# Package ‘NPMLEmix’

December 6, 2020

<table>
<thead>
<tr>
<th>Type</th>
<th>Package</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Two-Groups Mixture Model with Covariates</td>
</tr>
<tr>
<td>Version</td>
<td>1.2</td>
</tr>
<tr>
<td>Date</td>
<td>2020-12-05</td>
</tr>
<tr>
<td>Depends</td>
<td>R (&gt;= 3.0.1), Rmosek</td>
</tr>
<tr>
<td>Maintainer</td>
<td>Nabarun Deb <a href="mailto:nd2560@columbia.edu">nd2560@columbia.edu</a></td>
</tr>
<tr>
<td>Description</td>
<td>We develop three procedures for estimation in a two-groups model with covariates. One of them is a nonparametric maximum likelihood estimation based approach when the signal distribution is an infinite Gaussian location mixture and the signal proportion is a logistic function of the available covariates. Two other functions - marg1() and marg2() have also been implemented for inference in the above framework. All these methods can be used for inference in multiple hypotheses testing. For more information, see the paper: Deb, Saha, Guntuboyina and Sen (2019), &quot;Two-component Mixture Model in the Presence of Covariates&quot; <a href="">arXiv:1810.07897v2</a>.</td>
</tr>
<tr>
<td>License</td>
<td>GPL-2</td>
</tr>
<tr>
<td>Encoding</td>
<td>UTF-8</td>
</tr>
<tr>
<td>Author</td>
<td>Nabarun Deb [aut, cre], Sujayam Saha [ctb], Adityanand Guntuboyina [ctb], Bodhisattva Sen [ctb]</td>
</tr>
<tr>
<td>LazyData</td>
<td>true</td>
</tr>
<tr>
<td>RoxygenNote</td>
<td>7.1.1</td>
</tr>
<tr>
<td>NeedsCompilation</td>
<td>no</td>
</tr>
<tr>
<td>SystemRequirements</td>
<td>MOSEK (<a href="http://www.mosek.com">http://www.mosek.com</a>) and MOSEK license.</td>
</tr>
<tr>
<td>URL</td>
<td><a href="https://github.com/NabarunD/NPMLEmix">https://github.com/NabarunD/NPMLEmix</a></td>
</tr>
<tr>
<td>Imports</td>
<td>Rutils, dplyr, spatstat, methods, latentpdf, Matrix, FDRreg, Rcpp, REBayes, CAMAN, progress, pbapply, Hmisc, pracma, mosaic</td>
</tr>
<tr>
<td>Repository</td>
<td>CRAN</td>
</tr>
<tr>
<td>Date/Publication</td>
<td>2020-12-06 05:50:02 UTC</td>
</tr>
</tbody>
</table>
Description

The NPMLEmix-package fits nonparametric Gaussian location mixture models for z-scores arising out of several hypotheses, while taking into account any available covariate information. It also provides three important functions: marg1(), marg2() (both based on marginal likelihoods), npmleEM() (based on joint data likelihood) for inference in multiple testing.

Details

\[(Y_1, X_1), \ldots, (X_n, Y_n)\] are i.i.d. samples drawn from the model,

\[Y | X = x \sim (1 - \pi^*(x))\phi(y) + \pi^*(x) \int_\theta \phi(y - \theta) dG(\theta), \quad X \sim m_X(\cdot)\]

where \(\pi^*(\cdot)\) represents the logistic link function, \(\phi(\cdot)\) is the standard Gaussian density and \(G(\cdot)\) is some unknown probability measure on the real line. Usually, \(\pi^*(\cdot)\) is referred to as the signal proportion, \(\phi_1(\cdot)\) is called the signal density and \(G(\cdot)\) is called mixing distribution. The \(i^{th}\) local false discovery rate is then defined as

\[lfdr_i = \frac{(1 - \pi^*(X_i))\phi_0(Y_i)}{(1 - \pi^*(X_i))\phi_0(Y_i) + \pi^*(X_i)\phi_1(Y_i)}\]

All the principal functions estimate the unknown parameters - \(\pi^*(\cdot)\) and \(\phi_1(\cdot)\), and consequently the \(lfdr_i\)'s. The optimization algorithms use quasi-Newton routines such as the BFGS (Broyden-Fletcher-Goldfarb-Shanno) algorithm and the separable convex optimization routine available in the Rmosek optimization suite. The principal functions accept a vector of z-scores (\(Y\))’s and a covariate matrix \(X\) in their list of arguments. Read the documentations for each function to check whether or not to add a column of 1’s to \(X\) matrix.

