disp2D: R package for 2D Hausdorff and simplex dispersion orderings

G. Ayala[1], M.C. López-Díaz[2], M. López-Díaz[3] [*], L. Martínez-Costa[4]

[1]Universidad de Valencia. Dpto. de Estadística e Investigación Operativa. Av. Vicent Andrés Estellés, 1. E-46100 Burjasot, Valencia. Spain. Guillermo.Ayala@uv.es
[2]Universidad de Oviedo. Dpto. de Matemáticas. C/ Calvo Sotelo, s/n. E-33007 Oviedo. Spain. cld@uniovi.es
[3]Universidad de Oviedo. Dpto. de Estadística e I.O. y D.M. C/ Calvo Sotelo, s/n. E-33007 Oviedo. Spain. mld@uniovi.es
[4]Hospital Dr. Peset, Servicio de Oftalmología, Av. Gaspar Aguilar 90, E-46017 Valencia, Spain. martinez_luc@gva.es

Abstract

This document contains a description of the R package *disp2D* which is an implementation of the Hausdorff and simplex dispersion orderings in the 2D case. A complete description of the procedures implemented in this package can be found in Ayala et al. [2012]. Additional material related with the stochastic orderings considered here appears in López-Díaz [2006].

1 The introduction

First, we load the packages needed later. The package Barber et al. [2012] is required and the package Genz et al. [2012] is used in order to simulate data.

```
> library(disp2D)
> library(geometry)
> library(mvtnorm)
```

In order to illustrate the use of the package, we will use data generated with multivariate normal distribution.

Let us consider the $\mathbb{R}^d$-valued random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ with normal distributions, $\boldsymbol{X} \sim_{st} N(\mu_{\boldsymbol{X}}, \Sigma_{\boldsymbol{X}})$ and $\boldsymbol{Y} \sim_{st} N(\mu_{\boldsymbol{Y}}, \Sigma_{\boldsymbol{Y}})$, where $\Sigma_{\boldsymbol{X}} = AA^t$,

[*]corresponding author E-mail: mld@uniovi.es Phone: (34) 985 10 33 62 Fax: (34) 985 10 33 54

$A \in M_{d\times d}$ being a matrix whose values are randomly chosen with uniform distribution in the interval $(0,1)$, the super index ^{t} denoting the transpose matrix, and $\Sigma_{\boldsymbol{Y}} = \Sigma_{\boldsymbol{X}} + \lambda I_d$, with $\lambda \geq 0$. It is well-known that the eigenvalues of $\Sigma_{\boldsymbol{Y}}$ are those of $\Sigma_{\boldsymbol{X}}$ plus the value λ. Therefore, it holds that $\boldsymbol{X} \preceq_{sx} \boldsymbol{Y}$. Roughly speaking, larger values of λ will produce larger dispersion for the random vector $\boldsymbol{Y}$. In keeping with this framework, the following test is proposed

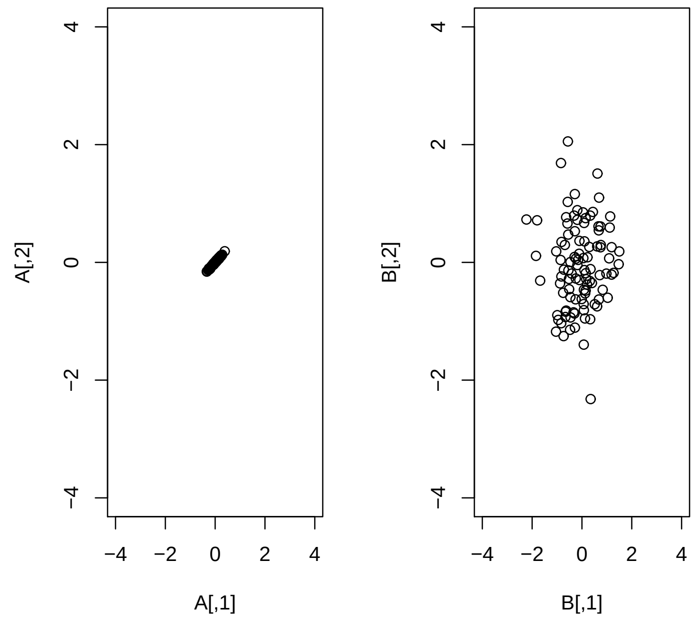
$$H_0 : \boldsymbol{X} \preceq_{sx} \boldsymbol{Y} \quad \text{against} \quad H_1 : \boldsymbol{X} \not\preceq_{sx} \boldsymbol{Y}.$$

