# Deux applications de la distance de Wasserstein à la statistique

Rencontres Poitiers-Bordeaux : ASMSA

---

M. Hallin, G. Mordant & J. Segers

ULB, UCLouvain & UCLouvain

ArXiv: 2003.06684

10 décembre 2020

# Table of contents

# Introduction

# Introduction

Fundamental questions of statistics and modelling :

- Does the observed data come from a specific (family of) model(s)?
- How dependent are two blocks of a random vector ?

# p-Wasserstein distance

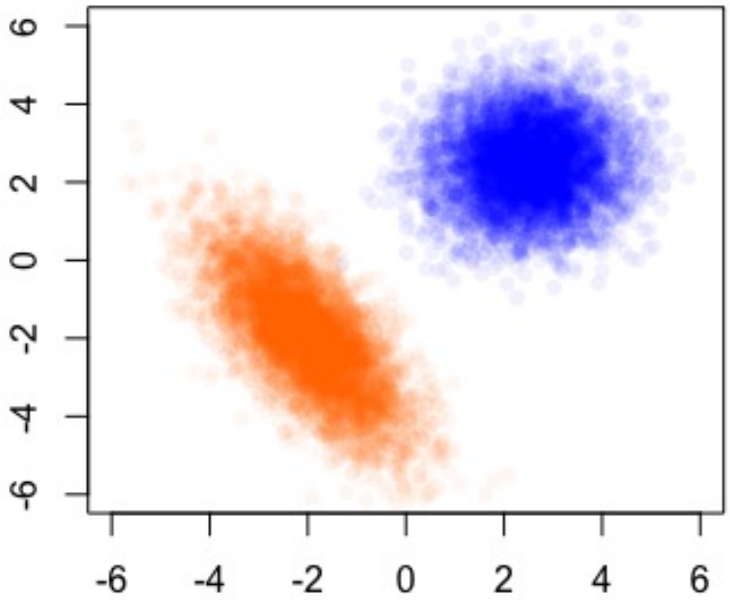
Define $\Gamma(\mathrm{P}, \mathrm{Q})$, the set of probability measures with marginals $\mathrm{P}$ and $\mathrm{Q}$ and $\mathcal{P}_p(\mathbb{R}^d)$, the set of Borel probability measures on $\mathbb{R}^d$ with finite $p$-th moment.

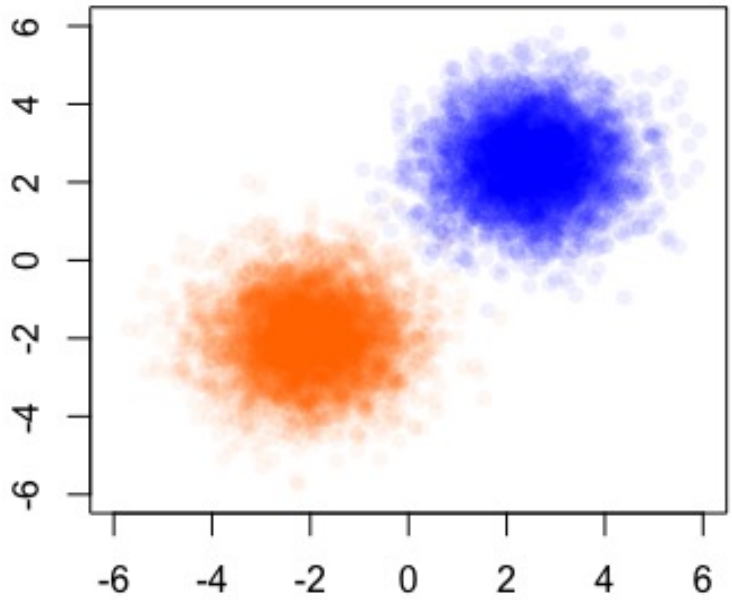
**Definition** (Villani, 2008)

The p-Wasserstein distance between $\mathrm{P}$ and $\mathrm{Q} \in \mathcal{P}_p(\mathbb{R}^d)$, is defined as

$$W_p^p(\mathrm{P}, \mathrm{Q}) = \inf_{\gamma \in \Gamma(\mathrm{P}, \mathrm{Q})} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p \mathrm{d}\gamma(x, y),$$

# Wasserstein distance



# Goodness-of-fit tests

## Meta-approach

- Find a metric on a space of probability distributions
- Construct an empirical version of this metric
- Find its distribution under the null hypothesis to perform the test

# Formal definition of the test

Consider an i.i.d. sample $\{X_i\}_{i=1}^n$ and its empirical measure $\hat{\mathrm{P}}_n$.

**Hypothesis**

$$\mathcal{H}_0^n : X_1 \sim \mathrm{P}_0 \quad \text{vs.} \quad \mathcal{H}_1^n : X_1 \sim \mathrm{P} \neq \mathrm{P}_0.$$

The test statistic based on the empirical Wasserstein distance is

$$T_{n,p} := W_p^p(\hat{\mathrm{P}}_n, \mathrm{P}_0).$$

The test has the form

$$\phi_{\mathrm{P}_0}^n = \begin{cases} 1 & \text{if } T_{n,p} > c(\alpha, n, \mathrm{P}_0, p) \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Formally, $c(\alpha, n, \mathrm{P}_0, p)$ is defined as

$$c(\alpha, n, \mathrm{P}_0, p) := \inf_c \{c : \mathrm{P}_0^{\otimes n}[T_{n,p} > c] \leq \alpha\}.$$

# Properties

**Lemma**

The test based on $T_{n,p}$ is consistent against fixed alternatives.

