka (2000).

For initialization, it is useful to derive a closed-form approximate MLE. Stirling's approximation to Γ gives

$$\begin{aligned} \frac{\Gamma(s)\exp(s\sum_k m_k \log \bar{p}_k)}{\prod_k \Gamma(s m_k)} &\approx \left(\frac{s}{2\pi}\right)^{(k-1)/2} \prod_k m_k^{1/2} \exp(s \sum_k m_k \log \frac{\bar{p}_k}{m_k}) &(63)\\ \hat{s} &\approx \frac{(k-1)/2}{-\sum_k m_k \log \frac{\bar{p}_k}{m_k}} &(64)\end{aligned}$$

4 Estimating variance from counts

The likelihood is

$$p(D|s) \propto \prod_i \left(\frac{\Gamma(s)}{\Gamma(n_i + s)} \prod_k \frac{\Gamma(n_{ik} + s m_k)}{\Gamma(s m_k)} \right) \tag{65}$$

The derivatives are

$$\frac{d \log p(D|s)}{ds} = \sum_i \Psi(s) - \Psi(n_i + s) + \sum_k m_k \Psi(n_{ik} + s m_k) - m_k \Psi(s m_k) \tag{66}$$

$$\frac{d^2 \log p(D|s)}{ds^2} = \sum_i \Psi'(s) - \Psi'(n_i + s) + \sum_k m_k^2 \Psi'(n_{ik} + s m_k) - m_k^2 \Psi'(s m_k) \tag{67}$$

A convergent fixed-point iteration is

$$s^{new} = s \frac{\sum_{ik} m_k \Psi(n_{ik} + s m_k) - m_k \Psi(s m_k)}{\sum_i \Psi(n_i + s) - \Psi(s)} \tag{68}$$

(the proof is similar to (31)). However, it is very slow. We can get a fast second-order method as follows. When s is small, i.e. the gradient is positive, use the approximation

$$\log p(D|s) \approx a \log(s) + cs + k \tag{69}$$

$$a = -s_0^2 f''(s_0) \tag{70}$$

$$c = f'(s_0) - a/s_0 \tag{71}$$

to get the update

$$s^{new} = -a/c = s/(1 + \frac{f'(s)}{s f''(s)}) \tag{72}$$

except when $c \geq 0$, in which case the solution is $s = \infty$. When s is large, i.e. the gradient is negative, use the approximation

$$\log p(D|s) \approx \frac{a}{2v^2} + \frac{c}{v} \tag{73}$$

$$a = s^3(s f''(s) + 2 f'(s)) \tag{74}$$

$$c = -(s^2 f'(s) + a/s) \tag{75}$$

to get the update

$$s^{new} = -a/c = s - \frac{f'(s)}{f''(s) + 3 f'(s)/s} \tag{76}$$

For large s, the value of a tends to be numerically unstable. If $s f''(s) + 2 f'(s)$ is within machine epsilon, then it is better to substitute the limiting value:

$$a \to \sum_i \frac{n_i(n_i - 1)(2n_i - 1)}{6} - \sum_{ik} \frac{n_{ik}(n_{ik} - 1)(2n_{ik} - 1)}{6 m_k^2} \tag{77}$$

A even faster update for large s is possible by using a richer approximation:

$$\log p(D|s) \approx c \log\left(\frac{s}{s+b}\right) + \frac{e}{s+b} \tag{78}$$

$$c = \sum_{ik} \delta(n_{ik} > 0) - \sum_i \delta(n_i > 0) \tag{79}$$

$$
\begin{aligned}
e &= -\frac{s+b}{s}(s(s+b)f'(s) - cb) &(80)\\
b &= \mathrm{RootOf}(a_2 b^2 + a_1 b + a_0) &(81)\\
a_2 &= s^3(sf''(s) + 2f'(s)) &(82)\\
a_1 &= 2s^2(sf''(s) + f'(s)) &(83)\\
a_0 &= s^2 f''(s) + c &(84)
\end{aligned}
$$

The approximation comes from setting c equal to its asymptotic value and then choosing (b, e) to match the first two derivatives of f. The resulting update is

$$
s^{new} = \frac{cb^2}{e - cb} = \left(\frac{1}{s} - \frac{f'(s)(s+b)^2}{cb^2}\right)^{-1} \tag{85}
$$

Note that a_2 is equivalent to a above and should be corrected for stability via the same method.