Functions

The principal functions in the NPMLEmix-package: marg1(), marg2(), npmleEM().
makedata

Simulates data from the aforementioned model

Description
This function can be used to simulate observations from the aforementioned model, if $G(\cdot)$ is chosen as a finite Gaussian mixture. It returns the true local false discovery rates which determine the optimal multiple testing procedure.

Usage
makedata(n, x, sx, atoms, probs, variances)

Arguments
n  Number of z-scores to be generated.
x  $n \times p$ data matrix. Do not add an additional column of 1’s.
sx  The vector of coefficients for the logistic function. The first entry will be considered as the intercept term by default. Requires compatibility with x. See Details.
atoms  The vector of means for each component of the mixing distribution.
probs  The probability vector for the mixing distribution.
variances  The vector of variances for each component of the mixing distribution. Requires compatibility with atoms and probs. See Details.

Details
Given $X = x$, a Bernoulli($\pi^*(x)$) sample is drawn. If the outcome is 1 (0), a z-score is drawn from $\phi_1(\cdot)$ ($\phi(\cdot)$). All the observations corresponding to a Bernoulli outcome 1 (0) are termed as non-null observations (null observations).
The length of sx should be 1 more than the number of columns of the data matrix x.
The vectors - atoms, probs and variances must have the same length.
Value

The output is a list with the following entries:

- **y**: The vector of simulated z-scores.
- **x**: The input data matrix.
- **pix**: The vector of signal proportions.
- **f0y**: The vector of standard Gaussian densities evaluated at simulated z-scores.
- **f1y**: The vector of signal densities evaluated at simulated z-scores.
- **den**: The vector of conditional densities evaluated at simulated z-scores.
- **localfdr**: The vector of local false discovery rates evaluated at simulated z-scores. Note that the local FDR can be interpreted as one minus the posterior probability that a given observation is non-null.
- **ll**: The average conditional log-likelihood.
- **nnind**: The indices corresponding to non-null observations.

References


Examples

```r
x=cbind(runif(1000),runif(1000))
n=1000
atoms=c(-2,0,2)
probs=c(0.48,0.04,0.48)
variances=c(1,16,1)
sx=c(-3,1.5,1.5)
### Generating the data ###
st=makedata(n,x,sx,atoms,probs,variances)
### Output the vector of local false discovery rates ###
st$localfdr
```

---

Implements a profile likelihood based algorithm for estimating signal proportion and density.
Description
This function estimates the signal proportion and the signal density by using the marginal distribution of $Y$, followed by a profile likelihood based approach. It returns the vector of estimated local false discovery rates and the corresponding rejection set at a prespecified level for the false discovery rate.

Usage
marg1(y, x, blambda = 1e-06/length(y), level = 0.05)

Arguments
- **y**: The observed vector of z-scores.
- **x**: The $n \times p$ data matrix, where $n$ must be equal to the length of $y$. If you are interested in the intercept, you must add a column of 1’s to $x$.
- **blambda**: The tolerance threshold while implementing a quasi-Newton approach for estimating the signal proportion. Default is set to $1e-6/\text{length}(y)$. We recommend not changing it unless absolutely sure.
- **level**: The level at which the false discovery rate is to be controlled. Should be a scalar in $[0, 1]$. Default set to 0.05.

