Package 'EFDR'

October 12, 2022

Type Package

Title Wavelet-Based Enhanced FDR for Detecting Signals from Complete or Incomplete Spatially Aggregated Data

Version 1.2

Date 2021-04-18

Suggests knitr, rmarkdown, markdown, ggplot2, RCurl, fields, gridExtra, animation

Description Enhanced False Discovery Rate (EFDR) is a tool to detect anomalies in an image. The image is first transformed into the wavelet domain in order to decorrelate any noise components, following which the coefficients at each resolution are standardised. Statistical tests (in a multiple hypothesis testing setting) are then carried out to find the anomalies. The power of EFDR exceeds that of standard FDR, which would carry out tests on every wavelet coefficient: EFDR choose which wavelets to test based on a criterion described in Shen et al. (2002). The package also provides elementary tools to interpolate spatially irregular data onto a grid of the required size. The work is based on Shen, X., Huang, H.-C., and Cressie, N. 'Nonparametric hypothesis testing for a spatial signal.' Journal of the American Statistical Association 97.460 (2002): 1122-1140.

Imports copula, Matrix, methods, foreach (>= 1.4.2), doParallel (>= 1.0.8), waveslim (>= 1.7.5), parallel, gstat (>= 1.0-19), tidyr (>= 0.1.0.9000), dplyr (>= 0.3.0.2), sp (>= 1.0-15)

URL https://github.com/andrewzm/EFDR/

Depends R (>= 3.5.0)

VignetteBuilder knitr

License GPL (>= 2)

Encoding UTF-8

NeedsCompilation no

RoxygenNote 7.1.1

Author Andrew Zammit-Mangion [aut, cre], Hsin-Cheng Huang [aut]

Maintainer Andrew Zammit-Mangion <andrewzm@gmail.com>

Repository CRAN

Date/Publication 2021-04-18 05:50:03 UTC

R topics documented:

`df.to.mat` *Change xyz data-frame into a Z image*

Description

Given a data frame with fields `x`, `y` and `z`, `df.to.mat` uses the `x` and `y` coordinates to rearrange `z` into a rectangular matrix image `Z`.

Usage

```
df.to.mat(df)
```

Arguments

`df`	data frame with fields `x`, `y` and `z`

Details

This function requires that *all* pixels in the image are defined, that is `df$x` and `df$y` must be the column outputs of the function `expand.grid(x0,y0)` where `x0`, `y0` are axes values. Note that `x0` and `y0` do not need to be regularly spaced.

Value

matrix image

Examples

```
df <- data.frame(expand.grid(1:10,1:10))
names(df) <- c("x","y")
df$z <- rnorm(nrow(df))
Z <- df.to.mat(df)
```

`diagnostic.table` *2x2 diagnostic table*

Description

Returns the a 2x2 table resulting from diagnostic evaluation. The cells contain the number of true negatives, true positives, false negatives and false positives.

Usage

```
diagnostic.table(reject.true, reject, n)
```

Arguments

`reject.true`	indices of the true alternative hypotheses
`reject`	indices of the rejected null hypotheses
`n`	total number of tests

Value

2x2 matrix

References

Noel Cressie and Sandy Burden (2015). "Evaluation of diagnostics for hierarchical spatial statistical models." Contribution to K. V. Mardia Festschrift, Wiley, Chichester, forthcoming.

Examples

```
set.seed(1)
wf = "la8"
J = 3
n = 64
h = 0.5
Z <- test_image(h = h, r = 14, n1 = n)$z
sig <- wav_th(Z, wf=wf, J=J, th = h)

Z <- Z + rnorm(n^2)*0.5
m1 <- test.bonferroni(Z, wf="la8",J=3, alpha = 0.05)
m2 <- test.fdr(Z, wf="la8",J=3, alpha = 0.05)

cat("Bonferroni diagnostic table: ",sep="\n")
```

```
diagnostic.table(sig,m1$reject_coeff,n = n^2)
cat("FDR diagnostic table: ",sep="\n")
diagnostic.table(sig,m2$reject_coeff,n = n^2)
```

EFDR *Wavelet-Based Enhanced FDR for Signal Detection in Noisy Images*

Description

Enhanced False Discovery Rate (EFDR) is a tool to detect anomalies in an image. The image is first transformed into the wavelet domain in order to decorrelate any noise components, following which the coefficients at each resolution are standardised. Statistical tests (in a multiple hypothesis testing setting) are then carried out to find the anomalies. The power of EFDR exceeds that of standard FDR, which would carry out tests on every wavelet coefficient: EFDR choose which wavelets to test based on a criterion described in Shen et al. (2002). The package also provides elementary tools to interpolate spatially irregular data onto a grid of the required size. The work is based on Shen, X., Huang, H.-C., and Cressie, N. 'Nonparametric hypothesis testing for a spatial signal.' Journal of the American Statistical Association 97.460 (2002): 1122-1140.