4.1 Large dimensionality

If m_k is roughly uniform and the dimensionality is large, then $\alpha_k << 1$ and we can use the approximations

$$
\begin{aligned}
\Gamma(n_k + \alpha_k) &\approx \Gamma(n_k) &(86)\\
\Gamma(\alpha_k) &\approx 1/\alpha_k &(87)\\
p(\mathbf{x}|s) &\approx \frac{\Gamma(s)}{\Gamma(n+s)} \prod_{n_k>0} s m_k \Gamma(n_k) &(88)\\
&\propto \frac{\Gamma(s) s^{\hat{K}}}{\Gamma(n+s)} &(89)
\end{aligned}
$$

where $\hat{K}$ is the number of unique observations in $\mathbf{x}$. The approximation does not hold if s is large, which can happen when $\mathbf{m}$ is a good match to the data. But if the dimensionality is large enough, the data will be too sparse for this to happen. The derivatives become

$$
\begin{aligned}
\frac{d \log p(D|s)}{ds} &\approx \sum_i \Psi(s) - \Psi(n_i + s) + \hat{K}_i/s &(90)\\
\frac{d^2 \log p(D|s)}{ds^2} &\approx \sum_i \Psi'(s) - \Psi'(n_i + s) - \hat{K}_i/s^2 &(91)
\end{aligned}
$$

Newton iteration can be used as long as the maximum for s is not on the boundary of $(0, \infty)$. These cases occur when $\hat{K} = 1$ and $\hat{K} = n$.

When the gradient is zero, we have

$$
\hat{K} = s(\Psi(n+s) - \Psi(s)) = E[K|s, n] \tag{92}
$$

A convergent fixed-point iteration is

$$s^{new} = \frac{\sum_i \hat{K}_i}{\sum_i \Psi(n_i + s^{old}) - \Psi(s^{old})} \tag{93}$$

Proof Use the bound

$$\begin{aligned} \frac{\Gamma(s)}{\Gamma(n+s)} &\geq \frac{\Gamma(\hat{s})\exp((\hat{s}-s)b)}{\Gamma(n+\hat{s})} && (94) \\ b &= \Psi(n+\hat{s}) - \Psi(\hat{s}) && (95) \end{aligned}$$

to get

$$p(D|s) \geq -s\sum_i b_i + \sum_i \hat{K}_i \log s + (\text{const.}) \tag{96}$$

leading to (93).

Let $\hat{s}(\hat{K}, n)$ be the resulting ML estimate. Since $\hat{s}(n/2, n) \approx 0.8(n/2 - 1)$, a reasonable initial guess when $N = 1$ is

$$\hat{s} \approx \frac{0.8}{2 - 0.55}\left(\frac{n}{n - \hat{K}} - 0.55\right)(\hat{K} - 1) \tag{97}$$

Instead of maximizing s, we can compute its posterior, which follows from (89). With a uniform prior, posterior expectations can be approximated for large n by the following formulas when $N = 1$:

$$\begin{aligned} E[s] &\approx \begin{cases} \hat{s}(\hat{K}+1, n) & \text{if } \hat{K}/n < 1/2 \\ \hat{s}(\hat{K}+2, n) & \text{otherwise} \end{cases} && (98) \\ E\left[\frac{s}{s+n}\right] &\approx \frac{\hat{s}(\hat{K}+1, n)}{\hat{s}(\hat{K}+1, n) + n} && (99) \end{aligned}$$

