# Package ‘FDRreg’

February 19, 2015

**Type** Package

**Title** False discovery rate regression

**Version** 0.1

**Date** 2014-02-24

**Author** James G. Scott, with contributions from Rob Kass and Jesse Windle

**Maintainer** James G. Scott <james.scott@mccombs.utexas.edu>

**Description** Tools for FDR problems, including false discovery rate regression. See corresponding paper: ``False discovery rate regression: application to neural synchrony detection in primary visual cortex.'' James G. Scott, Ryan C. Kelly, Matthew A. Smith, Robert E. Kass.

**License** GPL (>= 3)

**Imports** Rcpp (>= 0.11.0), mosaic (>= 0.8-10)

**Depends** fda (>= 2.4.0), splines (>= 3.0.2)

**LinkingTo** Rcpp, RcppArmadillo

**NeedsCompilation** yes

**Repository** CRAN

**Date/Publication** 2014-03-05 16:42:00

## R topics documented:

FDRreg-package *False discovery rate regression*

## Description

Tools for FDR problems, including false discovery rate regression. Fits models whereby the local false discovery rate may depend upon covariates, either via a linear or additive logistic regression model.

## Details

| | |
|---|---|
| Package: | FDRreg |
| Type: | Package |
| Version: | 1.0 |
| Date: | 2014-02-25 |
| License: | GPL (>=3) |

The workhouse function is FDRreg(z,X, ...), where z is an observed vector of z statistics, and X is a matrix of covariates. Do not add a column of ones to X to get an intercept term; the function does that for you, just like R's base lm() and glm() functions.

## Author(s)

Author: James G. Scott, with contributions from Rob Kass and Jesse Windle.

Maintainer: James G. Scott <james.scott@mccombs.utexas.edu>

## References

False discovery rate regression: application to neural synchrony detection in primary visual cortex. James G. Scott, Ryan C. Kelly, Matthew A. Smith, Pengcheng Zhou, and Robert E. Kass. arXiv:1307.3495 [stat.ME].

## Examples

```
library(FDRreg)

# Simulated data
P = 2
N = 10000
betatrue = c(-3.5,rep(1/sqrt(P), P))
X = matrix(rnorm(N*P), N,P)
psi = crossprod(t(cbind(1,X)), betatrue)
wsuccess = 1/{1+exp(-psi)}

# Some theta's are signals, most are noise
gammatrue = rbinom(N,1,wsuccess)
```

```
table(gammatrue)

# Density of signals
thetatrue = rnorm(N,3,0.5)
thetatrue[gammatrue==0] = 0
z = rnorm(N, thetatrue, 1)
hist(z, 100, prob=TRUE, col='lightblue', border=NA)
curve(dnorm(x,0,1), add=TRUE, n=1001)

## Not run:
# Fit the model
fdr1 <- FDRreg(z, covars=X, nmc=2500, nburn=100, nmids=120, nulltype='theoretical')


# Show the empirical-Bayes estimate of the mixture density
# and the findings at a specific FDR level
Q = 0.1
plotFDR(fdr1, Q=Q, showfz=TRUE)

# Posterior distribution of the intercept
hist(fdr1$betasave[,1], 20)

# Compare actual versus estimated prior probabilities of being a signal
plot(wsuccess, fdr1$priorprob)

# Covariate effects
plot(X[,1], log(fdr1$priorprob/{1-fdr1$priorprob}), ylab='Logit of prior probability')
plot(X[,2], log(fdr1$priorprob/{1-fdr1$priorprob}), ylab='Logit of prior probability')

# Local FDR
plot(z, fdr1$localfdr, ylab='Local false-discovery rate')

# Extract findings at level FDR = Q
myfindings = which(fdr1$FDR <= Q)

## End(Not run)
```

---

FDRreg *False Discovery Rate Regression*

---

## Description

Estimate an empirical-Bayes false-discovery rate regression model for test statistics z and regressors X.