From Varadarajan's theorem (1958)

$$\hat{\mathrm{P}}_n \to_w \mathrm{P}_0 \quad \text{a.s.}$$

This combined with convergence of moments implies convergence of the Wasserstein distance. Thus,

$$W_p^p(\hat{\mathrm{P}}_n, \mathrm{P}_0) = o_{\mathrm{P}_0}(1), \quad n \to \infty.$$

Otherwise, if $\mathrm{P} \neq \mathrm{P}_0$

$$W_p^p(\hat{\mathrm{P}}_n, \mathrm{P}_0) \to W_p^p(\mathrm{P}, \mathrm{P}_0) > 0, \quad n \to \infty.$$

# More general alternatives

**Hypothesis**

$$\mathcal{H}_0^n : X_1 \sim \mathrm{P} \in \mathcal{M} \quad \text{vs.} \quad \mathcal{H}_1^n : X_1 \sim \mathrm{P} \notin \mathcal{M},$$

where $\mathcal{M} := \{\mathrm{P}_\theta : \theta \in \Theta\}$, with $\Theta$ a metric space.

For general parametric families, we suggest using

$$T^*_{n,p} := W_p^p(\hat{\mathrm{P}}_n, \mathrm{P}_{\hat{\theta}_n}).$$

The test will then have the form

$$\phi^n_{\mathcal{M}} = \begin{cases} 1 & \text{if } T^*_{n,p} > c^*(\alpha, n, \hat{\theta}_n, p) \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

# Consistency for parametric families

Consider

(a) $\mathcal{K}(\Theta)$ the collection of compact subsets of $\Theta$;

(b) the map $\Theta \to \mathcal{P}_p^d : \theta \mapsto \mathrm{P}_\theta$, one-to-one and $W_p$-continuous;

(c) an estimator $\hat{\theta}_n$ weakly consistent locally uniformly in $\theta \in \Theta$.

**Properties**

(i) $T_{n,p}^* \to 0$ in $\mathcal{P}_\theta^n$-probability uniformly in $\theta \in \Theta$.

(ii) For every $\mathrm{P} \in \mathcal{P}^d \setminus \mathcal{M}$ : $\exists K \in \mathcal{K}(\Theta)$ with $\mathrm{P}^n[\hat{\theta}_n \in K] \to 1$

$$\lim_{n\to\infty} \mathrm{P}^n[\phi_{\mathcal{M}}^n = 1] = 1.$$

# Computations

# Semi-discrete setting

One can compute $W_p^p(\hat{\mathrm{P}}_n, \mathrm{Q})$, relying on a dual formulation.

# Semi-discrete setting

One can compute $W_p^p(\hat{\mathrm{P}}_n, \mathrm{Q})$, relying on a dual formulation.

The computation is equivalent to solving

$$\sup_{\psi \in \mathbb{R}^n} \left\{ F(\psi) := \frac{1}{n} \sum_j \psi_j + \int_{V_\psi(j)} \left( \|x - X_j\|^p - \psi_j \right) \, \mathrm{dQ} \right\}$$

where, for $p = 2$, $V_\psi(j)$ are Laguerre cells

$$V_\psi(j) := \left\{ x : \|x - X_j\|^2 - \psi_j \leq \|x - X_i\|^2 - \psi_i, \forall i \right\}.$$

# Semi-discrete setting

One can compute $W_p^p(\hat{\mathrm{P}}_n, \mathrm{Q})$, relying on a dual formulation.

