Jackknifing

for genetic analysis of pedigree data

David Duffy

Queensland Institute of Medical Research
Brisbane, Australia

The delete-1 jackknife (Quenouille, Tukey)

Jackknife estimate of bias

$$b_J = \frac{n-1}{n}(\sum_{i=1}^{n} T_{n-1,i} - nT_n)$$

Jackknife-corrected estimator

$$T_J = nT_n - (n-1)\sum_{i=1}^{n} T_{n-1,i}$$

Jackknife variance estimator

$$v_J = \frac{n-1}{n}\sum_{i=1}^{n}(T_{n-1,i} - {}^{1}\!/_{n}\sum_{j=1}^{n} T_{n-1,j})^2$$

where T_n is the estimator based on the entire dataset of n values, and $T_{n-1,i}$ the estimator based on the dataset after leaving out the ith observation.
The ith pseudovalue is $nT_n - (n-1)T_{n-1,i}$.

What does one get?

- An estimator with reduced bias (ML VC analysis)
- An “automatic” “nonparametric” estimate of the sampling variance
- Cross-validation type model diagnostics (pseudo-values)
- (Sampling density estimation)

The jackknife for complex data

Jackknife estimation is used in arenas where traditional approaches are too difficult or too expensive.

- Sample Survey data
- Time Series data
- Clustered data (GEE1, GEE2, Ziegler; Yan and Fine)
- More complex dependent data

The grouped jackknife for complex data

It is known that the standard delete-1 jackknife can be inconsistent for these types of data. Deleting larger groups at a time gets around these problems:

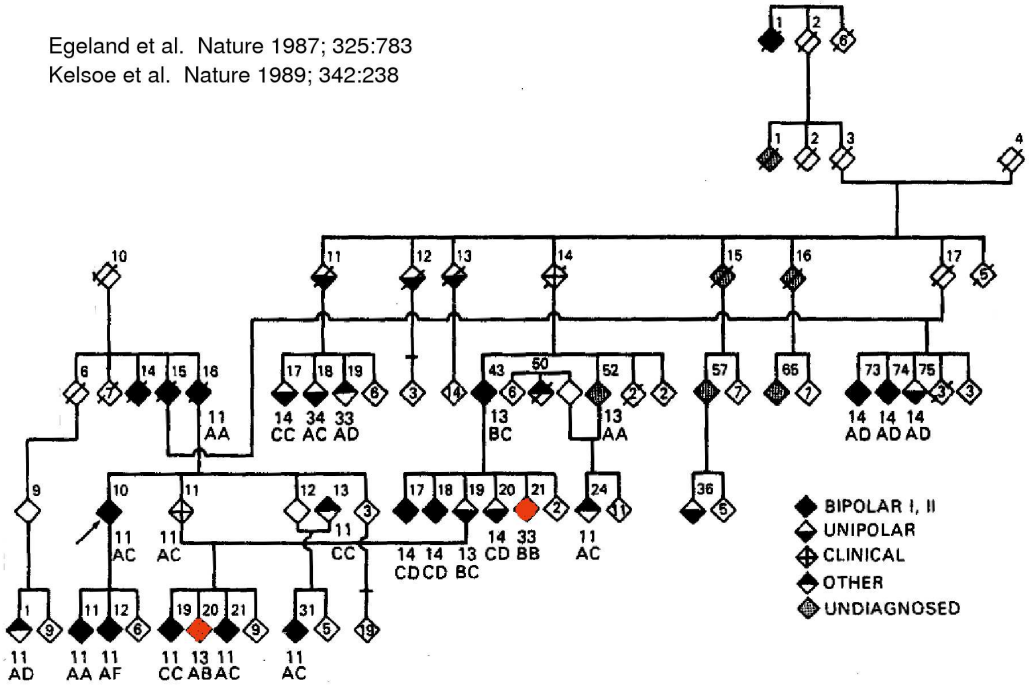
- Grouping can reflect the sampling design
- Grouping can reflect the natural structure of the data
- Groups can be approximately *i.i.d.*
- Also reduce computations (along with one-step jackknife etc)

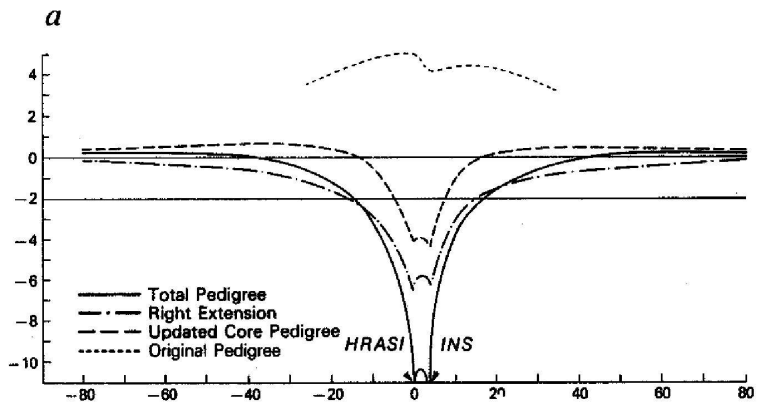
Genetic applications of leave-one-out diagnostics

Evidence for genetic linkage of a marker locus to a phenotype is usually derived via maximum likelihood methods, and the strength of evidence expressed as the decimal log likelihood ratio (following Barnard and Morton).
Several examples are known of marked sensitivity of linkage results to alteration of only one or two datapoints (1-2% of N) eg Egeland *et al* (Nature 1987; 325:783). That is, T is not always very smooth,
Therefore, it is a standard approach to carry out delete-1 type calculations to pinpoint influential individuals.