## Usage

```
FDRreg(z, covars, nulltype = 'empirical', type = 'linear', nmc = 10000, nburn = 500,
nmids = 150, densknots = 10, regknots = 5)
```

### Arguments

| | |
|---|---|
| `z` | An N dimensional vector; z_i is the test statistic for observation i. |
| `covars` | An N x P dimensional design matrix; x_i is the ith row. This is assumed not to have a column of ones representing an intercept; just like in lm() and glm(), this will be added by the fitting algorithm. |
| `nulltype` | Choices are 'empirical' for an empirical null using Efron's central-matching method, or 'theoretical' for a standard normal null. |
| `type` | Choices are 'linear' for a standard logistic regression, or 'additive' for an additive logit model, in which case each column of covars is expanded using a b-spline basis. |
| `nmc` | The number of MCMC iterations saved. Defaults to 10,000. |
| `nburn` | The number of initial MCMC iterations discarded as burn-in. Defaults to 500. |
| `nmids` | How many bins should be used in the estimation of the marginal density f(z)? Defaults to 150. |
| `densknots` | How many knots should be used to estimate the marginal density f(z) via spline-based Poisson regression? Defaults to 10; the function will warn you if it looks like you've used too few, using a simple deviance statistic. |
| `regknots` | Used only if type='additive'. How many knots should be used to estimate each partial regression function f_j(x_j)? Defaults to 5. |

### Details

This model assumes that a z-statistic z arises from

$$f(z_i) = w_i f^1(z) + (1 - w_i) f^0(z),$$

where f^1(z) and f^0(z) are the densities/marginal likelihoods under the alternative and null hypotheses, respectively, and where w_i is the prior probability that z_i is a signal (non-null case). Efron's method is used to estimate f(z) nonparametrically; f^0(z) may either be the theoretical (standard normal) null, or an empirical null which can be estimated using the middle 25 percent of the data. The prior probabilities w_i are estimated via logistic regression against covariates, using the Polya-Gamma Gibbs sampler of Polson, Scott, and Windle (JASA, 2013).

### Value

| | |
|---|---|
| `z` | The test statistics provided as the argument z. |
| `localfdr` | The corresponding vector of local false discovery rates (lfdr) for the elements of z. localfdr[i] is simply 1 minus the fitted posterior probability that z[i] comes from the non-null (signal) population. It is important to remember that localfdr is not necessarily monotonic in z, because the regression model allows the prior probability that z[i] is a signal to change with covariates x[i]. |
| `FDR` | The corresponding vector of cut-level false discovery rates (FDR) for the elements of z. Used for extracting findings at a given FDR level. FDR[i] is the estimated false discovery rate for the cohort of test statistics whose local fdr's are at least as small as localfdr[i] — that is, the z[j]'s such that localfdr[j] <= localfdr[i]. |

| | |
|---|---|
| `X` | The design matrix used in the regression. This will include an added column for an intercept, along with the spline basis expansion if type='additive'. |
| `grid` | Length nmids: equally-spaced midpoints of the histogram bins used to estimate f(z) via Poisson spline regression. |
| `breaks` | Length nmids: the breakpoints of the histogram used to estimate f(z) via Poisson spline regression. |
| `grid.fz` | Length nmids: the estimated value of f(z) at the histogram midpoints. |
| `grid.f0z` | Length nmids: the estimated value of f^0(z), the assumed (either theoretical or empirical) null density at the histogram midpoints. |
| `grid.zcounts` | Length nmids: The number of z-scores that fell into each histogram bin. |
| `dnull` | The estimated (or assumed) null density at each of the observed z scores; dnull[i] corresponds to z[i]. |
| `dmix` | The estimated marginal density f(z) at each point z[i]. This should look like a good, smooth fit to the histogram of z. |
| `empirical.null` | A list with two members mu0 and sig0, representing the mean and standard deviation of the empirical null estimated using Efron's central-matching method. Always returned, but only used if nulltype='empirical'. |
| `betasave` | A matrix of posterior draws. Each row is a single posterior draw of the vector of regression coefficients corresponding to the columns of the returned X. |
| `priorprob` | The estimated prior probability of being a signa for each observation z_i. Here priorprob[i] = P(z_i is non-null). |
| `postprob` | The estimated posterior probabilities of being a signal each observation z_i: postprob[i] = P(z_i is non-null \| data), and localfdr[i] = 1-postprob[i]. |
| `fjindex` | A list of indices of length ncol(covars), where covars is the matrix of covariates you fed in. Mainly useful if type='additive', in which case fjind[[j]] gives you a vector of indices telling you which columns of the returned X and betasave correspond to the basis expansion of the original design matrix covars[,j]. |

## References


J.G. Scott, R. Kelly, M.A. Smith, P. Zhou, and R.E. Kass (2013). False discovery rate regression: application to neural synchrony detection in primary visual cortex. arXiv:1307.3495 [stat.ME].

