# FDVAR: R code for variance of the number of false discoveries
# Beta version

Art B. Owen

May 2004

**Abstract**

This report documents a beta version of code for computing the variance of the false discovery rate, when tests are based on correlations. It provides some orientation bridging the gap between the research article and the R material itself.

## Context

The R functions described here are designed to implement the computations described in Owen (2004).

Let $\mathcal{Y}$ be an $n$ by $d$ matrix of numbers, and let $\mathcal{X}$ be an $n$-vector of numbers. Suppose that the $n$ vectors $(X_i, Y_{i1}, \ldots, Y_{id})$ of length $d+1$ are independent and identically distributed. Suppose also that $\mathcal{X}$ is independent of $\mathcal{Y}$. For $j = 1, \ldots, d$ let $\hat{\rho}_j$ be the sample correlation between $\mathcal{X}$ and the $j$'th column of $\mathcal{Y}$. The corresponding unknown true correlation is $\rho_j$.

The functions accept missing values in $\mathcal{Y}$. Sample correlations are computed by leaving out pairs with one or more missing value. When missing values comprise a significant proportion of the elements of $\mathcal{Y}$, the result will be very conservative, as described below. In extreme cases some pairs of columns of $\mathcal{Y}$ may have 2 or fewer elements in common even when both have many elements in common with $\mathcal{X}$. The slow but direct simulation alternative described below is better able to handle such cases. It may be even better to impute the missing values from other values of $\mathcal{Y}$. Such imputation maintains independence of $\mathcal{X}$ and $\mathcal{Y}$.

For $j = 1, \ldots, d$ let $N_j$ represent the null hypothesis that $\rho_j = 0$. We consider left, right and central alternatives to $N_j$ denoted $L_j$, $R_j$, and $C_j$, under which $\rho_j < 0$, $\rho_j > 0$ and $|\rho_j| > 0$ respectively. We reject $N_j$ in favor of $R_j$ when

$$\hat{\rho}_j > r_n^{1-\alpha} = \mathrm{qBeta}\left(1-2\alpha, \frac{1}{2}, \frac{n-2}{2}\right)^{1/2}, \tag{1}$$

where qBeta denotes the quantile function of the Beta distribution. We reject $N_j$ in favor of $L_j$ when $\hat{\rho}_j < -r_n^{1-\alpha}$ and we reject $N_j$ in favor of $C_j$ if we reject either $L_j$ or $R_j$ at the level $\alpha/2$.

If $N_1, \ldots, N_d$ all hold, then the expected number of rejected hypotheses is $d\alpha$. The variance of the number of rejected hypotheses for alternative $A \in \{L, R, C\}$ is

$$V_{n,\alpha,A} = d\alpha(1-\alpha) + \sum_{j=1}^{d} \sum_{\substack{j'=1 \\ j' \neq j}}^{d} \left( C_{n,\alpha}^{A}(\hat{\rho}_{jj'}) - \alpha^2 \right) \tag{2}$$

where $C^A$ is given by

$$C_{n,\alpha}^{L}(\rho) = \begin{cases} 0 & 2(r_n^{1-\alpha})^2 \geq \rho + 1, \\ \frac{n-3}{4\pi} \int_{\frac{2r^2}{\rho+1}}^{1} (1-\omega)^{(n-5)/2} \Big( 2\text{acos}\Big( \frac{r}{\sqrt{\omega}} \Big) - \text{acos}(\rho) \Big) d\omega & else, \end{cases} \tag{3}$$

and $C_{n,\alpha}^{R}(\rho) = C_{n,\alpha}^{L}(\rho)$ and $C_{n,\alpha}^{C}(\rho) = 2(C_{n,\alpha}^{L}(\rho/2) + C_{n,\alpha}^{L}(-\rho/2))$.

Equation 2 uses all $d(d-1)/2$ pairs of off-diagonal correlations $\hat{\rho}_{jj'}$. In many applications $d$ is very large and $O(nd^2)$ work to compute the correlations is too much. In such cases a sample of off-diagonal correlations is required. One approach is to select two random non-overlapping sets of $d_1$ and $d_2$ columns of $\mathcal{Y}$ respectively and compute the $d_1 \times d_2$ correlations among them. Another approach is to sample $M$ off-diagonal correlations by simple random sampling (with or without replacement) from the $d(d-1)/2$ off-diagonal correlations among columns of $\mathcal{Y}$. The latter sample has easier-to-study sampling properties. The former runs in much less time and memory and allows much larger numbers of off-diagonal correlations to be sampled. When sampling $M$ off-diagonal correlations, the sampling properties of correlations sampled without replacement are better than those sampled with replacement, but sampling with replacement takes more time and space.

# Functions

The functions for computing the variance of the number of false discoveries are listed in Table 1.

| | |
|---|---|
| corsampler | Samples correlations |
| critcor | Critical correlation $r_n^{1-\alpha}$ of equation (1) |
| directsimvar | Direct simulation of variance of number of false discoveries |
| fdvar | Variance of number of false discoveries of equation (2) |
| pairsig | Function $C$ of equation (3) |
| varhattaylor | Taylor approximation to variance of number of false discoveries |

Table 1: Function names for `fdvar` R code

## corsampler

Given a matrix $\mathcal{Y}$, obtain a sample of the off-diagonal correlations among columns of $\mathcal{Y}$. By default, the function samples two non-overlapping sets, one with $d_1$ columns and the other with $d_2$ columns, and returns the corresponding $d_1 \times d_2$ "block" of the correlation matrix. The default for $d_2$ is $d_1$ and that for $d_1$ is 500. An alternative strategy, on setting `byblock=FALSE` is to take a sample of size $m$ from the $d(d-1)/2$ off-diagonal correlations. By default that sample is taken with replacement. Both defaults are significantly faster than the alternatives allowing more correlations to be sampled.