Genetic applications of delete-1 diagnostics II

Egeland et al. Nature 1987; 325:783
Kelsoe et al. Nature 1989; 342:238





QIMR

Variance components linkage analysis

A current approach for quantitative trait locus (QTL)

$$y = a + d + q + e, \qquad a \sim N(0, R\sigma_A^2)$$

$$d \sim N(0, K\sigma_D^2)$$

$$q \sim N(0, \hat{\Pi}\sigma_Q^2)$$

$$e \sim N(0, I\sigma_E^2)$$

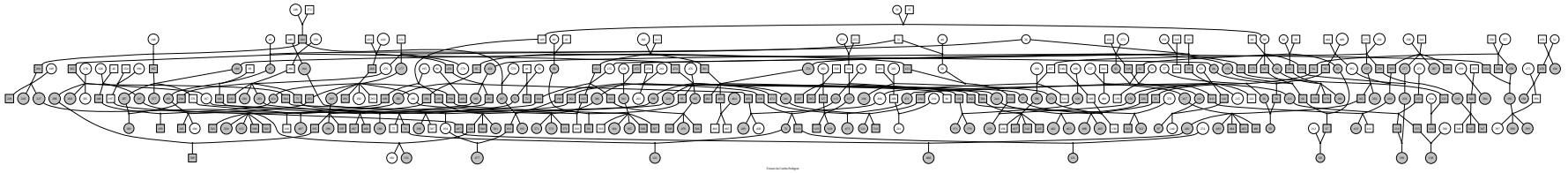
where R is the numerator relationship matrix for the pedigree,
K contains the coefficients of fraternity (Δ_7), and
$\hat{\Pi}$ is the matrix of average identity-by-descent at the map location of interest.

$$LL = -0.5[\log(\det(S)) + (y - \mu)^T S^{-1}(y - \mu)]$$

Genetic applications of grouped jackknife

In experimental crosses or studies of nuclear families, there is a natural grouping for deletion.
In comparisons with other methods, the jackknife performs well.
Arvesen and Schmitz (1970), Knapp et al (1989) applied to variance components analysis (nested ANOVA).
QTL Cartographer (Basten, Weir and Zeng 1994, 2002).
Roff and Preziosi (1994,2002)
Szyda et al (2001) for variance components analysis using half-sib families (breed crossing).
In GEE genetic applications of Ziegler er al (2000) and Yan and Fine (2004), the jackknife variance estimator approaches the sandwich estimator.

What to do with big pedigrees?



This is a pedigree I am currently analysing. There are 216 individuals with serum Immunoglobulin E measured.

It is not clear which are the optimal groups or group sizes for jackknifing.

However, it is possible that the variance estimator may be less affected than the bias estimator by a poor choice.

A possibly useful view of VC analysis is as a nonlinear regression (or EE problem), stacking S as a vector. In this case, removing one individual removes a cluster of n variances and covariances. Bootstrapping has been applied here.

A little simulation

Model (gives familial correlations similar to those for IgE level):

- Single QTL with two equifrequent alleles
- No dominance (genotype means 4,5,6), so $V_Q = 0.5$
- $V_E = 1$
- No residual familial correlations , so $V_A = 0$

Procedure:

- Gene drop (simulate) genotypes within Tristan pedigree
- Generate IgE values for 216 phenotyped individuals
- Generate marker completely linked to QTL
- Perform VC linkage analysis (AS319) on complete data
- And on jackknife samples from pedigree

Delete-d jackknife with random subsampling

The usual jackknife variance estimator for grouped data with randomly defined groups is:

$$v_{JG} = \frac{m-1}{m}\sum_{i=1}^{m}(T_{n-g,i} - \tfrac{1}{m}\sum_{j=1}^{m}T_{n-g,j})^2$$