N.G. Polson, J.G. Scott, and J. Windle (2013. Bayesian inference for logistic models using Polya-Gamma latent variables. Journal of the American Statistical Association (Theory and Methods) 108(504): 1339-49 (2013). arXiv:1205.0310 [stat.ME].

Efron (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. J. Amer. Statist. Assoc. 99, 96-104.

Efron (2005). Local false discovery rates. Preprint, Dept. of Statistics, Stanford University.


## Examples

```
library(FDRreg)

# Simulated data
```

```
P = 2
N = 10000
betatrue = c(-3.5,rep(1/sqrt(P), P))
X = matrix(rnorm(N*P), N,P)
psi = crossprod(t(cbind(1,X)), betatrue)
wsuccess = 1/{1+exp(-psi)}

# Some theta's are signals, most are noise
gammatrue = rbinom(N,1,wsuccess)
table(gammatrue)

# Density of signals
thetatrue = rnorm(N,3,0.5)
thetatrue[gammatrue==0] = 0
z = rnorm(N, thetatrue, 1)
hist(z, 100, prob=TRUE, col='lightblue', border=NA)
curve(dnorm(x,0,1), add=TRUE, n=1001)

## Not run:
# Fit the model
fdr1 <- FDRreg(z, covars=X, nmc=2500, nburn=100, nmids=120, nulltype='theoretical')


# Show the empirical-Bayes estimate of the mixture density
# and the findings at a specific FDR level
Q = 0.1
plotFDR(fdr1, Q=Q, showfz=TRUE)

# Posterior distribution of the intercept
hist(fdr1$betasave[,1], 20)

# Compare actual versus estimated prior probabilities of being a signal
plot(wsuccess, fdr1$priorprob)

# Covariate effects
plot(X[,1], log(fdr1$priorprob/{1-fdr1$priorprob}), ylab='Logit of prior probability')
plot(X[,2], log(fdr1$priorprob/{1-fdr1$priorprob}), ylab='Logit of prior probability')

# Local FDR
plot(z, fdr1$localfdr, ylab='Local false-discovery rate')

# Extract findings at level FDR = Q
myfindings = which(fdr1$FDR <= Q)

## End(Not run)
```

---

`plotFDR` *Plot an FDR regression model.*

---

## Description

Plots the results of a fitted FDR regression model from FDRreg.

## Usage

```
plotFDR(fdrr, Q=0.1, showrug=TRUE, showfz=TRUE, showsub=TRUE)
```

## Arguments

| | |
|---|---|
| fdrr | A fitted model object from FDRreg. |
| Q | The desired level at which FDR should be controlled. Defaults to 0.1, or 10 percent. |
| showrug | Logical flag indicating whether the findings at the specified FDR level should be displayed in a rug plot beneath the histogram. Defaults to TRUE. |
| showfz | Logical flag indicating the fitted marginal density f(z) should be plotted. Defaults to TRUE. |
| showsub | Logical flag indicating whether a subtitle should be displayed describing features of the plot. Defaults to TRUE. |

## Details

It is important to remember that localfdr (and therefore global FDR) is not necessarily monotonic in z, because the regression model allows the prior probability that z[i] is a signal to change with covariates x[i].

## Value

No return value.

## Examples

```
library(FDRreg)

# Simulated data
P = 2
N = 10000
betatrue = c(-3.5,rep(1/sqrt(P), P))
X = matrix(rnorm(N*P), N,P)
psi = crossprod(t(cbind(1,X)), betatrue)
wsuccess = 1/{1+exp(-psi)}

# Some theta's are signals, most are noise
gammatrue = rbinom(N,1,wsuccess)
table(gammatrue)

# Density of signals
thetatrue = rnorm(N,3,0.5)
thetatrue[gammatrue==0] = 0
```

```
z = rnorm(N, thetatrue, 1)
hist(z, 100, prob=TRUE, col='lightblue', border=NA)
curve(dnorm(x,0,1), add=TRUE, n=1001)

## Not run:
# Fit the model
fdr1 <- FDRreg(z, covars=X, nmc=2500, nburn=100, nmids=120, nulltype='theoretical')
# Show the empirical-Bayes estimate of the mixture density
# and the findings at a specific FDR level
Q = 0.1
plotFDR(fdr1, Q=Q, showfz=TRUE)

## End(Not run)
```

# Index