A modification to the jackknife to deal with adjacent blocks

Nick Patterson

February 21, 2020

1 Introduction

When we carry out a block jackknife [2, 1], which has become a standard tool in the Reich Lab, we want to compute a standard error for the estimates of some quantity. Examples where we use this are in the f-statistics of [4] and in the computation of a standard error for F_{st} in *smartpca* ([3]).. The standard weighted jackknife computes *pseudovalues* by deleting each block of data in turn, and then computing an estimate from the remaining data. An underlying assumption is that the data deleted from different blocks are independent, and so the deviations of the pseudovalues from the overall estimator are independent. In genetics, this will rarely be exactly the case as LD will extend across block boundaries, but more seriously with our new technique *rolloff* we are computing statistics involving (many) pairs of snps, and so if we delete all data in a block we are also deleting data from a neighboring block if a pair of SNPs crosses a block boundary. Thus we can expect that the 'errors' in our pseudovalues will correlate for neighboring blocks. In this short note, we sketch a remedy. Our idea, for which we give more detail below, is to convolve our pseudovalues with an appropriate vector such that the resulting quantities are uncorrelated and then apply the jackknife to the resulting convolved data.

We suppose that we are carrying out a block jackknife and for block i we have a jackknifed estimate $\theta(i)$ of a parameter m from deleting block i. We also have a weight $w(i)$ for block i, most commonly proportional to the number of observations in the block. We number the blocks $0, 1, \ldots, N-1$.

Estimate m by

$$\tilde{m} = \frac{\sum_i w(i)\theta(i)}{\sum_i w(i)}$$

and form the residual $j(i) = \theta(i) - \tilde{m}$.

We will refer to the $j(i)$ as the *pseudoerrors*. In the case of interest, $j(i)$ is correlated at lag 1. It is convenient for the theory below, to regard the j as

situated on a circle, so that block $N-1$ is a neighbor of block 0, and ignore chromosomal end effects. Let r be an estimate of the lag-1 correlation.

$$r = \frac{\sum_i j(i)j(i+1)}{\sum_i j^2(i)}$$

Here and later we interpret block indices mod N. Suppose that r is significantly different from 0. We shall also suppose, as is realistic in practice, that the correlation of j at lag greater than 1 is 0. Thus the correlation matrix V of j is a circulant, whose first row is $1, r, 0, 0, \ldots, 0, r$. Our idea is to apply a circulant operator L to the pseudoerrors, such that the resulting quantities are uncorrelated. We therefore seek a symmetric circulant L such that

$$LVL = I$$

Fourier analysis and some standard signal theory (which I had to remind myself about!) helps us here. The Fourier transform diagonalizes V, and we can calculate that the eigenvalues of V are

$$l_k = 1 + 2r\cos\left(\frac{2\pi k}{N}\right) \tag{1}$$

L and V have the same eigenvectors and so the corresponding eigenvalue d_k for L is:

$$d_k = 1/\sqrt{l_k}$$

[This is related to 'spectral factorization' in Signal Theory]. Here we assume, as again is reasonable in practice, that $r < 0.5$, which implies that $l_k > 0$ so that V is positive definite. Our theory works in principle for arbitrary circulant positive definite V, but we know of no genetic applications of the block jackknife, where the correlation is non-zero for larger lag than 1.

Taking the inverse Fourier transform we recover L. We find that the first row of L is

$$c_0, c_1, \ldots, c_{N-1}$$

where c_i is given by

$$c(i) = 1/N \sum_k d_k \cos\left(\frac{2\pi k i}{N}\right)$$

Remaining rows of L are obtained by circularly shifting c. Let $S = \sum_i c(i)$. We can show that $S = 1/\sqrt{(1+2r)}$. Now apply L to our original data, defining

$$\begin{aligned} j'(i) &= \sum_k L(i,k)j(k) \\ &= \sum_k c(k+i)j(k) \\ &= \sum_k c(k)j(k-i) \end{aligned} \tag{2}$$

[If efficiency is an issue, we can carry out this convolution using Fourier transforms, so that the total work here is $O(N \log N)$.] The errors of the j' are uncorrelated, so we can apply the weighted jackknife to j' to estimate a mean M' and standard error σ'. However in effect we are producing estimates here for Sm not our original m and thus our final estimate is that the mean is M'/S with standard error σ'/S.

