AFNC

Adaptive False Negative Control (AFNC)

Z. John Daye and X. Jessie Jeng

AFNC: Performs the adaptive false negative control (AFNC), as descripted in Jeng et al. (2016). The AFNC can provide informative analysis of very high-dimensional datasets with weak signals, such as in the analysis of large-scale sequencing studies at the single locus level, by including a large proportion of signals with high confidence. The proportion of signals is estimated adaptively from the data.

This package requires a Fortran 90 compiler. Your build might fail if you don't have it.

Installation

Install AFNC from local source with

```
install.packages("AFNC_1.0.tar.gz", repos=NULL, type="source")
```

Install AFNC from GitHub with

```
library(devtools)
install_github("zjdaye/AFNC")
```

The AFNC

Statististical inference in high-dimensions can be described by three regions: the Signals region where strongly associated variables (SNPs) can be readily identified by controlling false positives, the Noise region where noncausal variables (SNPs) can be identified by controlling false negatives, and the indistinguishable region where causal and noncausal variables (SNPs) are inextricably mixed. In very high-dimensions and under weak signals, such as in next-generation sequencing (NGS) studies, the Signals region tends to be very narrow and can often be degenerate. The adaptive false negative control (AFNC) is proposed to allow a large proportion of causal variables (SNPs) to be retained with high probability by determining the variables (SNPs) that can be confidently dispatched as noncausal.

	Selected	Not selected	Total
Causal	TP	FN	s
Noncausal	FP	TN	$d-s$
	R	$d-R$	d

To characterize false negatives in very high-dimensions, the AFNC employs the signal missing rate (SMR), defined as

$$SMR(j) = P\left(\frac{FN(j)}{s} > \epsilon\right)$$

where FN(j) is the number of causal variables (SNPs) missed by selecting the top j ranked variables (SNPs) and $\epsilon > 0$ is a small constant. The SMR can be interpreted as the probability of neglecting at least a small proportion of causal variables (SNPs) among the top j ranked variables (SNPs). The SMR provides a robust measure of false negatives in high-dimensional applications where the number of causal variables (SNPs) s is expected to be small and, thus, the proportion FN/s is receptive to changes in the number of false negatives.

Users pre-specify the parameters α and β, where α is the level of family-wise error rate for false positive

control using the Bonferroni and β is the level of signal missing rate for false negative control using the AFNC. Smaller value of β corresponds to more stringent control on false negatives. Given α and β, the algorithm implemented in this package is as follows. (See Jeng et al. (2016) for further details.)

1. Perform association tests and obtain ordered p-values from the test statistics, such that the p-values are ordered at decreasing significance, using the **association.test** function.
2. The signal proportion estimator $\hat{\pi}$ and estimated number of signals $\hat{s} = \hat{\pi} \cdot d$ are obtained using the **estimate.signal.proportion** function.
3. Two cutoff positions, t_α and T_{fn}, are determined to separate the Signal, Indistinguishable, and Noise regions using the **AFNC** function. (See Figure 1 of Jeng et al. (2016) for illustration of the Signal, Indistinguishable, and Noise regions of inference.)
4. Finally, variables (SNPs) with ordered p-values ranked at or before t_α are selected by Bonferroni for family-wise false positive control. Variables (SNPs) with ordered p-values ranked at or before T_{fn} are selected by the AFNC procedure for adaptive false negative control using the **AFNC** function.

The AFNC threshold T_{fn} asymptotically controls the signal missing rate at level β for an arbitrarily small constant ϵ that does not change with the number of variables (SNPs) d, allowing the method to be robust under increasing dimensions. The AFNC selects variables (SNPs) with signficances ranked at or before that of T_{fn} to adaptively encompass a large proportion of causal variables (SNPs) with high probability $\approx 1 - \beta$. Thus, decreasing β increases the probability of encompassing nearly all the causal variables (SNPs), but increases the number of selected variables (SNPs) R and, in turn, the number of false positives.

See Jeng et al. (2016) for detailed descriptions and explanations, in addition to comprehensive simulation results and applications to NGS studies.

Example use

The following example performs AFNC given a vector of p-values.

```
set.seed(1)
cd = estimate.cd(d=length(p.value), alpha=0.05)
afnc = AFNC(p.value, alpha=0.05, beta=0.1, cd=cd)$afnc
selected = afnc$index # selected variables
selected.p.value = afnc$p.value # p-values of selected variables
```

The following example performs association test given a matrix of predictors and response.

```
# Simulate response and predictors
set.seed(1); d = 10000; n = 2000
X = array(rnorm(n*d),c(n,d))
y = X[,1:50] %*% (1:50/10) + rnorm(n)

# Performs Wald's test
obj = association.test(X, y, method="Wald")
p.value = obj$p.value; test.stat = obj$test.stat
```

Reference

Jeng, X.J., Daye, Z.J., Lu, W., and Tzeng, J.Y. (2016) Rare Variants Association Analysis in Large-Scale Sequencing Studies at the Single Locus Level.

Licenses

The AFNC package as a whole is distributed under [GPL-3 (GNU General Public License version 3)[http://www.gnu.org/licenses/gpl-3.0.en.html].