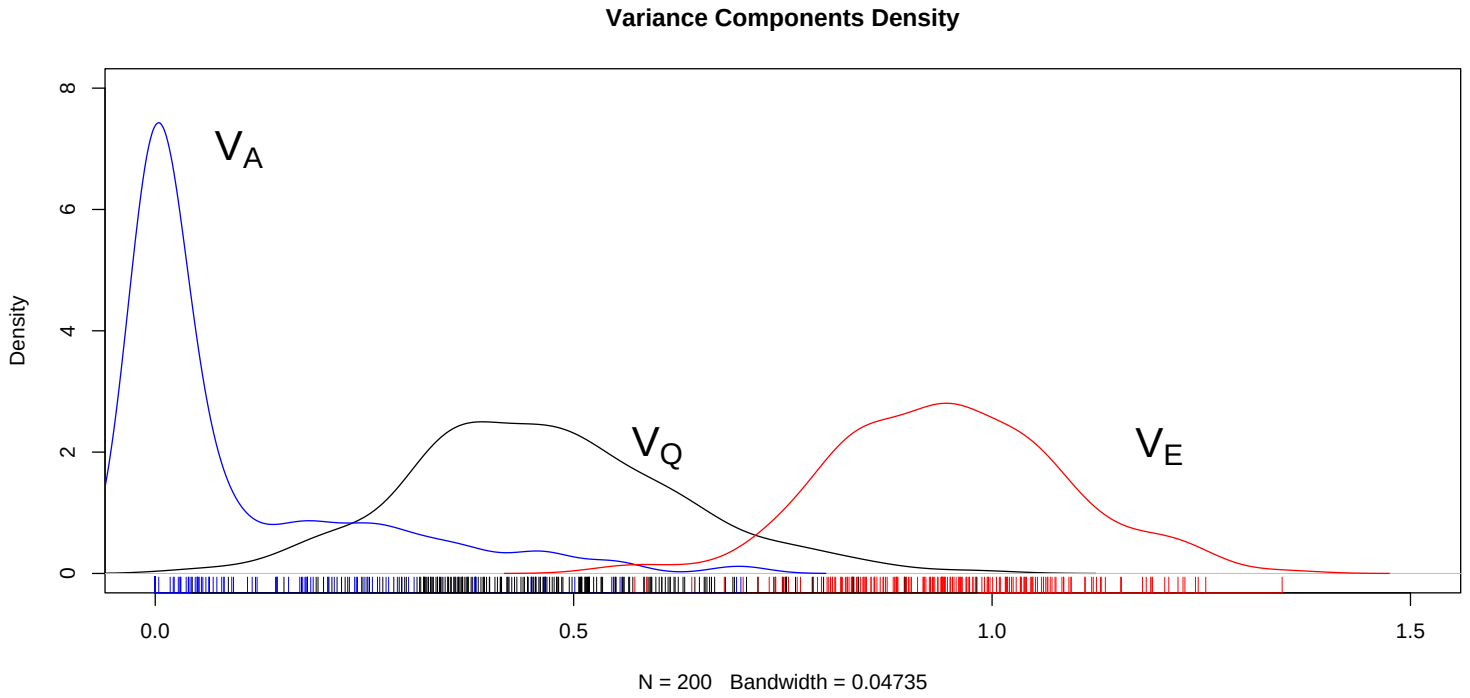
where the n data points have been divided into m groups of size g.

The delete-d jackknife variance estimator with random subsamples is slightly different:

$$v_{JD} = \frac{n-d}{dm}\sum_{i=1}^{m}(T_{n-d,i} - \tfrac{1}{m}\sum_{j=1}^{m}T_{n-d,j})^2$$

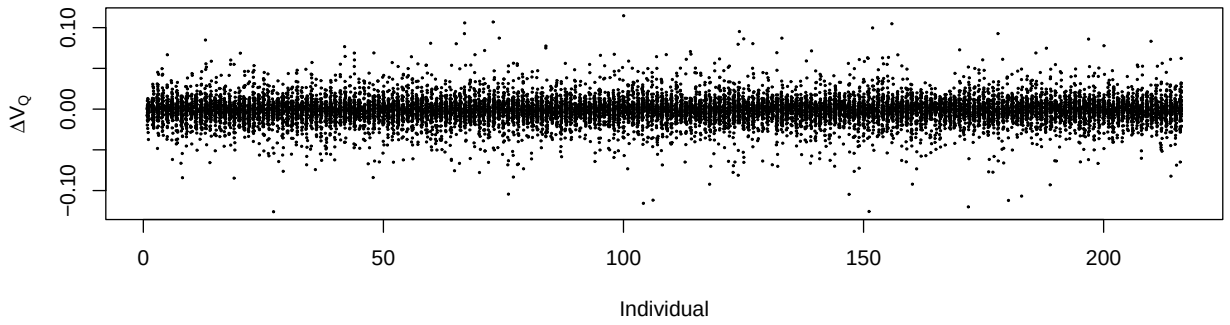
where m draws of size d from the n data have been made.

Results



QIMR

Results



VC	MLE	Emp SD	JSE	Delete-10 JSE	Delete-1 bias
A	0.079	0.130	0.238	0.204	+0.216
Q	0.478	0.151	0.210	0.199	-0.268
E	0.962	0.140	0.194	0.181	0.017

QIMR

Conclusions

- The delete-1 bias correction tends to overcompensate
- The delete-1 variance estimators are larger than the empirical variances
- The delete-10 variance estimators are a slight improvement
- For this particular pedigree structure, the influence of individuals reflects their trait value, rather than their position in the pedigree

Software used: http://www.qimr.edu.au/davidD