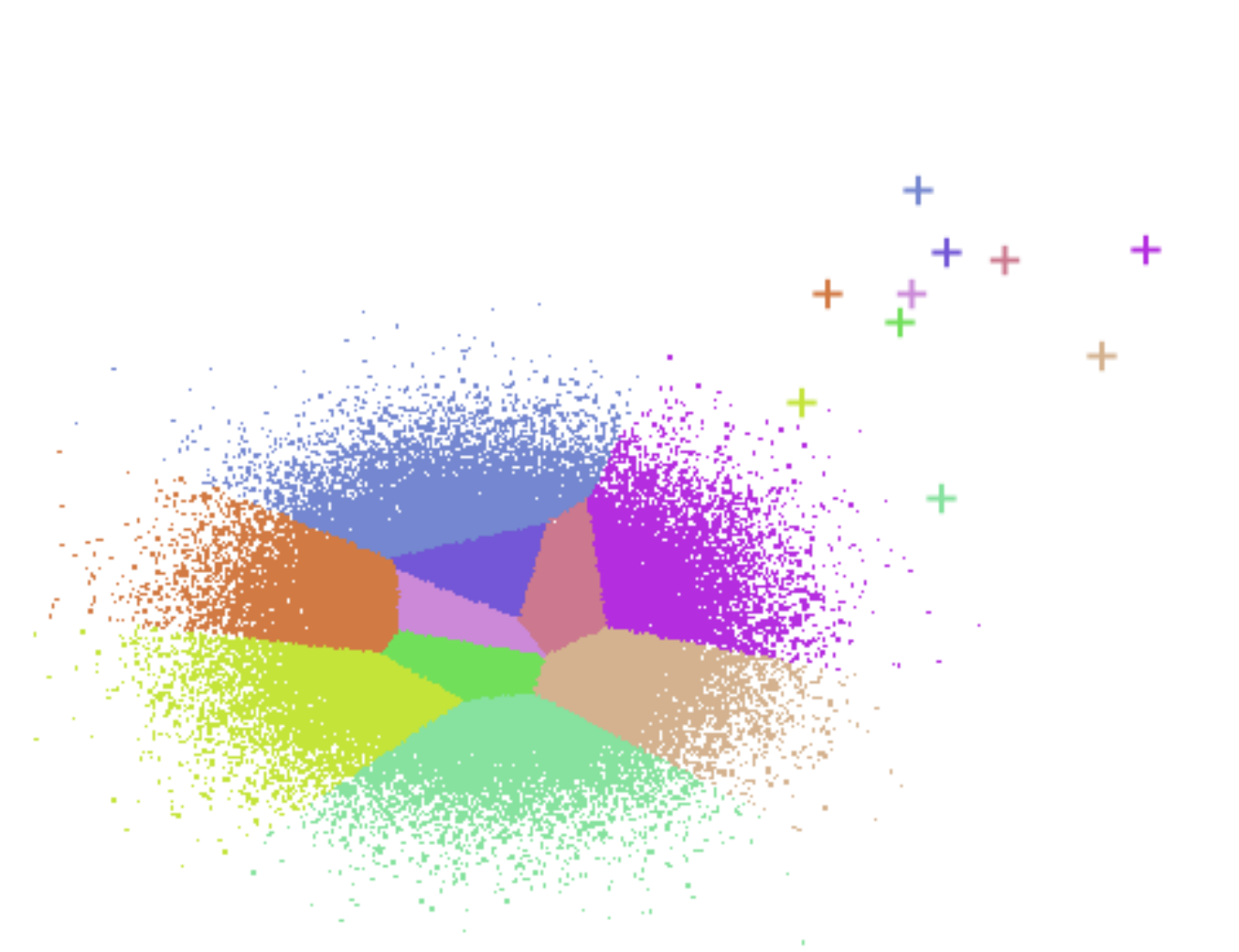
The computation is equivalent to solving

$$\sup_{\psi \in \mathbb{R}^n} \left\{ F(\psi) := \frac{1}{n} \sum_j \psi_j + \int_{V_\psi(j)} \left( \|x - X_j\|^p - \psi_j \right) \, \mathrm{dQ} \right\}$$

where, for $p = 2$, $V_\psi(j)$ are Laguerre cells

$$V_\psi(j) := \left\{ x : \|x - X_j\|^2 - \psi_j \leq \|x - X_i\|^2 - \psi_i, \forall i \right\}.$$

Algorithms by Mérigot (2016), Genevay et al. (2016), Hartmann and Schumacher (2017), ...



# Simulation results

# Comparisons

- Rippl–Munk–Sturm (2016)
  For $X \sim \mathcal{N}(m_1, \Sigma_1)$ and $Y \sim \mathcal{N}(m_2, \Sigma_2)$,

$$W_2^2(X, Y) = \|m_1 - m_2\|^2 + \operatorname{tr}\left[\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right].$$