---

fdrpower *Power function*

Description

Returns the power of the multiple hypothesis test, by finding the proportion of the correctly rejected null hypotheses.

Usage

```
fdrpower(reject.true, reject)
```

Arguments

reject.true	indices of the true alternative hypotheses
reject	indices of the rejected null hypotheses

Value

Single value (proportion)

References

Shen, X., Huang, H.-C., and Cressie, N. 'Nonparametric hypothesis testing for a spatial signal.' Journal of the American Statistical Association 97.460 (2002): 1122-1140.

Examples

```
set.seed(1)
wf = "la8"
J = 3
n = 64
h = 0.5
Z <- test_image(h = h, r = 14, n1 = n)$z
sig <- wav_th(Z, wf=wf, J=J, th = h)

Z <- Z + rnorm(n^2)*0.5
m1 <- test.bonferroni(Z, wf="la8",J=3, alpha = 0.05)
m2 <- test.fdr(Z, wf="la8",J=3, alpha = 0.05)

cat(paste0("Bonferroni power: ",fdrpower(sig,m1$reject_coeff)))
cat(paste0("FDR power: ",fdrpower(sig,m2$reject_coeff)))
```

---

nei.efdr *Find wavelet neighbourhood*

Description

Given an image, this function first computes the 2d DWT and then returns a matrix of size `N` by `b` where `N` is the number of wavelets and `b` is the number of neighbours per wavelet. Two wavelets are deemed to be neighbours according to the metric of Shen, Huang and Cressie (2002). The distance metric is a function of the spatial separation, the resolution and the orientation.

Usage

```
nei.efdr(Z, wf = "la8", J = 2, b = 11, parallel = 1L)
```

Arguments

`Z`	image of size `n1` by `n2` where both `n1,n2` have to be powers of two
`wf`	type of wavelet to employ. Please see `waveslim::wave.filter` for a full list of wavelet names
`J`	number of resolutions to employ in the wavelet decomposition
`b`	number of neighbours to consider in EFDR
`parallel`	number of cores to use with parallel backend; needs to be an integer less than the number of available cores

Value

matrix of size `N` by `b`

References

Shen, X., Huang, H.-C., and Cressie, N. 'Nonparametric hypothesis testing for a spatial signal.' Journal of the American Statistical Association 97.460 (2002): 1122-1140.

Examples

```
image <- matrix(rnorm(64),8,8)
```

`regrid` *Regrid ir/regular data*

Description

Given a data frame with fields `x, y` and `z`, `regrid` returns a data frame with fields `x, y` and `z`, this time with `x, y` arranged on a regular grid of size `n2` by `n1`.

Usage

```
regrid(
  df,
  n1 = 128,
  n2 = n1,
  method = "idw",
  idp = 0.5,
  nmax = 7,
  model = "Exp"
)
```

Arguments

`df`	data frame with fields `x`, `y` and `z`
`n1`	image length in pixels
`n2`	image height in pixels
`method`	method to be used, see details
`idp`	inverse distance power
`nmax`	when using inverse distance weighting, the number of nearest neighbours to consider when interpolating using idw. When using conditional simulation, the number of nearest observations to used for a kriging simulation
`model`	the model type when using conditional simulation (use `gstat::vgm()` to list all possible models)

Details

There are three supported methods for regridding. The first, "idw", is the inverse-distance-weighting method. The function overlays a grid over the data. The cells are constructed evenly within the bounding box of the data and filled with interpolated values using the inverse weighting distance metric with power `idp`. `nmax` determines the maximum number of neighbours when using the distance weighting. With this method, interpolation uses the inverse distance weight function `gstat` in the `gstat` package. Refer to the package `gstat` for more details and formulae.

The second method "cond_sim" uses conditional simulation to generate a realisation of the unobserved process at the grid points. This is a model-based approach, and the variogram model may be selected through the parameter `model`. The exponential variogram is used by default. For a complete list of possible models use `gstat::vgm()`. For a tutorial on how the conditional simulation is carried out see the `gstat` vignette.