## critcor

This function computes the critical value of the sample correlation at which the hypothesis of zero population correlation is rejected. It requires a sample size $n$ a critical value `alph` and the side of the alternative hypothesis, "l", "r", or by default "c".

## directsimvar

This function computes the variance of the number of false discoveries by direct simulation. It requires a distribution from which samples are to be correlated with columns of `ydata`. Of course `ydata` must also be provided. The distribution can be specified as a quantile function `qsrc`. If `qsrc` requires further arguments, they should be provided in the call to `directsimvar`. If a vector `xvals` is provided, then the desired distribution is assumed to be resampling from `xvals`, which should then have length equal to the number of rows in `ydata`. The direct simulation is repeated $m$ times (default $m = 1000$). Arguments `side` and `alph` are as for `critcor`.

When `paired=TRUE` (the dafault) the simulation is coupled to one in which Gaussian random vectors are correlated with columns of `ydata`. When `plot=TRUE` the $m$ sampled numbers of false discoveries are displayed, as a histogram when `paired=FALSE`, or as a scatterplot versus the Gaussian results when `paired=TRUE`.

By default, the function returns the sample mean and variance of the number of false discoveries. When `paired=TRUE` the mean and variance for the Gaussian case are also returned as is the correlation between the Gaussian and non-Gaussian values. If `raw=TRUE` then the simulated numbers of false discoveries are returned as a vector when `paired=FALSE`, or as a two-column matrix when `paired=TRUE`.

## fdvar

This function computes the variance of the number of false disocveries using equation (2). It requires $n$, $d$, $\alpha$ (via `alph`), an indicator (`side` = "l" or "r" or "c") of which alternative is being considered and a sample of the off-diagonal correlations among columns of $\mathcal{Y}$.

The correlations may be provided as the return value from a histogram (via `odcorhist`), as a sample of correlations (via `odcors`), or by providing the raw data (via `ydata`) and the number of correlations to sample (via `rootmcor`). By default a 1000 by 1000 block of off-diagonal correlations is sampled. If a histogram of off-diagonal correlations is not provided, then one is created using `nbrk` break points (default 500). As these are off-diagonal correlations, the argument `odcors` should never be set to the entire correlation matrix of `ydata` as that matrix includes diagonal terms equal to 1, which the function takes account of directly.

The variance estimate can be negative. By default the function gives a warning when this happens. If too few sampled correlations are provided, then the variance estimate is not reliable. The function gives a warning when the variance estimate appears to be under-sampled. To turn off these warnings set `iwarn=FALSE`.

Undersampled variance estimates are much more likely when `alph` is relatively large and when `side` is not equal to "c". Negative variance estimates are much more common when `side` is not equal to "c". It is recommended to run `fdvar` several times to see how stable the answers are.

### pairsig

This function uses adaptive quadrature to find the probability that the two correlations between a Gaussian $X$ and a pair of columns from $\mathcal{Y}$ are simultaneously significantly non-zero. It requires the correlation between the two columns of $\mathcal{Y}$ via `rho`, the sample size $n$, the level of the test via `alph`, and the type of alternative via `side`. If the adaptive quadrature returns with any message other than "OK", the function gives a warning. To suppress such warnings use the option `iwarn=FALSE`. This function is not vectorized. It requires scalar inputs and gives a scalar output.

### varhattaylor

This function uses a Taylor approximation to estimate the variance of the number of false discoveries. It requires as input $d$, $n$, `alph`, `side` as well as two moments: the mean off-diagonal correlation $\eta$ and the mean square off-diagonal correlation $\tau^2$. Note that $\tau^2$ is *not* the variance of the off-diagonal correlations. Taylor approximation is accurate when the off-diagonal correlations are all quite small. In examples with modestly large off-diagonal correlations, it can severely underestimate the desired variance.

## Missing values

Suppose that there are $n_{jj'}$ observations $i = 1, \ldots, n$ for which both $Y_{ij}$ and $Y_{ij'}$ are available. Similarly let $n_j$ denote the number of observations for which both

$X_i$ and $Y_{ij}$ are available. The correlation $\hat{\rho}_{jj'}$ between columns $j$ and $j'$ of $\mathcal{Y}$ is computed using $n_{jj'}$ observations and $\hat{\rho}_j$ is computed using $n_j$ observations.

Assuming that missingness is unrelated to expression, the variance of $\hat{\rho}_{jj'}$ is roughly $(1-\rho_{jj'}^2)/n_{jj'}$ The histogram of correlations $\hat{\rho}_{jj'}$ will be very wide if there are significantly many missing values. Correlations far from zero tend to be the ones that inflate the variance of the number of false discoveries. They will be overrepresented in the histogram making the results conservative.

It is also true that the correlation between $\hat{\rho}_j$ and $\hat{\rho}_{j'}$ is no longer equal to $\hat{\rho}_{jj'}$ when the observation sets used to compute $\hat{\rho}_j$ and $\hat{\rho}_{j'}$ do not coincide. In an extreme case $n_{jj'} = 2 \ll n$ so $\hat{\rho}_{jj'} = \pm 1$ while the true correlation between $\hat{\rho}_j$ and $\hat{\rho}_{j'}$ could be nearly zero.

# References

Owen, A. B. (2004). Variance of the number of false discoveries. Technical report, Stanford University, Department of Statistics.