  Test based on plug-ins of $\hat{m}_1$ and $\widehat{\Sigma}_1$

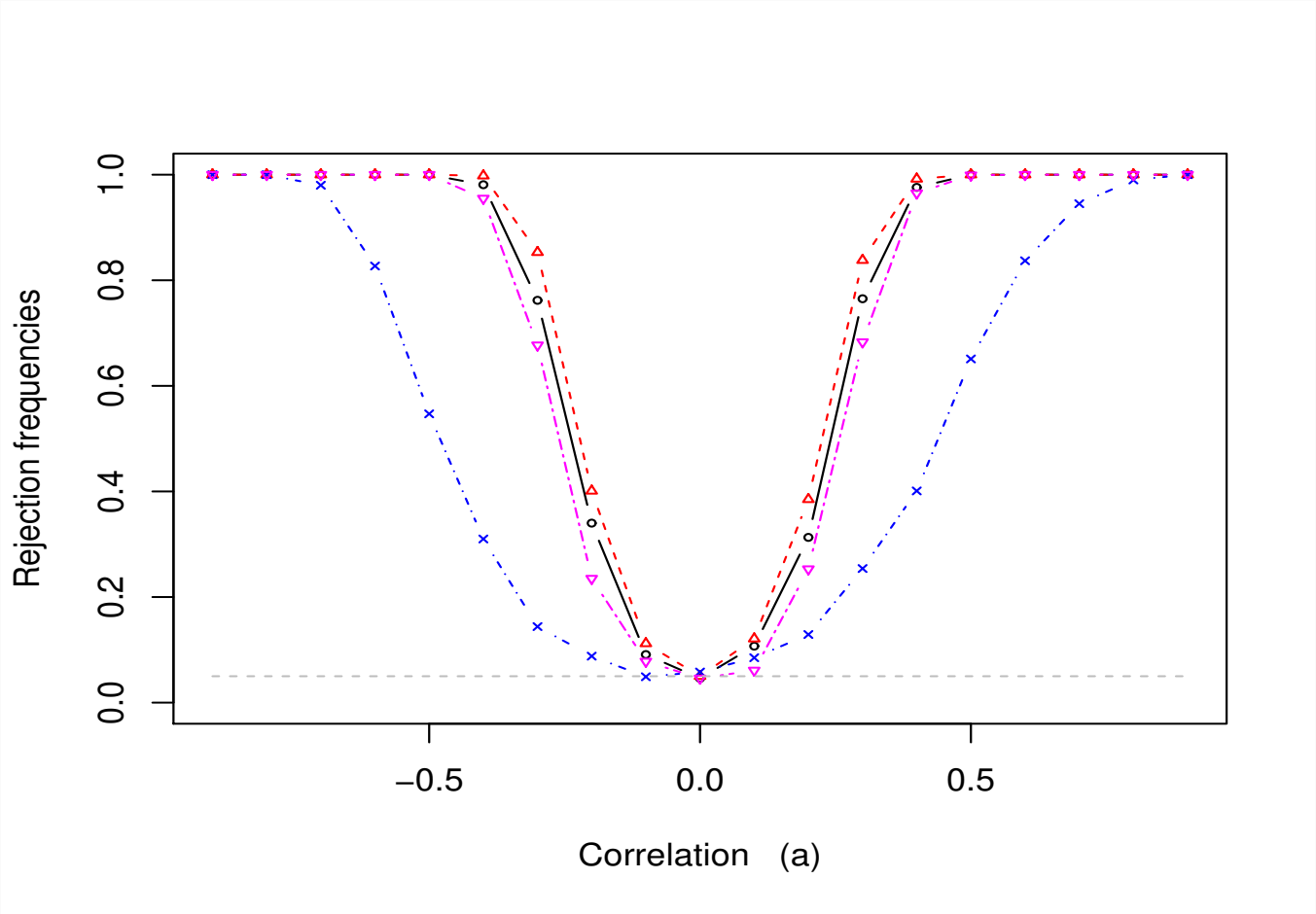
# Comparisons

- Rippl–Munk–Sturm (2016)
  For $X \sim \mathcal{N}(m_1, \Sigma_1)$ and $Y \sim \mathcal{N}(m_2, \Sigma_2)$,

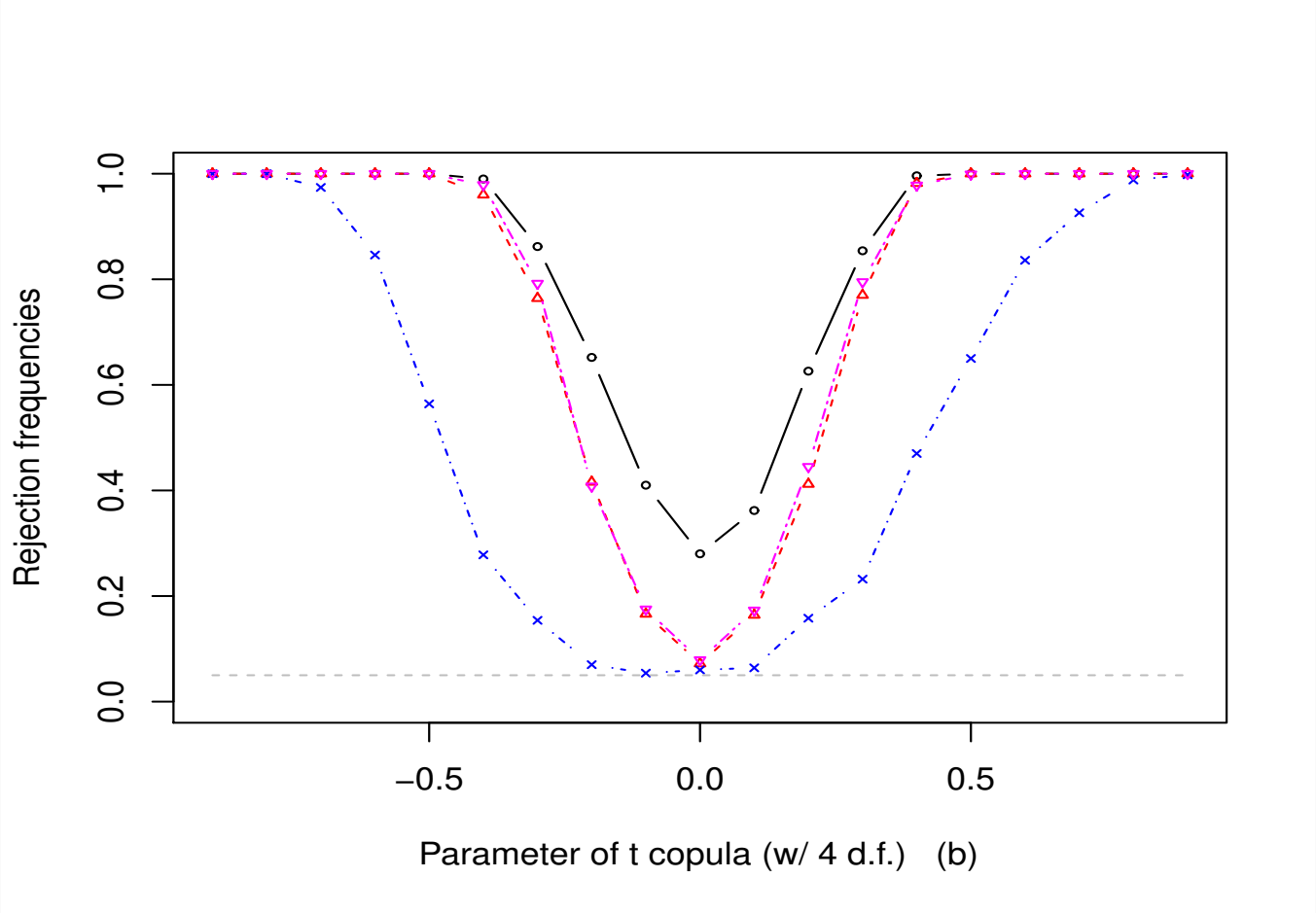
  $$W_2^2(X, Y) = \|m_1 - m_2\|^2 + \mathrm{tr}\left[\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right].$$