Details
Note that the marginal distribution of $Y$ based on the aforementioned model is same as that in a standard two-groups model (Efron 2008, see References). Fixing $\bar{\pi} = \mathbb{E}[^\pi](X)$, the signal density $\phi_1(\cdot)$ is estimated using the Rmosek optimization suite. The primary idea is to approximate the mixing distribution $G(\cdot)$ using $\max\{100, \sqrt{\bar{n}}\}$ many components, each having a suitable Gaussian distribution. The signal proportion is then estimated using the BFGS algorithm. Finally, the algorithm chooses the best value of $\bar{\pi}$ based on a profile likelihood approach.

Value
This function returns a list consisting of the following:
- **p**: The estimated prior probabilities, i.e., $\hat{\pi}(\cdot)$ evaluated at the data points.
- **b**: The estimates for the coefficient vector in the logistic function.
- **f1y**: The vector of estimated signal density evaluated at the data points.
- **kwo**: This is a list with four items - i. **atoms**: The vector of means for the Gaussian distributions used to approximate $G(\cdot)$, ii. **probs**: The vector of probabilities for each Gaussian component used to approximate $G(\cdot)$, iii. **f1y**: Same as f1y above, iv. **ll**: The average of the logarithmic values of f1y.
- **localfdr**: The vector of estimated local false discovery rates evaluated at the data points.
- **den**: The vector of estimated conditional densities evaluated at the data points.
- **ll**: The log-likelihood evaluated at the estimated optima.
- **rejset**: The vector of 1s and 0s where 1 indicates that the corresponding hypothesis is to be rejected.
pi0  The average of the entries of the vector p.
ll_list  The vector of profile log-likelihoods corresponding to a pre-determined set of
grid points for \( \bar{\pi} \). The highest element of this vector is the output in \( ll \).

References
Koenker, R. and Mizera, I., 2014. Convex optimization, shape constraints, compound decisions,

marg2()

Implements a non-linear least squares based algorithm for estimating
signal proportion and density

Description
This function estimates the signal proportion and the signal density by using the conditional mean 
\( Y|X = x \), followed by a non-linear least squares regression based approach. It returns the vector of 
estimated local false discovery rates and the corresponding rejection set at a prespecified level for 
the false discovery rate.

Usage

marg2(y, x, nlslambda = 1e-06/length(y), level = 0.05)

Arguments

y  The observed vector of z-scores.
x  The \( n \times p \) data matrix, where \( n \) must be equal to the length of \( y \). If you are
interested in the intercept, you must add a column of 1’s to \( x \).
nlslambda  The tolerance threshold while implementing a quasi-Newton approach for the
non-linear least squares problem. Default is set to \( 1e - 6/\text{length}(y) \). We recom-

level  The level at which the false discovery rate is to be controlled. Should be a scalar
in \([0, 1]\). Default set to 0.05.

Details
Note that the conditional mean of \( Y|X \) based on the aforementioned model is a non-linear function
of the parameters, i.e., the logistic coefficients and the mean of the marginal distribution of \( Y \),
\( \mu^* = E[Y] \). This is a non-convex optimization problem in the parameters and is solved by varying
\( \mu^* \) over a predetermined grid, and optimizing over the logistic coefficients. This is the estimate of
\( \pi^*(\cdot) \) from the marg2() method. The estimate of \( \phi_1(\cdot) \) is obtained as in the marg1() method by using
the Rmosek optimization suite, and the same discrete approximation to the mixing distribution \( G(\cdot) \).
Value

This function returns a list consisting of the following:

- **p**: The estimated prior probabilities, i.e., \( \hat{\pi}(\cdot) \) evaluated at the data points.
- **b**: The estimates for the coefficient vector in the logistic function.
- **fly**: The vector of estimated signal densities evaluated at the data points.
- **kwo**: This is a list with four items - i. **atoms**: The vector of means for the Gaussian distributions used to approximate \( G(\cdot) \), ii. **probs**: The vector of probabilities for each Gaussian component used to approximate \( G(\cdot) \), iii. **fly**: Same as fly above, iv. **ll**: The average of the logarithmic values of fly.
- **localfdr**: The vector of estimated local false discovery rates evaluated at the data points.
- **den**: The vector of estimated conditional densities evaluated at the data points.
- **ll**: The log-likelihood evaluated at the estimated optima.
- **rejset**: The vector of 1s and 0s where 1 indicates that the corresponding hypothesis is to be rejected.
- **pi0**: The average of the entries of the vector **p**.
- **ll_list**: The vector of profile log-likelihoods corresponding to a pre-determined set of grid points for \( \mu^* \). The highest element of this vector is the output in **ll**.