Applying the large K approximation to the LOO likelihood gives

$$\begin{aligned} t &= \sum_{ik} \delta(n_{ik} - 1) \qquad \text{(number of singletons)} && (100) \\ f(s) &= t\log s - \sum_i n_i \log(n_i - 1 + s) && (101) \\ \frac{df(s)}{ds} &= \frac{t}{s} - \sum_i \frac{n_i}{n_i - 1 + s} && (102) \end{aligned}$$

For $N = 1$:

$$\begin{aligned} s &= \frac{t(n-1)}{n-t} && (103) \\ \frac{s}{s+n} &= \frac{t(n-1)}{n^2 - t} \approx \frac{t}{n} && (104) \end{aligned}$$

5 Estimating mean from probability vectors

Now suppose we know the variance parameter s and we want to estimate the mean $\mathbf{m}$. The likelihood is

$$p(D|\mathbf{m}) \propto \left(\prod_k \frac{\exp(sm_k \log \bar{p}_k)}{\Gamma(sm_k)}\right)^N \tag{105}$$

Reparameterize to get the gradient:

$$\begin{aligned} m_k &= \frac{z_k}{\sum_k z_k} && (106)\\ \frac{d \log p(D|\mathbf{m})}{dz_k} &= \frac{Ns}{\sum_k z_k}\left(\log \bar{p}_k - \Psi(sm_k) - \sum_k m_k \left(\log \bar{p}_k - \Psi(sm_k)\right)\right) && (107)\end{aligned}$$

The MLE can be computed by iterating the fixed-point equations

$$\begin{aligned} \Psi(\alpha_k) &= \log \bar{p}_k - \sum_k m_k^{old}\left(\log \bar{p}_k - \Psi(sm_k^{old})\right) && (108)\\ m_k^{new} &= \frac{\alpha_k}{\sum_k \alpha_k} && (109)\end{aligned}$$

This update converges very quickly.

6 Estimating mean from counts

The likelihood is

$$p(D|\mathbf{m}) \propto \prod_{ik} \frac{\Gamma(n_{ik} + sm_k)}{\Gamma(sm_k)} \tag{110}$$

The maximum can be computed by iterating the fixed-point equation

$$m_k^{new} \propto m_k \sum_i \left(\Psi(n_{ik} + sm_k) - \Psi(sm_k)\right) \tag{111}$$

(the proof is similar to (31)).

For Newton-Raphson, reparameterize to get

$$\begin{aligned} m_K &= 1 - \sum_{k=1}^{K-1} m_k && (112)\\ g_k = \frac{d \log p(D|\mathbf{m})}{dm_k} &= s \sum_i \Psi(n_{ik} + sm_k) - \Psi(sm_k) - \Psi(n_{iK} + sm_K) + \Psi(sm_K) && (113)\\ \frac{d^2 \log p(D|\mathbf{m})}{dm_k^2} &= s^2 \sum_i \Psi'(n_{ik} + sm_k) - \Psi'(sm_k) + \Psi'(n_{iK} + sm_K) - \Psi'(sm_K) && (114)\\ \frac{d^2 \log p(D|\mathbf{m})}{dm_k m_j} &= s^2 \sum_i \Psi'(n_{iK} + sm_K) - \Psi'(sm_K) && (115)\end{aligned}$$

For large s, the search should be initialized at $m_k \propto \sum_i n_{ik}$, since this is the exact optimum as $s \to \infty$.

References

[1] N. L. Johnson and S. Kotz. *Distributions in statistics: Continuous univariate distributions.* New York, Hougton Mifflin, 1970.

[2] T. P. Minka. Beyond Newton's method.
`http://vismod.www.media.mit.edu/~tpminka/papers/newton.html`

[3] A. Naryanan. Algorithm AS 266: Maximum Likelihood Estimation of the Parameters of the Dirichlet Distribution, *Applied Statistics*, Volume 40, Number 2, pages 365-374, 1991.
`http://www.psc.edu/~burkardt/dirichlet.html`

[4] G. Ronning. Maximum-likelihood estimation of Dirichlet distributions. *Journal of Statistical Computation and Simulation* 32(4), pp215–221, 1989.