  Test based on plug-ins of $\hat{m}_1$ and $\widehat{\Sigma}_1$
- Khmaladze (2016)
  Test based on $\sup_{x \in \mathbb{R}^d} |v_G(x)|$ where $v_G(x)$ is a transformed Brownian bridge.

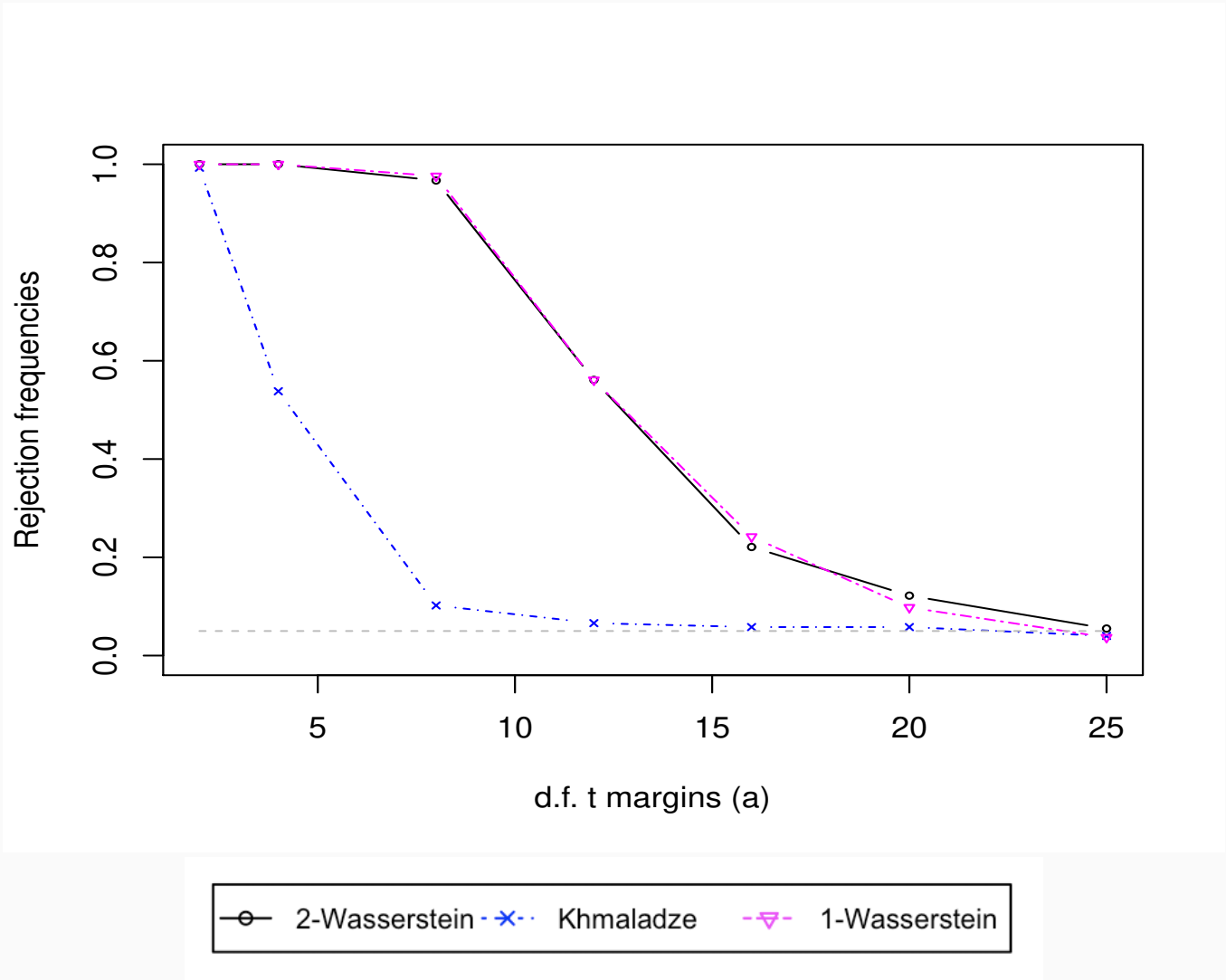
# $\mathcal{N}(0, I_2)$ null hypothesis



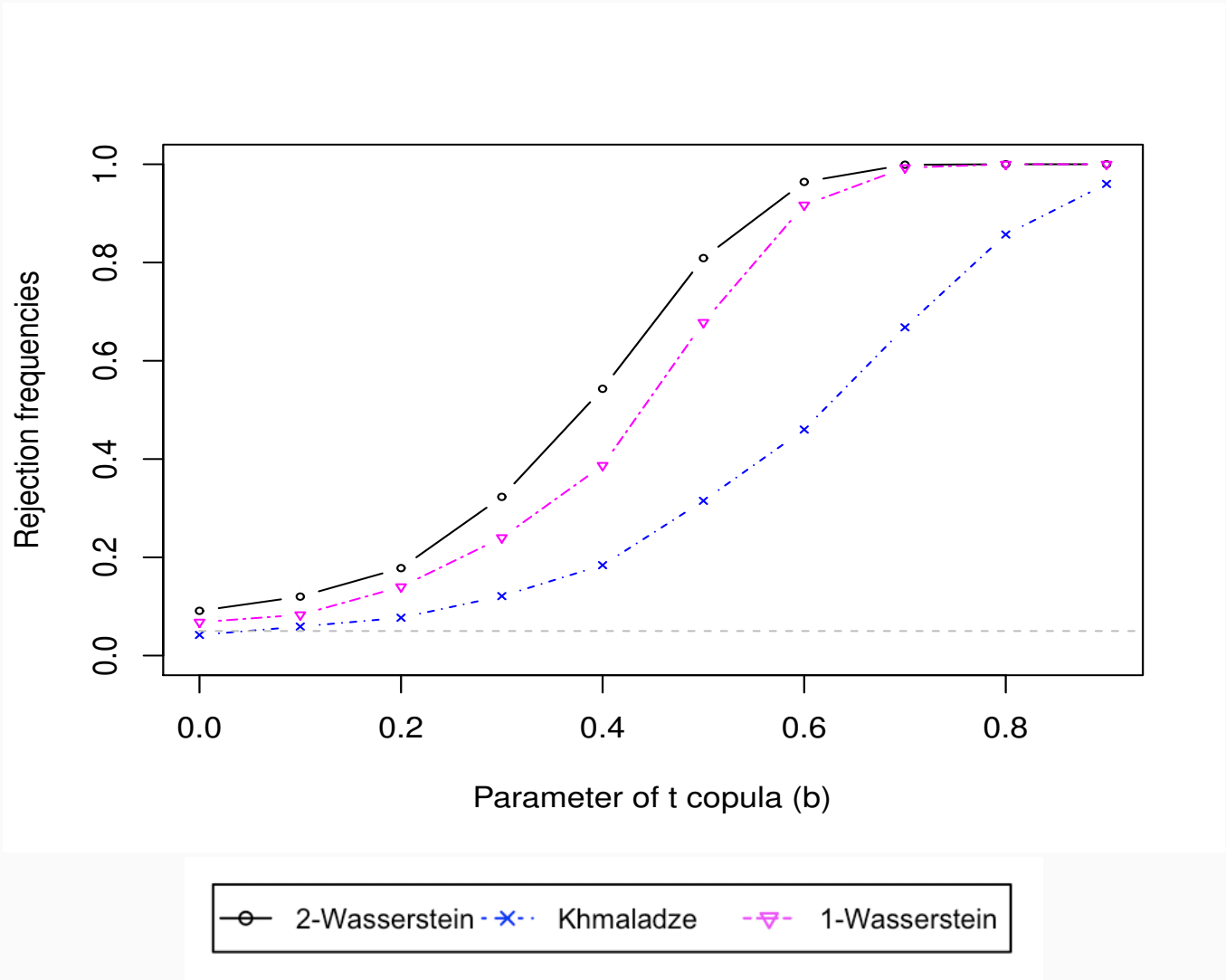
# $\mathcal{N}(0, I_2)$ null hypothesis



# $t_{25}^{\otimes 5}$ null hypothesis



# $t_{25}^{\otimes 5}$ null hypothesis



# Location-scatter

Consider $\hat{m}$ and $\hat{S}$ an estimator of the location and the scatter matrix of the distribution, respectively. Define

$$\tilde{\mathrm{P}}_n := n^{-1} \sum_{i=1}^{n} \delta_{\hat{S}^{-1/2}(X_i - \hat{m})},$$