References


---

**npmleEM**

*Implements the full likelihood approach based on the EM algorithm for estimating signal proportion and density*

**Description**

This function estimates the signal proportion and the signal density by using the full likelihood of the sample, followed by an EM algorithm based approach. It returns the vector of estimated local false discovery rates and the corresponding rejection set at a prespecified level for the false discovery rate.

**Usage**

```r
npmleEM(y, x, level = 0.05, initp = 1)
```
Arguments

y
The observed vector of z-scores.

x
The \( n \times p \) data matrix, where \( n \) must be equal to the length of \( y \). If you are interested in the intercept, you must add a column of 1’s to \( x \).

level
The level at which the false discovery rate is to be controlled. Should be a scalar in \([0, 1]\). Default set to 0.05.

initp
The initialization method for the EM algorithm. It should be either 1, 2, 3 or 4. 
1 indicates a marg1() initialization, 2 indicates a marg2() initialization, 3 indicates a FDRreg() initialization (see Details and References) and 4 chooses that initialization among marg1(), marg2() and FDRreg() which yields the highest sample likelihood. Default is set to 1.

Details

The key observation in the full likelihood approach is that the M-step of the EM algorithm results in two decoupled optimization problems, one involving \( \pi^*(\cdot) \) and the other involving \( \phi_1(\cdot) \). These two individual problems are then solved using the BFGS algorithm and the Rmosek optimization suite, as has been discussed previously in the Details sections of the methods marg1() and marg2(). The FDRreg() method was introduced in Scott et al (2015). We recommend using the version of the FDRreg() package available in https://github.com/jgscott/FDRreg/tree/master/R_pkg.

Value

This function returns a list consisting of the following:

atoms
The vector of means for the Gaussian distributions used to approximate \( G(\cdot) \).

probs
The vector of probabilities for each Gaussian component used to approximate \( G(\cdot) \).

f1y
The vector of estimated signal densities evaluated at the data points.

f0y
The vector of null densities evaluated at the data points.

b
The estimates for the coefficient vector in the logistic function.

p
The estimated prior probabilities, i.e., \( \hat{\pi}(\cdot) \) evaluated at the data points.

ll
The log-likelihood evaluated at the estimated optima.

rejset
The vector of 1s and 0s where 1 indicates that the corresponding hypothesis is to be rejected.

den
The vector of estimated conditional densities evaluated at the data points.

localfdr
The vector of estimated local false discovery rates evaluated at the data points.

References


**Description**

This function accepts a vector of local false discovery rates from a family of hypotheses and a level parameter, to compute the rejection set.

**Usage**

```r
reject_set(locfdr, level = 0.1)
```

**Arguments**

- `locfdr`: The vector of local false discovery rates (actual or estimated) corresponding to a family of hypotheses.
- `level`: The level at which the false discovery rate is to be controlled. Should ideally be a scalar in \([0, 1]\).

**Details**

The problem of optimal inference in multiple hypotheses testing has been widely studied in literature. In particular, this function adopts the framework and algorithm proposed in Basu et al. See References.

**Value**

A vector of 1s and 0s with 1s indicating the hypotheses which are to be rejected.

**References**


**Examples**

```r
dx=cbind(runif(1000),runif(1000))
n=1000
atoms=c(-2,0,2)
probs=c(0.48,0.04,0.48)
variances=c(1,16,1)
sx=c(-3,1.5,1.5)
stdata=makedata(n,x,sx,atoms,probs,variances)
### Obtain the rejection set ###
reject=reject_set(stdata$lo)
```
Index

makedata, 3
marg1, 4
marg2, 6

npmleEM, 7
NPMLEmix-package, 2

reject_set, 9