ST758, Homework 4

Due Nov 12, 2013

When multivariate count data exhibit over-dispersion, the Dirichlet-multinomial distribution is preferred to the multinomial distribution. In the Dirichlet-multinomial model, the multinomial probabilities $\boldsymbol{p} = (p_1, \ldots, p_d)$ follow a Dirichlet distribution with parameter vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$, $\alpha_j > 0$, and density

$$\pi(\boldsymbol{p}) = \frac{\Gamma(|\boldsymbol{\alpha}|)}{\prod_{j=1}^d \Gamma(\alpha_j)} \prod_{j=1}^d p_j^{\alpha_j - 1},$$

where $|\boldsymbol{\alpha}| = \sum_{j=1}^d \alpha_j$.

1. For a multivariate count vector $\boldsymbol{x} = (x_1, \ldots, x_d)$ with batch size $|\boldsymbol{x}| = \sum_{j=1}^d x_j$, show that the probability mass function for Dirichlet-multinomial distribution is

$$\begin{aligned} f(\boldsymbol{x} \mid \boldsymbol{\alpha}) &= \int_{\Delta_d} \binom{|\boldsymbol{x}|}{\boldsymbol{x}} \prod_{j=1}^d p_j^{x_j} \pi(\boldsymbol{p}) \, d\boldsymbol{p} \\ &= \binom{|\boldsymbol{x}|}{\boldsymbol{x}} \frac{\prod_{j=1}^d \Gamma(\alpha_j + x_j)}{\Gamma(|\boldsymbol{\alpha}| + |\boldsymbol{x}|)} \frac{\Gamma(|\boldsymbol{\alpha}|)}{\prod_{j=1}^d \Gamma(\alpha_j)} \\ &= \binom{|\boldsymbol{x}|}{\boldsymbol{x}} \frac{\prod_{j=1}^d (\alpha_j)_{x_j}}{(|\boldsymbol{\alpha}|)_{|\boldsymbol{x}|}}, \end{aligned}$$

   where Δ_d is the unit simplex in d dimensions, $|\boldsymbol{\alpha}|$ equals $\sum_{j=1}^d \alpha_j$, and $(a)_k = \prod_{i=0}^{k-1}(a+i)$ denotes a rising factorial. (Hint: $\Gamma(a+k)/\Gamma(a) = (a)_k$.)

2. Given independent data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, show that the log-likelihood is

$$\begin{aligned} L(\boldsymbol{\alpha}) &= \sum_{i=1}^n \ln \binom{|\boldsymbol{x}_i|}{\boldsymbol{x}_i} + \sum_{i=1}^n \sum_{j=1}^d \sum_{k=0}^{x_{ij}-1} \ln(\alpha_j + k) - \sum_{i=1}^n \sum_{k=0}^{|\boldsymbol{x}_i|-1} \ln(|\boldsymbol{\alpha}| + k) \\ &= \sum_{i=1}^n \ln \binom{|\boldsymbol{x}_i|}{\boldsymbol{x}_i} + \sum_{i=1}^n \sum_{j=1}^d [\ln \Gamma(\alpha_j + x_{ij}) - \ln \Gamma(\alpha_j)] - \sum_{i=1}^n [\ln \Gamma(|\boldsymbol{\alpha}| + |\boldsymbol{x}_i|) - \ln \Gamma(|\boldsymbol{\alpha}|)]. \end{aligned}$$

   Is the log-likelihood a concave function?

3. Write an `R` function to compute the density and/or log-density of the Dirichlet-multinomial distribution. The interface should be `ddirmult(x, alpha, log=FALSE)`. Please vectorize your code. The input `x` and `alpha` are allowed to be n-by-d matrices. In this case, the function should return a vector of Dirichlet-multinomial probabilities $f(\boldsymbol{x}_i|\boldsymbol{\alpha}_i)$ (if `log=FALSE`) or log-probabilities $\ln f(\boldsymbol{x}_i|\boldsymbol{\alpha}_i)$ (if `log=TRUE`) for $i = 1, \ldots, n$.

4. Read in the 'allele.dat', the counts of eight alleles of the `HUMTH01` locus on chromosome 11 from four separate Houston subpopulations of whites, blacks, Chicanos, and Asians. The first column is the allele names (numbers of tandem repeat units) and the remaining four

columns contain the allele counts in the four subpopulations. This comprises a data set of $n = 4$ observed count vectors. Evaluate the log-likelihood of this data at parameter values $\boldsymbol{\alpha} = (1, 1, 1, 1, 1, 1, 1, 1)$ and $\boldsymbol{\alpha}' = (0.11, 4.63, 7.33, 2.97, 5.32, 5.26, .27, .10)$ using your function `ddirmult()`.

5. Derive the score function $\nabla L(\boldsymbol{\alpha})$, observed information matrix $-d^2 L(\boldsymbol{\alpha})$, and expected Fisher information matrix $\mathbf{E}[-d^2 L(\boldsymbol{\alpha})]$ for the Dirichlet-multinomial distribution.

6. Comment on why Fisher scoring method is inefficient for computing MLE in this example.

7. What structure does the observed information matrix possess that can facilitate the evaluation of the Newton direction? Is the observed information matrix always positive definite? What remedy can we take if it fails to be positive definite?

8. Discuss how to choose a good starting point. Implement this as the default starting value in your function below. (Hint: Method of moment estimator may furnish a good starting point.)

9. Write a function for finding MLE of Dirichlet-multinomial distribution given iid observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, using safeguarded Newton's method. The interface should be `dirmultfit(x, alpha0=NULL, maxiters=100, tolfun=1e-6)`. The arguments are: `x` a n-by-d matrix of counts, `alpha0` a d vector of starting point (optional), `maxiters` the maximum allowable Newton iterations (default 100), `tolfun` the tolerance for relative change in objective values (default 1e-6). The return value should be a list containing: `maximum` the log-likelihood at MLE, `estimate` the MLE, `gradient` the gradient at MLE, `hessian` the Hessian at MLE, `se` a d vector of standard errors, `iterations` the number of iterations performed.

10. Read in the 'allele.dat', the counts of eight alleles of the `HUMTH01` locus on chromosome 11 from four separate Houston subpopulations of whites, blacks, Chicanos, and Asians. This comprises a data set of $n = 4$ observed count vectors. Find the MLE $\hat{\boldsymbol{\alpha}}_{\text{MLE}}$ using your function. Use the following starting values: (i) default starting point derived in part 8, (ii) $\boldsymbol{\alpha}^{(0)} = (100, 100, \ldots, 100)$, (iii) $\boldsymbol{\alpha}^{(0)} = (0.01, 0.01, \ldots, 0.01)$.

11. As $\boldsymbol{\alpha}/|\boldsymbol{\alpha}| \to \boldsymbol{p}$, the Dirichlet-multinomial distribution converges to a multinomial with parameter $\boldsymbol{p}$. Therefore multinomial can be considered as a special Dirichlet-multinomial with $|\boldsymbol{\alpha}| = \infty$. Perform a likelihood ratio test (LRT) whether Dirichlet-multinomial offers a significantly better fit than multinomial for this data set.