Although *rolloff* was the initial motivation here, this theory is likely to yield a slight improvement in standard error estimates in other cases, such as the F_{st} computations of *smartpca* where LD across block boundaries introduces some unwanted correlations, though we believe that in most of these cases the effects are small.

2 A specification

*void weightjack(double *est, double *sig, double mean, double *jmean, double *jwt, int g) ;*

Is an old routine in nicklib. We propose
*void weightjackr(double *est, double *sig, double mean, double *jmean, double *jwt, int g, double *prho)*
Here *mean, jmean, jwt, g* are inputs, g is the number of blocks, jwt are weights, and $jmean$ are the jackknifed estimates deleting each block.

Step 1: Estimate rho (the correlation)
Set

$$m' = \frac{\sum_{i=0}^{g-1} jmean[i]}{g}$$

and $j^\star[k] = jmean[k] - m'$.
$j^\star$ of course has mean 0. Now set

$$*prho = \rho = \frac{\sum_{i=0}^{g-1} j^\star[i] j^\star[i+1]}{\sum_{i=0}^{g-1} j^{\star 2}[i]}$$

where we read indices mod g. We require that $\rho < \frac{1}{2}$ and in practice if $\rho < 0$ we should set $\rho = 0$ and just apply *weightjack*
Step 2: Apply our theory
Set

$$l_k = 1 + 2\rho \cos\left(\frac{2\pi k}{g}\right) \tag{3}$$

and

$$d_k = 1/\sqrt{l_k}$$

for $k = 0, 1, \ldots g-1$. Then calculate:

$$c(i) = 1/g \sum_k d_k \cos\left(\frac{2\pi k i}{g}\right)$$

As a check, set

$$S = \sum_i c(i)$$

S should equal $1/\sqrt{(1+2\rho)}$. Now we apply equation (2) Set

$$j'(i) = \frac{\sum_k c(k) j(k-i)}{S}$$

where again we interpret indices mod g. The division by S leaves the mean of j' the same as the mean of j. Now we apply the weighted jackknife to j' to obtain the standard error.

What should the weights $w'(i)$ be for the jackknife? This is not entirely clear. For ρ small, simply setting $w'(i) = w(i)$ should not work badly. A more refined formula is obtained by thinking of $w(i)$ as an estimated variance of $j(i)$. Then we can estimate the covariance of $j(i), j(i+1)$ as

$$q(i) = \rho(w(i)w(i+1))^{\frac{1}{2}}$$

Now set

$$\begin{aligned} w'(i) &= S^2 Var(j'(i)) \\ &= Var(\sum_k c(k) j(k-i)) \\ &= \sum_k c^2(k) w(k-i) + 2\sum_k c(k)c(k+1)q(k-i) \end{aligned}$$

which will be non-negative.

Acknowledgement: Pier Palamara, Srefan Groha and Arti Tandon have worked on this with me.

References

[1] F.M.T.A. Busing, E. Meijer, and R. van der Leeden. Delete-m jackknife for unequal m. *Statistics and Computing*, 9:3–8, 1999.

[2] H R Künsch. The jackknife and the bootstrap for general stationary observations. *Ann. Statist.*, 17:1217–1241, 1989.

[3] A. Price, N. Patterson, R. Plenge, M. Weinblatt, N. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38:904–909, 2006.

[4] D. Reich, K. Thangaraj, N. Patterson, A. L. Price, and L. Singh. Reconstructing Indian population history. *Nature*, 461:489–494, Sep 2009.