

## MULTI-LIST METHODS USING INCOMPLETE LISTS IN CLOSED POPULATIONS

Jason Sutherland and Carl James Schwarz<sup>1</sup>

### ABSTRACT

Multi-list methods have become a common application of capture-recapture methodology to estimate the size of human populations. Multi-list methods having been successfully applied to estimating prevalence of diabetes, human immunodeficiency virus (HIV) and drug abuse. A key assumption in multi-list methods is that individuals have a unique “tag” that allows them to be matched across all lists. In some cases, this may not be true. For example, a subset of the lists may use health insurance number to cross-match, while another subset of lists may use date of birth, while only a few lists may have both keys. This paper develops multi-list methodology that relaxes the assumption of a single tag common to all lists.

KEY WORDS: Multi-list methods, capture-recapture methodology

### RÉSUMÉ

Les méthodes de listes multiples sont devenues une application commune de la méthodologie de capture-recapture pour l'estimation de la taille de populations humaines. Elles ont été utilisées avec succès pour estimer la fréquence du diabète, du virus d'immunodéficience humain (VIH) et de la consommation de stupéfiants. Une hypothèse importante des méthodes de listes multiples est que les individus doivent avoir une « étiquette » unique, ce qui permet de les identifier dans toutes les listes. Ce n'est pas toujours le cas. Par exemple, un sous-ensemble des listes peut utiliser le numéro d'assurance-santé pour faire l'appariement, tandis qu'un autre sous-ensemble se sert de la date de naissance, et seulement quelques listes ont les deux étiquettes. Cet article développe une méthodologie de listes multiples qui relâche l'hypothèse d'une étiquette unique commune à toutes les listes.

MOTS CLÉS : Méthodes de listes multiples; méthodologie capture-recapture

### 1. INTRODUCTION

Multi-list methods have become a common application of capture-recapture methodology to estimate the size of human populations. Multi-list methods having been successfully applied to estimating prevalence of diabetes, human immunodeficiency virus (HIV) and drug abuse. A key assumption in multi-list methods is that individuals have a unique “tag” that allows them to be matched across all lists. In some cases, this may not be true. For example, a subset of the lists may use health insurance number to cross-match, while another subset of lists may use date of birth, while only a few lists may have both keys. This paper develops multi-list methodology that relaxes the assumption of a single tag common to all lists.

There are other capture-recapture methods that address difficulties matching individuals across lists; these include corrections for tag loss and adjustments for tag mismatches across lists. Tag loss adjustments were discussed by Seber (1982) and Seber and Felton (1981) and more recently tag errors by Schwarz and Stobo (1999). Huakau (2001), Lee (2002) and Lee et al. (2001) have advanced methods in the area of tag mismatches. Our methodology assumes that there is no tag loss and that no errors occur in matching, but does not assume that all tags are available on each list.

Similar to many capture-recapture methods reviewed by Schwarz and Seber (1999), our approach assumes that the population is closed. We also assume that each tag is sufficient to identify an individual. Under these assumptions, the estimates are found using estimating functions. An example illustrates the application of estimating functions to

---

<sup>1</sup> Jason Sutherland (sutherli@stat.sfu.ca) and Carl James Schwarz, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada, V5A 1S6

estimating the prevalence of diabetes. The results of the model fitting are compared against other analyses of the same data. The last section is a simulation that investigates the performance of the methodology.

## 2. NOTATION

### 2.1 Parameters

Let  $N =$  Population size;  
 $\beta_0 =$  Intercept;  
 $\beta_k =$  Effect of List  $k$ ,  $k = 1, \dots, K$ ;  
 $\beta_{jk} =$  Interaction effect between List  $j$  and List  $k$ .

The notation above is extended to higher order interactions in a similar fashion. The vector of parameters is written  $\beta$ .

### 2.2 Statistics

To help illustrate the definition and construction of the statistics, it will be helpful to consider the small example of four lists in Table 1. Although list 3 has both Tag A and Tag B, lists 1 and 2 have Tag A only, and list 4 has Tag B only.