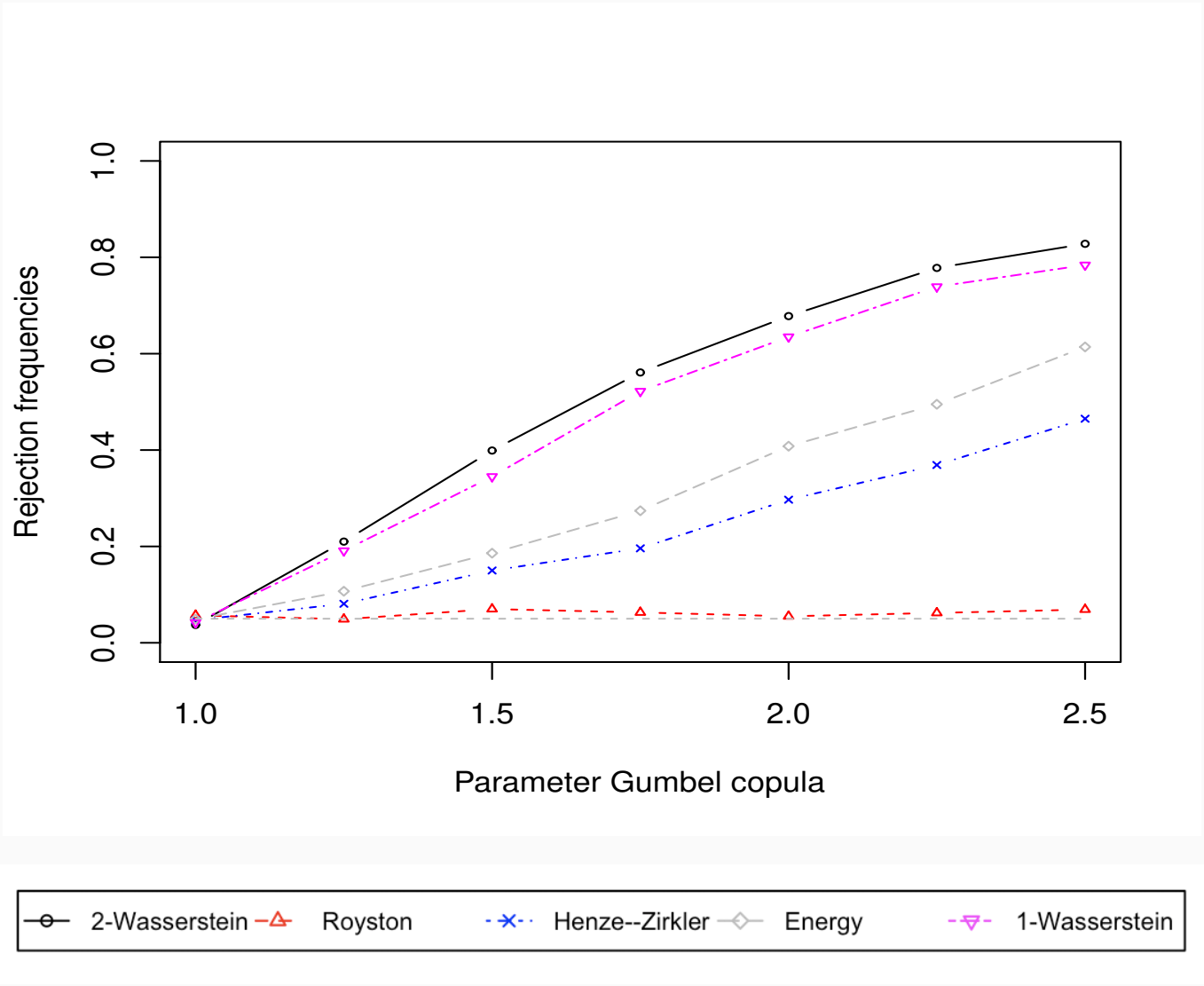
where $\hat{S}^{-1/2}$ is the Cholesky square-root. The test statistic based on the empirical Wasserstein distance is