A Proof of (8)

Use the bound

$$\Gamma(x) \geq \Gamma(\hat{x}) \exp((x - \hat{x})\Psi(\hat{x})) \tag{116}$$

to get

$$\frac{1}{N} \log p(D|\alpha) \geq (\sum_k \alpha_k)\Psi(\sum_k \alpha_k^{old}) - \sum_k \log \Gamma(\alpha_k) + \sum_k (\alpha_k - 1) \log \bar{p}_k + (\text{const.}) \tag{117}$$

leading to (8).

B Proof of (31)

Use the bound

$$\frac{\Gamma(x)}{\Gamma(n+x)} \geq \frac{\Gamma(\hat{x}) \exp((\hat{x} - x)b)}{\Gamma(n+\hat{x})} \tag{118}$$

$$b = \Psi(n + \hat{x}) - \Psi(\hat{x}) \tag{119}$$

and the bound

$$\frac{\Gamma(n+x)}{\Gamma(x)} \geq cx^a \qquad \text{if } n \geq 1 \tag{120}$$

$$a = (\Psi(n+\hat{x}) - \Psi(\hat{x}))\hat{x} \tag{121}$$

$$c = \frac{\Gamma(n+\hat{x})}{\Gamma(\hat{x})}\hat{x}^{-a} \tag{122}$$

to get

$$\log p(D|\alpha) \geq -(\sum_k \alpha_k - 1)\sum_i b_i + \sum_k a_{ik}\log\alpha_k + (\text{const.}) \tag{123}$$

leading to (31).

C Inverting the Ψ function

This section describes how to compute a high-accuracy solution to

$$\Psi(x) = y \tag{124}$$

for x given y. Given a starting guess for x, Newton's method can be used to find the root of $\Psi(x) - y = 0$. The Newton update is

$$x^{new} = x^{old} - \frac{\Psi(x) - y}{\Psi'(x)} \tag{125}$$

To start the iteration, use the following asymptotic formulas for $\Psi(x)$:

$$\Psi(x) \approx \begin{cases} \log(x - 1/2) & \text{if } x \geq 0.6 \\ -\frac{1}{x} - \gamma & \text{if } x < 0.6 \end{cases} \tag{126}$$

$$\gamma = -\Psi(1) \tag{127}$$

to get

$$\Psi^{-1}(y) \approx \begin{cases} \exp(y) + 1/2 & \text{if } y \geq -2.22 \\ -\frac{1}{y+\gamma} & \text{if } y < -2.22 \end{cases} \tag{128}$$

With this initialization, five Newton iterations are sufficient to reach fourteen digits of precision.

D EM for estimation from counts

Any algorithm for estimation from probability vectors can be turned into an algorithm for estimation from counts, by treating the $\mathbf{p}_i$ as hidden variables in EM. The E-step computes a posterior distribution over $\mathbf{p}_i$:

$$q(\mathbf{p}_i) \sim \mathcal{D}(n_{ik} + \alpha_k) \tag{129}$$

and the M-step maximizes

$$E[\sum_i \log p(\mathbf{p}_i|\alpha)] = N \log \Gamma(\sum_k \alpha_k) - N \sum_k \log \Gamma(\alpha_k) + N \sum_k (\alpha_k - 1) \log \bar{p}_k \tag{130}$$

$$\text{where } \log \bar{p}_k = \frac{1}{N} \sum_i E[\log p_{ik}] \tag{131}$$

$$= \frac{1}{N} \sum_i \Psi(n_{ik} + \alpha_k^{old}) - \Psi(n_i + \sum_k \alpha_k^{old}) \tag{132}$$

This is the same optimization problem as in section 1, with a new definition for $\bar{p}$. It is not necessary or desirable to reach the exact maximum in the M-step; a single Newton step will do. The Newton step will end up using the old Hessian (9) but the new gradient (30). Compared to the exact Newton algorithm, this uses half as much computation per iteration.