**Table 1 - Four list, two tag example.**

<u>List</u>	<u>Tags Available for Matching</u>	
List 1	Tag A	
List 2	Tag A	
List 3	Tag A	Tag B
List 4		Tag B

Given the structure of available tags in Table 1, a population member could be matched across all lists; matching Tag A on lists 1, 2 and 3, and matching Tag B on lists 3 and 4. It is not possible to match individuals on lists 1 and 4 who are not also present on list 3.

Multi-list methods commonly begin with the  $2^K - 1$  contingency table containing the observed counts of individuals on various combinations of lists. Let the vector  $\omega = \{\omega_1, \dots, \omega_K\}$  represent the ‘‘capture-history’’ of each individual. We define  $\omega_k$  as

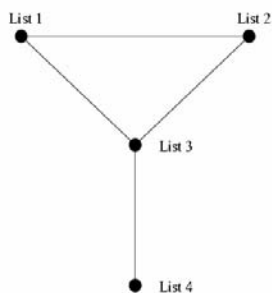
$$\omega_k = \begin{cases} 1 & \text{if individual } i \text{ is known to be present on List } k; \\ 0 & \text{if individual } i \text{ is known not to be present on List } k; \\ \cdot & \text{Unknown whether individual } i \text{ is present/absent on List } k. \end{cases}$$

The count of individuals with the same capture history  $\omega$  is the statistic  $Y_\omega$ . For example,  $Y_{1111}$  is the count of individuals present on lists 1, 2, 3 and 4.

However, not all history vectors can be observed. In Table 1, the statistics  $Y_{1101}$  and  $Y_{1100}$  are not observable because records on lists 1 and 2 cannot be matched with list 4 without list 3 acting as a ‘‘link’’. Let the vector  $\mathbf{Y}$  of length  $L$  represent the observable statistics.

The set of observable statistics is sometimes difficult to determine. The process is facilitated by drawing a graph of the list and tag structure. Vertices (lists) are connected by an edge if they share a tag. An ‘internal’ vertex is one that is connected to more than one vertex, while a ‘leaf’ is a vertex that is connected to a single vertex. The graph of Table 1 is illustrated in Figure 1. There are 12 observable statistics in our example comprising  $\mathbf{Y}$ . They are  $\omega \in \{\{100\cdot\}, \{010\cdot\}, \{110\cdot\}, \{1010\}, \{1011\}, \{0110\}, \{0111\}, \{0010\}, \{0011\}, \{1110\}, \{1111\}, \{\cdot\cdot 01\}\}$ .

**Figure 1 - Graphical representation of Table 1. Four list, two tag example.**



### 3. MODEL DEVELOPMENT

In a closed, multi-list setting, population size is often estimated using the well developed log-linear modelling framework (Feinberg, 1972 and Cormack, 1989). In a two-list setting, the simple Petersen estimator is used.

In our example in Table 1, we would be able to apply a log-linear model only to counts derived from capture histories based on matches across lists 1, 2 and 3 using Tag A. However, the log-linear model ignores information provided by unmatched records on List 4 with Tag B, statistic  $Y_{\cdot 01}$ .

Alternatively, the Petersen estimate could be formed based on counts from lists 3 and 4 using Tag B. In this example,  $n_{List\ 3} = Y_{1111} + Y_{1110} + Y_{0111} + Y_{1011} + Y_{1010} + Y_{0110} + Y_{0011} + Y_{0010}$ , while  $n_{List\ 4} = Y_{1111} + Y_{0111} + Y_{0110} + Y_{0011} + Y_{\cdot 01}$  and  $m_{Lists\ 3\ and\ 4} = Y_{1111} + Y_{0111} + Y_{1011} + Y_{0011}$ , but this ignores information provided by unmatched records on List 1, 2 and 3, statistics  $Y_{100\cdot}$ ,  $Y_{010\cdot}$  and  $Y_{110\cdot}$ .

Our model starts with a vector  $\mathbf{Z}$  of the underlying ‘complete’ counts (of length  $2^K - 1$ ), representing the (unobservable) counts derived as though all ‘tags’ were available on each list. The same notation is used to denote capture histories for statistics  $Z_{\omega}$ . As is the case with log-linear models, assume that

$$\begin{aligned} Z &\sim \text{Poisson}(\mu_Z) \\ \log(\mu_Z) &= \mathbf{X} \beta, \end{aligned}$$

where the design matrix  $\mathbf{X}$  corresponds to the full  $2^K - 1$  cells counts,  $\mathbf{Z}$ , and the vector of parameters,  $\beta$ , represent list and list interaction effects.