The third method "median_polishing" applies a median polish to the data. First, a grid is overlayed. If more than one data point is present in each grid box, the mean of the data is taken. Where there is no data, the grid box is assigned a value of NA. This gridded image is then passed to the function `medpolish` which carried out Tukey's median polish procedure to obtain an interpolant of the form $z(s) = \mu + a(s1) + b(s2)$ where $s1$ is the x-axis and $s2$ is the y-axis. Missing points in the gridded image are then replaced with $z(s)$ evaluated at these points. This method cannot be used if all rows and columns do not contain at least one data point.

Value

data frame with fields `x,y,z`

Examples

```
df <- data.frame(x = runif(200),y = runif(200),z=rnorm(200))
df.gridded <- regrid(df, n1=10)
```

---

test.efdr.condsim *Test for anomalies in wavelet space via conditional simulation*

Description

Test for anomalies using `EFDR` and conditional simulation. The noisy image can be partially observed, or/and aggregated at different resolutions

Usage

```
test.efdr.condsim(
  Zvec,
  H,
  n1,
  n2,
  rho_est_method = c("CPL", "MOM"),
  iter.cs = 100,
  wf = "la8",
  J = 2,
  alpha = 0.05,
  n.hyp = 100,
  b = 11,
  iteration = 200,
  parallel = 1L
)
```

Arguments

Zvec	vector of observations such that $Ztilde = H.Z$
H	matrix mapping the fine-resolution image Z in vector form to Ztllde. Must have as many rows as Ztilde and n1 x n2 columns
n1	number of rows in fine-resolution image
n2	number of columns in fine-resolution image
rho_est_method	method with which to estimate the level of exchangeability rho; can be either "CPL" (copula model) or "MOM" (method of moments)
iter.cs	number of conditional simulations to carry out
wf	type of wavelet to employ. Defaults to 'la8', the Daubechies orthonormal compactly supported wavelet of length L = 8 (Daubechies, 1992), least asymmetric family. Other options include 'haar' (Haar wavelet), 'fk8' (Fejer-Korovkin wavelet with L=8) and 'mb8' (minimum-bandwidth wavelet with L=8). Please type 'waveslim::wave.filter' in the console for a full list of wavelet names
J	number of resolutions to employ in wavelet decomposition
alpha	significance level at which tests are carried out
n.hyp	number of hypotheses tests to carry out with EFDR. If a vector is supplied, the optimal one from the set of proposed number of tests is chosen
b	the number of neighbours to consider in EFDR
iteration	number of Monte Carlo iterations to employ when determining which of the proposed number of tests in n.hyp is the optimal number of tests
parallel	number of cores to use with parallel backend; needs to be an integer less than or equal to the number of available cores

Value

List with three fields:

filtered the discrete wavelet transform containing the anomalous wavelet coefficients in the signal

Z the image containing the anomalous wavelets in the signal

reject_coeff indices of wavelets under which the null hypothesis of no anomaly was rejected

pvalue_ordered ordered p-values under the null hypothesis. The column names indicate the wavelet to which the p-value belongs

nhat the number of tests carried out.

References

Daubechies, I. (1992) Ten Lectures on Wavelets, CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM: Philadelphia.

Shen, X., Huang, H.-C., and Cressie, N. 'Nonparametric hypothesis testing for a spatial signal.' Journal of the American Statistical Association 97.460 (2002): 1122-1140.

Examples

```
## Set up experiment
n <- 32       # 32 x 32 images
r <- 10       # signal of size 10 x 10
h <- 5        # intensity of 5
grid <- 8     # aggregated to 8 x 8 image
parallel <- 4 # use 4 cores

## Simulate the pixel-level data
raw_grid <- expand.grid(x = seq(1, n), y = seq(1, n))
df <- data.frame(raw_grid)                          # spatial grid
dd <- as.matrix(dist(raw_grid, diag = TRUE))      # distance matrix
Sigma <- exp(-dd/5)                                 # cov. fn.
diag(Sigma) <- 1                                    # fix diagonal
L <- t(chol(Sigma))                                 # lower Cholesky factor
mu <- matrix(0, n, n)                               # zero mean
mu[(n/2-r/2):(n/2+r/2), (n/2-r/2):(n/2+r/2)] <- h # add signal
Z <- mu + matrix(L %*% rnorm(n^2), n, n)          # simulate data

## Construct H (aggregation) matrix
H <- matrix(0, grid^2, n^2)
for(i in 1:grid^2) {
  ind <- rep(rep(c(0L,1L,0L),
             c((n/grid)*((i-1)%%grid),n/grid,(n-n/grid-n/grid*((i-1)%%grid)))),
         n/grid)
  H[i,which(c(rep(0L,(ceiling(i/grid)-1)*n^2/grid),ind) == TRUE)] <- 1/(n/grid)^2
}

## Aggregate the signal
z_tilde <- c(H %*% c(Z))

## Run EFDR using conditional simulation
## Not run: out2 <- test.efdr.condsim(Zvec = z_tilde, H = H, n1 = n, n2 = n,
                                    parallel = parallel)
## End(Not run)
```

`test_image` *Create a test image*

Description

This function generates an image for test purposes. The image is that of a filled circle at the centre.