Let us generate two point sets from the model just considered. The multivariate normal data are generated using the package *mvtnorm* Genz et al. [2012], Genz and Bretz [2009].

```
> d = 2
> n = 100
> mu1 = rep(0,d)
> mu2 = mu1
> lambda = .5
> n1=n2=n= 100
> sigma1 = matrix(runif(d*d),nrow=d,ncol=d)
> sigma1 = t(sigma1) %*% sigma1
> sigma2 = sigma1 + lambda * diag(1,d)
> A = rmvnorm(n1,mean=mu1,sigma=sigma1)
> B = rmvnorm(n2,mean=mu2,sigma=sigma2)
```

Let us see the datasets

```
> par(mfrow=c(1,2))
> plot(A,xlim=c(-4,4),ylim=c(-4,4))
> plot(B,xlim=c(-4,4),ylim=c(-4,4))
> par(mfrow=c(1,1))
```



2 Exact algorithm for the 2D Hausdorff dispersion ordering

The exact version of the Hausdorff dispersion ordering can be evaluated using the following code.

```
> r = .1
> prob = rep(1/n,n)
> HA = exactHausdorff(A,prob,r)
> HB = exactHausdorff(B,prob,r)
```

The following plot displays the two cumulative distribution functions.

```
> plot(ecdf(HA$distance))
> lines(ecdf(HB$distance),lty=2)
```

Finally, we can test the usual stochastic ordering for the empirical distribution of the just evaluated distances by using the Kolmogorov-Smirnov test.

```
> ks.test(HA$distance,HB$distance,alternative="greater")

        Two-sample Kolmogorov-Smirnov test

data:  HA$distance and HB$distance
D^+ = 0.9226, p-value < 2.2e-16
alternative hypothesis: the CDF of x lies above that of y
```

Other possibilities could be the Wilcoxon test for paired samples.

3 Simplex dispersion ordering

A detailed explanation of this algorithm can be found in Ayala et al. [2012].

If $A = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_1}\}$ and $B = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{n_2}\}$, let $\{i_1, \ldots, i_{d+1}, i_{d+2}, \ldots, i_{2(d+1)}\}$ be a sample without replacement from $\{1, \ldots, n_1\}$, and $U = d_H(\mathrm{co}(\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_{d+1}}), \mathrm{co}(\boldsymbol{x}_{i_{d+2}}, \ldots, \boldsymbol{x}_{i_{2(d+1)}}))$. Therefore, s_1 independent extractions from the set $\{1, \ldots, n_1\}$ will produce a random sample of the corresponding bootstrap distribution $u_1, \ldots, u_{s_1}$. Replacing $\boldsymbol{x}$ by $\boldsymbol{y}$, we obtain $v_1, \ldots, v_{s_2}$, a random sample of the bootstrap distribution associated to the vector $\boldsymbol{y}$. Now, these values can be used for the proposed tests.

The function *simplex* provides us a sample of u's.

```
> d1 = simplex(A,bootstrap=TRUE,nresamples=100)
```

If we consider two different samples we can test the simplex dispersion ordering using

```
> dhx = simplex(A,bootstrap=TRUE,nresamples=10)
> dhy = simplex(B,bootstrap=TRUE,nresamples=10)
> ks.test(dhx,dhy,alternative="greater")

        Two-sample Kolmogorov-Smirnov test

data:  dhx and dhy
D^+ = 1, p-value = 4.54e-05
alternative hypothesis: the CDF of x lies above that of y

> ks.test(dhx,dhy,alternative="less")

        Two-sample Kolmogorov-Smirnov test

data:  dhx and dhy
D^- = 0, p-value = 1
alternative hypothesis: the CDF of x lies below that of y
```

References

G. Ayala, M.C. López-Díaz, M., M. López-Díaz, and L. Martínez-Costa. Methods and algorithms to test the simplex and hausdorff dispersion orders with a simulation study and an ophthalmological application. Technical report, Universidad de Oviedo, 2012.

C. B. Barber, Kai Habel, Raoul Grasman, Robert B. Gramacy, Andreas Stahel, and David C. Sterratt. *geometry: Mesh generation and surface tesselation*, 2012. URL `http://CRAN.R-project.org/package=geometry`. R package version 0.3-0.

Alan Genz and Frank Bretz. *Computation of Multivariate Normal and t Probabilities*, volume 195 of *Lecture Notes in Statistics*. Springer-Verlag, Heidelberg, 2009. ISBN 978-3-642-01688-2.

Alan Genz, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl, and Torsten Hothorn. *mvtnorm: Multivariate Normal and t Distributions. R package version 0.9-9992.*, 2012. URL `http://CRAN.R-project.org/package=mvtnorm`.

Miguel López-Díaz. An indexed multivariate dispersion ordering based on the Hausdorff distance. *Journal of Multivariate Analysis*, 97(7): 1623 – 1637, 2006. ISSN 0047-259X. doi: DOI:10.1016/j.jmva.2006.02.004. URL `http://www.sciencedirect.com/science/article/B6WK9-4JK4P9F-1/2/4da718a7f427528dcc70d68b38f33901`.