The observable statistics  $\mathbf{Y}$  can be related to the vector  $\mathbf{Z}$  through a matrix  $\mathbf{T}$ ,

$$\mathbf{Y} = \mathbf{T}\mathbf{Z},$$

where  $\mathbf{T}$  is defined as an  $L \times (2^K - 1)$  known matrix of indicator variables representing linear combinations of elements of  $\mathbf{Z}$ . A statistic  $Y_{\omega}$ ,  $\omega_k = \cdot$  for any  $k$ , is written as the sum of some  $Z_{\omega}$ ,  $Z_{\omega}$  determined by replacing the unobservable component(s) of  $\omega$  with 0 and 1. For example,  $Y_{110\cdot} = Z_{1101} + Z_{1100}$ .

A likelihood-based approach would be difficult to develop because of the possibility of double-counting individuals on lists where there is incomplete information. For example, the count  $Z_{1101}$  appears in both  $Y_{110\cdot}$  and  $Y_{\cdot 01}$ , as  $Y_{110\cdot} = Z_{1101} + Z_{1100}$  and  $Y_{\cdot 01} = Z_{1101} + Z_{1001} + Z_{0101} + Z_{0001}$ . We propose employing a system of estimating functions (Godambe, 1960) to estimate the population size.

Quasi-likelihood estimates of the vector of parameters are obtained as solutions of

$$\mathbf{D}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{E}[\mathbf{Y}]) = 0,$$

where  $\mathbf{E}[\mathbf{Y}]$  is the vector of first moments derived from  $\mathbf{E}[\mathbf{Y}] = \mathbf{T}\mathbf{E}[\mathbf{Z}] = \mathbf{T}\boldsymbol{\mu}_z$ .  $\mathbf{D}$  is a rectangular matrix of first partial derivatives of  $\mu_{Y_i}$  with respect to the parameter set (i.e.  $\mathbf{D} = \mathbf{T} \frac{\partial \mu_z}{\partial \boldsymbol{\beta}}$ ), while  $\mathbf{V}$  is the working covariance matrix of  $\mathbf{Y}$ . Refer to Wedderburn (1974) and McCullagh (1983) for a treatment of quasi-likelihood theory.

Treating the statistics of  $\mathbf{Y}$  as independent Poisson random variables results in a diagonal working covariance matrix  $\mathbf{V}$ . However, specifying the model as  $\mathbf{Y} = \mathbf{T}\mathbf{Z}$  suggests that the working covariance  $\mathbf{V}$  be written as  $\mathbf{T}\text{diag}(\mu_z)\mathbf{T}'$ . Parameter estimates are known to be consistent estimators even if  $\mathbf{V}$  is not the true covariance matrix of  $\mathbf{Y}$ .

Parameters estimates are obtained using the Newton-Raphson method, written as

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} + [\mathbf{D}\mathbf{V}^{-1}\mathbf{D}]^{-1}\mathbf{D}\mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}^{(i)}))$$

In each step, the working variance is evaluated at the current parameter estimates, or  $\mathbf{T}\text{diag}(\boldsymbol{\mu}(\boldsymbol{\beta}^{(i)}))\mathbf{T}'$ . The covariance of  $\boldsymbol{\beta}$  is consistently estimated by

$$\text{Var}(\boldsymbol{\beta}) = [\mathbf{D}\mathbf{V}^{-1}\mathbf{D}]^{-1}[\mathbf{D}'\mathbf{V}^{-1}V(\mathbf{Y})\mathbf{V}^{-1}\mathbf{D}] [\mathbf{D}\mathbf{V}^{-1}\mathbf{D}]^{-1},$$

where  $\mathbf{D}$  and  $\mathbf{V}$  are previously defined.  $V(\mathbf{Y})$  is based on the “true” distribution of the statistics of  $\mathbf{Y}$ .

