Package ‘emdist’
February 19, 2015

Version 0.3-1
Title Earth Mover’s Distance
Author Simon Urbanek <urbanek@research.att.com>, Yossi Rubner
Maintainer Simon Urbanek <simon.urbanek@r-project.org>
Description Package providing calculation of Earth Mover’s Distance (EMD).
License MIT
URL http://www.rforge.net/emd
Repository CRAN
Date/Publication 2012-12-02 07:14:50
NeedsCompilation yes

R topics documented:

emd ................................................................. 1

Index

emd

Earth Mover’s Distance

Description

emd computes Earth Mover’s Distance (related to 1st Mallows and Wasserstein distances) between distributions. emd and emdw use (weight,location) notation whereas emd2d compares two distributions represented as matrices over a grid.

Usage

emd(A, B, dist=“euclidean”, ...)
emdw(A, wA, B, wB, dist=“euclidean”, ...)
emd2d(A, B, xdist = 1, ydist = 1, dist=“euclidean”, ...)
emdr(A, B, extrapolate=NA, flows=FALSE, dist=“euclidean”, ...)
Arguments

- **A**: matrix A
- **B**: matrix B
- **extrapolate**: if set to 1 or 2 the mass of A or B respectively is used to extrapolate the distance by penalization using the mass quotient assuming the other signature is truncated and thus more unlikely to match. It has any effect only if the other specified signature has larger mass.
- **flows**: logical indicating whether flows should be returned in the "flows" attribute of the result.
- **wA**: weights for locations specified by A
- **wB**: weights for locations specified by B
- **xdist**: distance between columns (scalar) or a vector of positions of the columns
- **ydist**: distance between rows (scalar) of a vector of positions of the rows
- **dist**: distance to be used for the computation of the cost over the locations. Must be either "euclidean", "manhattan" or a closure taking two vectors and returning a scalar number. The latter case is much less efficient because it requires R evaluation for every possible combination of flows.
- **...**: additional parameters for future use, currently unused

Details

emd2d interprets the two matrices A and B as a distribution over a two-dimensional grid. The distance between the grid points in each direction is defined by xdist and ydist. Both matrices must have the same dimensionality.

emd uses first column of each matrix as the weights and the remaining columns as location coordinates in a up to four-dimensional space. A and B must have the same number of columns.

e mdw separates the weights from the location matrices but is otherwise identical to emd.

e md r uses the original EMD implementation by Yossi Rubner from Stanford. In case A and B are not densities, the weighted sum of flows is normalized by the smaller total mass of the two. The version of the emd package released on CRAN contains only this implementation and all other functions are just front-ends for the call to emd r.

Value

Earth Mover’s Distance between of the distributions A and B. If A and B are not distributions then A is the source and B is the target.

Note

This is an open-source version of the package which contains only the implementation by Yossi Rubner.

Author(s)

R code by Simon Urbanek, EMD code by Yossi Rubner.
Examples

A <- matrix(1:6 / sum(1:6), 2)
B <- matrix(c(0, 0, 0, 0, 0, 1), 2)
emd2d(A, B)
# if we bring the rows closer, the distance will be reduced
# since mass from the first row has to move down
emd2d(A, B, 0.1)

# use Manhattan distance instead
emd2d(A, B, dist="manhattan")
# same, but using R-side closure
emd2d(A, B, dist=function(x, y) sum(abs(x - y)))

# the positions can overlap - this is a degenerate case of that
emd2d(A, B, rep(0, 3), rep(0, 2))
# and just a sanity check
emd2d(A, A) + emd2d(B, B)

# and the weight/location code should, hopefully have the same results
A. <- matrix(c(1:6 / sum(1:6), 1, 2, 1, 2, 1, 2, 3, 3), 6)
B. <- matrix(c(1, 2, 3), 1)
stopifnot(emd(A., B.) == emd2d(A, B))
stopifnot(emd(A., B.) == emdw(A.[,-1], A.[,1], B.[,-1,drop=FALSE], B.[,1]))
Index

*Topic multivariate
  emd, 1

emd, 1
emd2d (emd), 1
emdr (emd), 1
emd w (emd), 1