

# Package ‘jewel’

April 6, 2023

**Title** Graphical models estimation from multiple sources

**Version** 2.0.0

**Description** Estimates networks of conditional dependencies (Gaussian graphical models) from multiple classes of data (similar but not exactly, i.e. measurements on different equipment, in different locations or for various subtypes). Package also allows to generate simulation data and evaluate the performance. Based on <https://doi.org/10.3390/math10213983>

**Depends** R (>= 3.6.0)

**Imports** Matrix, matrixcalc, MASS, SMUT, igraph, parallel, purrr

**URL** <https://github.com/annaplaksienko/jewel>

**License** GPL-2

**Encoding** UTF-8

**RoxygenNote** 7.2.3

**NeedsCompilation** no

**Author** Anna Plaksienko [aut, cre] (<<https://orcid.org/0000-0001-9607-057X>>),  
Claudia Angelini [aut] (<<https://orcid.org/0000-0001-8350-8464>>),  
Daniela De Canditiis [aut] (<<https://orcid.org/0000-0002-3022-3411>>)

**Maintainer** Anna Plaksienko <[anna@plaxienko.com](mailto:anna@plaxienko.com)>

## R topics documented:

constructWeights . . . . .	1
evaluatePerformance . . . . .	2
generateData_rewire . . . . .	3
jewel . . . . .	4

---

constructWeights	<i>Construct weights for _jewel_ minimization problem from prior information on vertices degrees.</i>
------------------	---

---

## Description

Function takes a numerical vector of vertices degrees and constructs weights with the rule  $W_{ij} = 1 / \sqrt{d_i * d_j}$  and then the whole matrix is normalized by the maximum.

**Usage**

```
constructWeights(d, K = NULL)
```

**Arguments**

- d** either one numerical vector or a list of  $K$  numerical vectors of length  $p$  with user-provided degrees of vertices for each class. If there is only one vector, we assume degrees are the same for all  $K$  classes. In that case parameter  $K$  (number of classes) must be provided. Note that for successful `_jewel_` estimation true degrees are not necessary: for example, user can provide a vector where known hubs have degree 10 and the rest of the vertices have degree 1.
- K** number of classes (i.e. datasets, i.e. desired graphs). By default it is `length(d)`. In `length(d) = 1`,  $K$  must be provided by the user.

**Value**

**W** - a list of  $K$  numeric matrices of the size  $p$  by  $p$

---

evaluatePerformance	<i>Evaluation of graph estimation method's performance if the true graph is known.</i>
---------------------	--

---

**Description**

Function compares adjacency matrices of the true and estimated simple graphs and calculates the number of true positives (correctly estimated edges), true negatives (correctly estimated absence of edges), false positives (edges present in the estimator but not in the true graph) and false negatives (failure to identify an edge).

**Usage**

```
evaluatePerformance(G, G_hat)
```

**Arguments**

- G** true graph's adjacency matrix.
- G\_hat** estimated graph's adjacency matrix. Must have the same dimensions as **G**.

**Value**

**performance** - a numeric vector of length 4 with TP, TN, FP, FN.

---

generateData_rewire	<i>Generate a set of scale-free graphs and corresponding datasets (using the graphs as their Gaussian graphical models)</i>
---------------------	---

---

## Description

Function first generates  $K$  scale-free graphs with  $p$  vertices. They have the same order and degree distribution and share most of the edges, but some edges may vary (user can control how many). Function then generates corresponding precision and covariance matrices, all of the size  $p$  by  $p$  (see the paper for the details of the procedure). Then for each  $l$ -th element of vector  $n$  it generates  $K$  data matrices, each of the size  $n_l$  by  $p$ , i.e., for the same underlying graphs we can generate several sets of  $K$  datasets with different sample sizes.