There are several approaches to estimating  $\text{Var}(\boldsymbol{\beta})$ . The model-based covariance estimator assumes that  $V(\mathbf{Y}) = \mathbf{V}$ , and  $\text{Var}(\boldsymbol{\beta})$  simplifies to  $[\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}]^{-1}$ . The model-based approach to estimating the covariance will likely mis-specify the form of the variance of  $\mathbf{Y}$ . If the statistics are correlated the covariance estimates will not reflect the true  $\text{Var}(\boldsymbol{\beta})$ . However, its simple application makes it an attractive alternative estimator.

The empirical, or robust covariance estimator, estimates the covariance  $V(\mathbf{Y})$  using cross-products of residuals for each record,  $[\mathbf{Y}_i - \boldsymbol{\mu}(\boldsymbol{\beta})][\mathbf{Y}_i - \boldsymbol{\mu}(\boldsymbol{\beta})]'$ , where  $\mathbf{Y}_i$  is a vector of indicator variables for each observed individual and  $\boldsymbol{\mu}(\boldsymbol{\beta})$  is the expectation. The expression  $\mathbf{D}'\mathbf{V}^{-1}V(\mathbf{Y})\mathbf{V}^{-1}\mathbf{D}$  is then summed over all observed records.

The non-parametric bootstrap standard error (Efron and Tibshirani, 1993) can also be used to estimate  $\text{Var}(\boldsymbol{\beta})$ . To establish a bootstrap sample,  $\mathbf{Y}^*$ , we re-sample records to create a ‘new’ sample of the same size as the original data. For each  $\mathbf{Y}^*$ , we estimate the parameters  $\boldsymbol{\beta}^*$  using the methodology described.

Many models can be fit with the log-linear framework provided that the number of parameters determining  $\boldsymbol{\beta}$  does not exceed  $L$ . For instance, we can model equal capture probabilities for all lists ( $\boldsymbol{\beta} = \{\beta_1 = \beta_2 = \beta_3 = \beta_4\}$ ), corresponding to  $M_0$  (reviewed by Chao, 2001). Also, we can fit equal second order interaction effects,  $\{\beta_{12} = \beta_{13} = \beta_{14} = \beta_{23} = \beta_{24} = \beta_{34}\}$ , the quasi-symmetric model, and compare it to that of  $M_t$ , different capture probabilities for each list,  $\boldsymbol{\beta} = \{\beta_0, \beta_1, \beta_2, \beta_3, \beta_4\}$ .

The Akaike information criterion (AIC) and the deviance ( $G^2$ ) are well established for log-linear model selection. Because we do not employ a likelihood function, we use alternative approaches. A Pearson statistic is asymptotically chi-square, although as Cormack (1989) notes, capture history data is often sparse and applying the statistic may not be appropriate if this is the case. For model selection purposes, we also compute the  $\text{QIC}_u$  statistic (Pan, 2001), a statistic analogous to the AIC for estimating functions that adjusts for the number of parameters in the model.

Once a model is selected and estimates of  $\boldsymbol{\beta}$  obtained, the population size is estimated (as in all log-linear multi-list methods) by estimating the number in the missing cell,  $Z_{000\dots 0}$ , and then adding this to the observed cell counts. In the

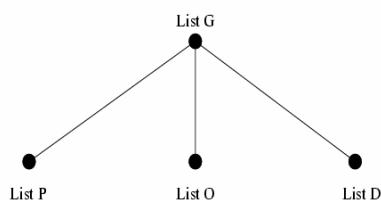
“corner-point” log-linear parameterization (refer to Agresti, 2002, Sections 8.1.3, 8.1.4, and 12.3.6),  $Z_{0000}$  is estimated by  $e^{\beta_0}$ . Then,  $N$  is estimated by the missing cell count cell count plus the observed cell counts.

### 3.1 Example

We used as an example the data available from the Auckland Diabetes Study data. Huakau (2001) describes the characteristics of the lists whose information was to used to estimate the prevalence of diabetes. In this example, four lists were available; 1,276 general practioners records (List G), 1,297 pharmacy records (List P), 12,792 outpatient records (List O) and 3,436 inpatient discharge records (List D). In the analyses of Seber, Huakau and Simmons (2001), Lee (2001 and 2002) and Huakau (2001), five ‘tags’ were split such that Tag A consisted of first name, surname and age, while Tag B consisted of date of birth, gender and address.