Usage

```
test_image(h = 1, r = 10, n1 = 64, n2 = 64)
```

Arguments

h	amplitude of the filled circle
r	radius of the circle (in pixels)
n1	image height in pixels
n2	image width in pixels

Value

List with two elements

z the test image

signal.grid the x-y grid in long table format

References

Shen, X., Huang, H.-C., and Cressie, N. 'Nonparametric hypothesis testing for a spatial signal.' Journal of the American Statistical Association 97.460 (2002): 1122-1140.

Examples

```
Z <- test_image()$z
```

---

wavelet-test *Test for anomalies in wavelet space*

Description

Test for anomalies using either bonferroni, FDR, EFDR or LOS in the wavelet domain using the 2D wavelet transform.

Usage

```
test.efdr(
  Z,
  wf = "la8",
  J = 2,
  alpha = 0.05,
  n.hyp = 100,
  b = 11,
  iteration = 200,
  parallel = 1L
)

test.fdr(Z, wf = "la8", J = 2, alpha = 0.05)

test.bonferroni(Z, wf = "la8", J = 2, alpha = 0.05)

test.los(Z, wf = "la8", J = 2, alpha = 0.05)
```

Arguments

Z	image of size n1 by 2n where n1,n2 have to be powers of two
wf	type of wavelet to employ. Defaults to 'la8', the Daubechies orthonormal compactly supported wavelet of length L = 8 (Daubechies, 1992), least asymmetric family. Other options include 'haar' (Haar wavelet), 'fk8' (Fejer-Korovkin wavelet with L=8) and 'mb8' (minimum-bandwidth wavelet with L=8). Please type 'waveslim::wave.filter' in the console for a full list of wavelet names
J	number of resolutions to employ in wavelet decomposition
alpha	significance level at which tests are carried out
n.hyp	number of hypotheses tests to carry out with EFDR. If a vector is supplied, the optimal one from the set of proposed number of tests is chosen
b	the number of neighbours to consider in EFDR
iteration	number of Monte Carlo iterations to employ when determining which of the proposed number of tests in n.hyp is the optimal number of tests
parallel	number of cores to use with parallel backend; needs to be an integer less than or equal to the number of available cores

Value

List with three fields:

filtered the discrete wavelet transform containing the anomalous wavelet coefficients in the signal

Z the image containing the anomalous wavelets in the signal

reject_coeff indices of wavelets under which the null hypothesis of no anomaly was rejected

pvalue_ordered ordered p-values under the null hypothesis. The column names indicate the wavelet to which the p-value belongs

nhat the number of tests carried out.

References

Daubechies, I. (1992) Ten Lectures on Wavelets, CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM: Philadelphia.

Shen, X., Huang, H.-C., and Cressie, N. 'Nonparametric hypothesis testing for a spatial signal.' Journal of the American Statistical Association 97.460 (2002): 1122-1140.

Examples

```
## See vignettes by typing vignette("EFDR_vignettes")
```

wav_th *Indices of wavelets exceeding a given threshold*

Description

This function is primarily used for testing the power of a method in the wavelet domain. Given an image, the discrete wavelet transform is found. The indices of the coefficients which exceed a certain threshold are then considered the 'signal' for testing purposes.

Usage

```
wav_th(Z, wf = "la8", J = 2, th = 1)
```

Arguments

Z	image of size n1 by n2 where n1,n2 have to be powers of two
wf	type of wavelet to employ. Please see waveslim::wave.filter for a full list of wavelet names
J	number of resolutions to employ in the wavelet decomposition
th	threshold

Value

Indices of wavelet coefficients in a vector

References

Shen, X., Huang, H.-C., and Cressie, N. 'Nonparametric hypothesis testing for a spatial signal.' Journal of the American Statistical Association 97.460 (2002): 1122-1140.

Examples

```
Z <- test_image(h = 0.5, r = 14, n1 = 64)$z
print(wav_th(Z,wf="la8",J=3,th=0.5))
```

Index