## Usage

```
generateData_rewire(
  K,
  p,
  n,
  power = 1,
  m = 1,
  perc = 0.05,
  int = NULL,
  makePlot = TRUE,
  verbose = TRUE
)
```

## Arguments

$K$	number of graphs/data matrices.
$p$	number of nodes in the true graphs.
$n$	a numerical vector of the sample sizes for each desired set of $K$ data matrices. Can be a vector of one element if the user wishes to obtain only one dataset of $K$ matrices.
power	a number, power of preferential attachment for the Barabasi-Albert algorithm for the generation of the scale-free graph. Bigger number means more connected hubs. The default value is 1.
$m$	number of edges to add at each step of Barabasi-Albert algorithm for generation of the scale-free graph. Resulting graph has $mp - (2m - 1)$ edges. The default value is 1.
perc	a number, tuning parameter for the difference between graphs. Number of trials to perform in the rewiring procedure of the first graph is $p * \text{perc}$ . Bigger the number, more different are the graphs.
int	a vector of two numbers, $a$ and $b$ . Entries of precision matrices are sampled from the uniform distribution on the interval $[-b, -a] + [a, b]$ . The default values are $a = 0.2$ , $b = 0.8$ .
makePlot	If <code>makePlot = FALSE</code> , plotting of the generated graphs is disabled. The default value is <code>TRUE</code> .
verbose	If <code>verbose = FALSE</code> , tracing information printing is disabled. The default value is <code>TRUE</code> .

**Value**

The following list is returned

- Graphs – a list of adjacency matrices of the K generated graphs.
- CommomGraph - a matrix, common part (intersection) of the K generated graphs.
- Data - a list of lists, for each sample size of the input vector n one obtains K data matrices, each of the size  $n_l$  by p.
- Sigma - a list of K covariance matrices of the size p by p.

---

jewel

---

*Estimate Gaussian graphical models from multiple datasets*


---

**Description**

This function estimates Gaussian graphical models (i.e. networks of conditional dependencies, direct connections between variables) given multiple datasets. We assume that datasets contain measurements of the same variables collected under different conditions (different equipment, locations, even sub-types of disease).

**Usage**

```

jewel(
  X,
  lambda1,
  lambda2 = NULL,
  Theta = NULL,
  W = NULL,
  tol = 0.01,
  maxIter = 10000,
  stability = FALSE,
  stability_nsubsets = 25,
  stability_frac = 0.8,
  verbose = TRUE
)

```

**Arguments**

- |         |   |
|---------|---|
| X       | a list of K numeric data matrices of $n_k$ samples and p variables ( $n_k$ can be different for each matrix).   |
| lambda1 | a number, first regularization parameter (of the common penalty).   |
| lambda2 | an optional number, second regularization parameter (of the class-specific penalty). If NULL, set to $\lambda_2 = \lambda_1 * 1.4$  |
| Theta   | an optional list of K regression coefficient matrices of the size p by p. User-provided initialization can be used for warm-start procedures. If NULL, initialized as all zeros.  |
| W       | an optional list of K weights matrices of the size p by p. User-provided initialization can be used when some vertices are believed to be hubs. If NULL, initialized as all ones. |

<code>tol</code>	an optional number, convergence threshold controlling the relative error between iterations. The default value is 0.01.
<code>maxIter</code>	an optional number, maximum allowed number of iterations. The default value is 10 000.
<code>stability</code>	if <code>stability = TRUE</code> , stability selection procedure to reduce the number of false positives will be applied. $n_k / 2$ samples are randomly chosen in each dataset <code>stability_nsubsets</code> times and then <code>__jewel__</code> method is applied to each subset. In the final estimate, we include only the edges that appear in at least <code>stability_frac</code> proportion of the subsets. By default this procedure is disabled since it increases the running time.
<code>stability_nsubsets</code>	an optional number, how many times to subsample datasets and apply <code>__jewel__</code> for stability selection procedure. The default value is 25.
<code>stability_frac</code>	an optional number, in what proportion of the stability results on subsampled data an edge has to be present to be included into the final estimate. The default value is 0.8.
<code>verbose</code>	if <code>verbose = FALSE</code> , tracing information printing is disabled. The default value is <code>TRUE</code> .

## Value

The following list is returned

- `CommonG` - an adjacency matrix of the common estimated graph (intersection of  $K$  estimated graphs).
- `G_list` - a list of  $K$  adjacency matrices for each estimated graph.
- `Theta` - a list of  $K$  estimated covariance matrices (when stability selection is disabled).
- `BIC` - a number, value of Bayesian information criterion for resulting graphs (when stability selection is disabled).