To develop an example for our methodology, we simulated a new list structure whereby Lists P, O and D do not share a tag, but each shares one with List G. Figure 2 shows the graph of the list structure of this example.

**Figure 2 - Graphical representation of lists of Auckland Diabetes Data.**



To establish our simulated dataset from the Diabetes Study, we aggregated over appropriate cells of  $\mathbf{Z}$ . For instance,  $Y_{01..} = Z_{0111} + Z_{0110} + Z_{0101} + Z_{0100}$ . The set of observable statistics,  $\mathbf{Y}$ , is available from the authors. For this reduced list structure, we could not use a log-linear model on all 4 lists, nor could we use a log-linear model on any subset of 3 lists. The approximately unbiased Petersen estimates formed using list pairs G-P, G-O and G-D are shown in Table 4. The Petersen estimator provided widely varying point estimates of the population size.

**Table 4 - Results of estimating population size of the Auckland Diabetes Study.**

Estimation Method	$N$	$se(N)$
Approx. Unbiased Petersen		
List G and P	14,412	1,201
List G and O	30,940	1,007
List G and D	27,260	1,935
Estimating Functions, model $\beta^{(6)}$	45,853	
Model Based		4,530
Robust		4,343
Bootstrap		4,008
All Information Available		
$\beta^{(6)}, 2^4 - 1$ log-linear model, both tags	55,168	3,459

We applied the methodology developed above. The results of our model fitting are available from the authors. We selected model  $\beta^{(6)}$  because its  $QIC_u$  statistic is the smallest. At this point, a model averaging approach, such as discussed by Hook and Regal (1997) and Buckland et al. (1997), could be used to account for model uncertainty in the estimate of  $N$ .

The estimated population size was  $N_{\beta^{(6)}} = 45,853$ . This estimate was substantially larger than the pairwise Petersen estimates. We calculated  $V(\mathbf{Y})$  using three methods. The model-based, the robust estimator and the bootstrap estimate of

the standard error were substantially larger than those from the Petersen estimates. The similarity of the model-based and robust estimates of variability suggests that our Poisson assumption may not be inappropriate.

Because all  $2^4 - 1$  cells of the contingency table were available, we fit log-linear models to the counts of  $\mathbf{Z}$  for comparison (Table 4). Using the same model as  $\beta^{(6)}$ , the population estimate based on the complete data was 55,168, which is substantially larger than our estimate.

Direct comparison of the results to the literature is difficult. For instance, the estimating functions approach assumes tags are not missed and errors do not occur in matching. However, in analysis of the same dataset, Lee et al. (2001) estimated 85% tag retention, suggesting that our method will overestimate true population size since we do not adjust for tag loss. The simulated dataset for this example is based on counts from the full  $2^K$  multiway contingency table ( $\mathbf{Y} = \mathbf{TZ}$ ), resulting in a loss of information which manifests as different inferences on population size.

#### 4. DISCUSSION

This paper developed a multi-list method which incorporates lists when not all lists are required to have the same tags. This situation is likely to appear in epidemiological settings when lists being used for matching only have subsets of tags available. An example demonstrates that it is easily applied to real data.

A simulation study showed that it has consistently closer to the “true” population size than log-linear models or the simple Petersen on subsets of lists. However, not surprisingly, the methodology is less precise than if all data were available.

The simulation illustrated a danger with multi-list modelling; missing lists can lead to ‘misleading’ models as a result of interaction effects not being properly detected. If this happens, population estimates can be severely compromised. This is evident with the Petersen estimator, where list interaction effects cannot be modelled, although this is also true for the log-linear estimate and the new estimator.

At first glance, the EM algorithm (Dempster et al, 1977) would appear to be another method that could be suitable for this problem. Using this approach, a likelihood could be developed assuming Poisson counts for  $\mathbf{Z}$ . Then,  $Z_{\omega}$  with missing values would be estimated by conditioning on the observed count  $Y_{\omega}$ . Calculating the expectations of the missing  $Z_{\omega}$  is difficult because the unobserved  $Z_{\omega}$  appear in more than one observed statistic. For example, from Table 1, in the

expectation step,  $Z_{1101}$  appears in  $Y_{110}$  and  $Y_{\cdot 01}$ , and  $Z_{1101} | Y_{110} \sim Bin\left(Y_{110}, \frac{Z_{1101}}{Z_{1101} + Z_{1100}}\right)$  and  $Z_{1101} | Y_{\cdot 01} \sim Bin\left(Y_{\cdot 01}, \frac{Z_{1101}}{Z_{1101} + Z_{1001} + Z_{0101} + Z_{0001}}\right)$ . It is unclear how to implement the E-step under these constraints.

In the development of our model, we assumed that we were able to match records across lists without error. Violation of this assumption may positively bias population estimates. We also assumed that there were no missing tags, so we did not correct for tag loss (Seber, 1982, Seber and Felton, 1981), transcription errors or misread tags (Seber, Huakau and Simmons, 2000), each a potential source of bias.

The closed population assumption is common in multi-list methods. If lists are compiled over long periods, this assumption may not be viable; leading to biased estimates of population size. In this case, methods for open populations should be investigated.

#### ACKNOWLEDGEMENTS

Research supported by the National Science and Engineering Research Council (NSERC) of Canada. The authors would like to thank the Canadian Institute for Health Information (CIHI) and NSERC.

## REFERENCES

- Agresti, A. (2002) *Categorical Data Analysis*, 2nd Edition Wiley: New York.
- Buckland, S.T, Burnham, K.P. and Augustin, N.H. (1997) Model selection: An integral part of inference. *Biometrics* **53**, 603-618.
- Cormack, R.M. (1989) Log-linear models for capture-recapture. *Biometrics* **45**, 395-413.
- Chao, A. (2001) An overview of closed capture-recapture models. *Journal of Agricultural, Biological and Environmental Statistics* **6**, 158-175.
- Chao, A. and Tsay, P.K. (1998) A sample coverage approach to multiple-system estimation with application to census undercount. *Journal of the American Statistician* **93**, 283-293.
- Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1-18.
- Efron, B., Tibshirani, R. (1993) *An Introduction to the Bootstrap*. New York: Chapman and Hall.
- Feinberg, S.E. (1972) The multiple recapture census for closed populations and incomplete  $2^k$  contingency tables. *Biometrika* **59**, 591-603.
- Godambe V.P. (1960) An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics* **31**, 1208-1212
- Hook, E.B., Regal, R.R. (1997) Validity of methods for model selection, weighting for model uncertainty, and small sample adjustments in capture-recapture estimation. *American Journal of Epidemiology* **145**, 1138-1144.
- Huakau, J.T. (2001) PhD dissertation. The University of Auckland. New Zealand.
- Lee, A. J., Seber, G.A.F., Holden, J.K., Huakau, J.T. (2001) Capture-recapture, epidemiology, and list mismatches: several lists. *Biometrics* **57**, 707-713.
- Lee, A. J. (2002) Effects of list errors on the estimation of population size. *Biometrics* **58**, 185-191.
- Pan, W. (2001) Akaike's information criteria in generalized estimating equations. *Biometrics* **57**, 120-125.
- McCullagh, P. (1983) Quasi-likelihood functions. *The Annals of Statistics* **11**, 59-67.
- Schwarz, C.J. and Stobo, W.T. (1999) Estimation and effects of tag-misread rates in capture-recapture studies. *Canadian Journal of Fisheries and Aquatic Sciences* **56**, 551-559
- Schwarz, C.J. and Seber, G.A.F. (1999). Estimating animal abundance: review III. *Statistical Science* **14**, 427-456.
- Seber, G.A.F. (1982) *The Estimation of Animal Abundance*, 2nd edition. London: Edward Arnold.
- Seber, G.A.F., Felton, R. (1981) Tag loss and the Petersen mark-recapture experiment. *Biometrika* **68**, 211-219.
- Seber, G.A.F., Huakau, J.T., and Simmons, D. (2000) Capture-recapture, epidemiology and list mismatches: two lists. *Biometrics* **56**, 1227-1232.
- Wedderburn, R.W.M. (1974) Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439-447.