$$\tilde{T}_{n,p} := W_p^p(\tilde{\mathrm{P}}_n, \mathrm{P}_0).$$

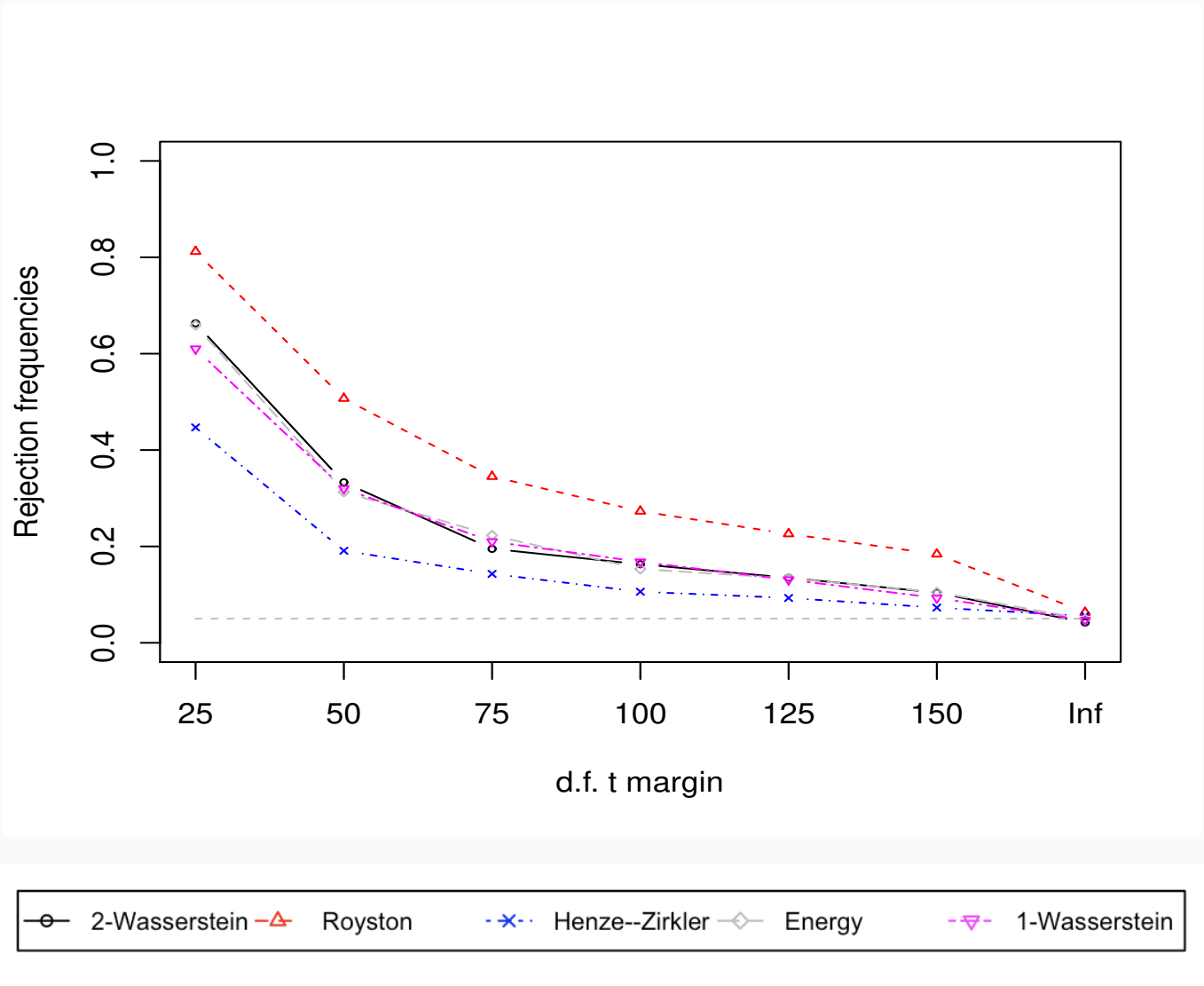
The test has the form

$$\phi_{\mathrm{P}_0}^n = \begin{cases} 1 & \text{if } \tilde{T}_{n,p} > \tilde{c}(\alpha, n, \mathrm{P}_0, p) \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

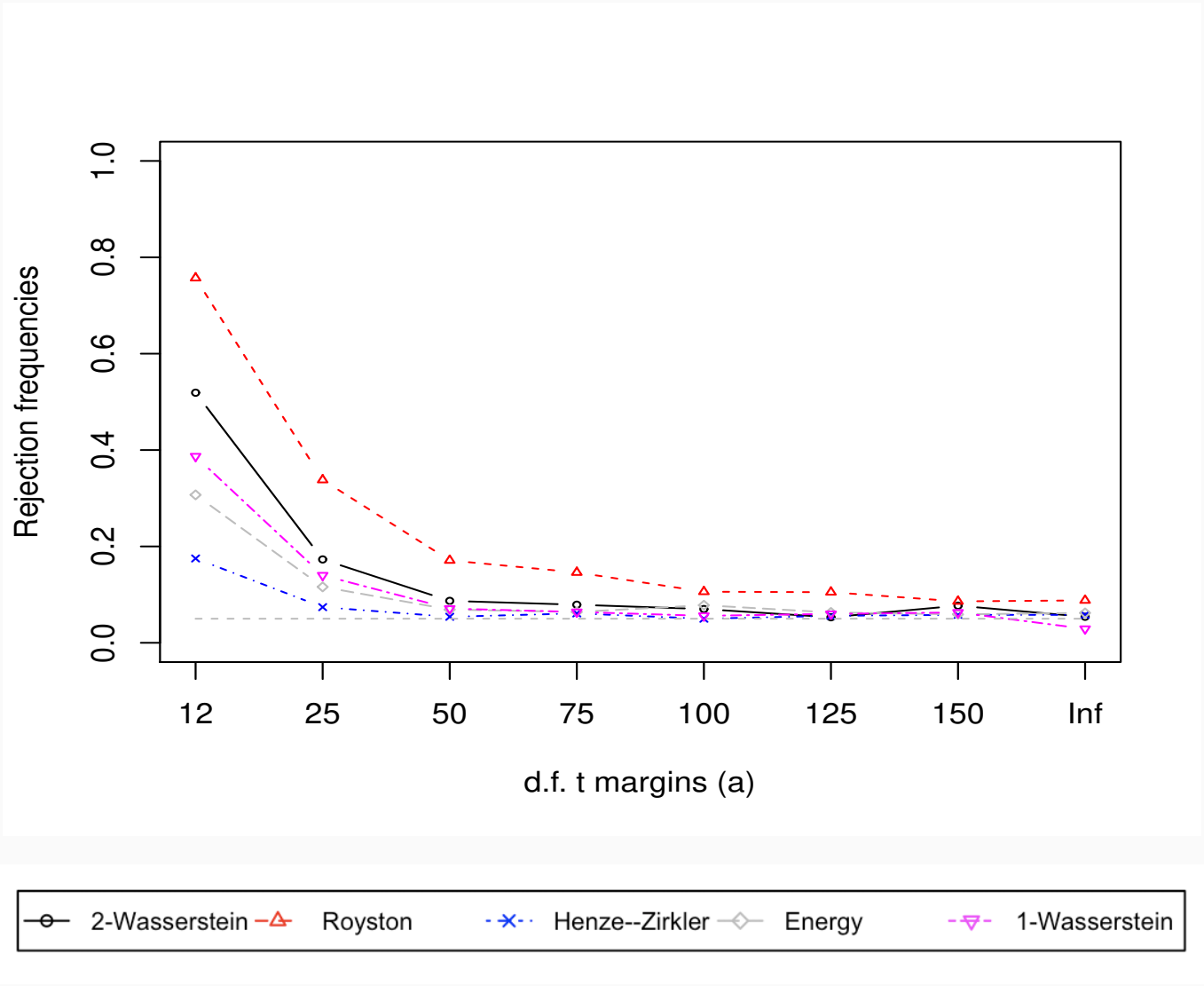
# Bivariate Gaussian family



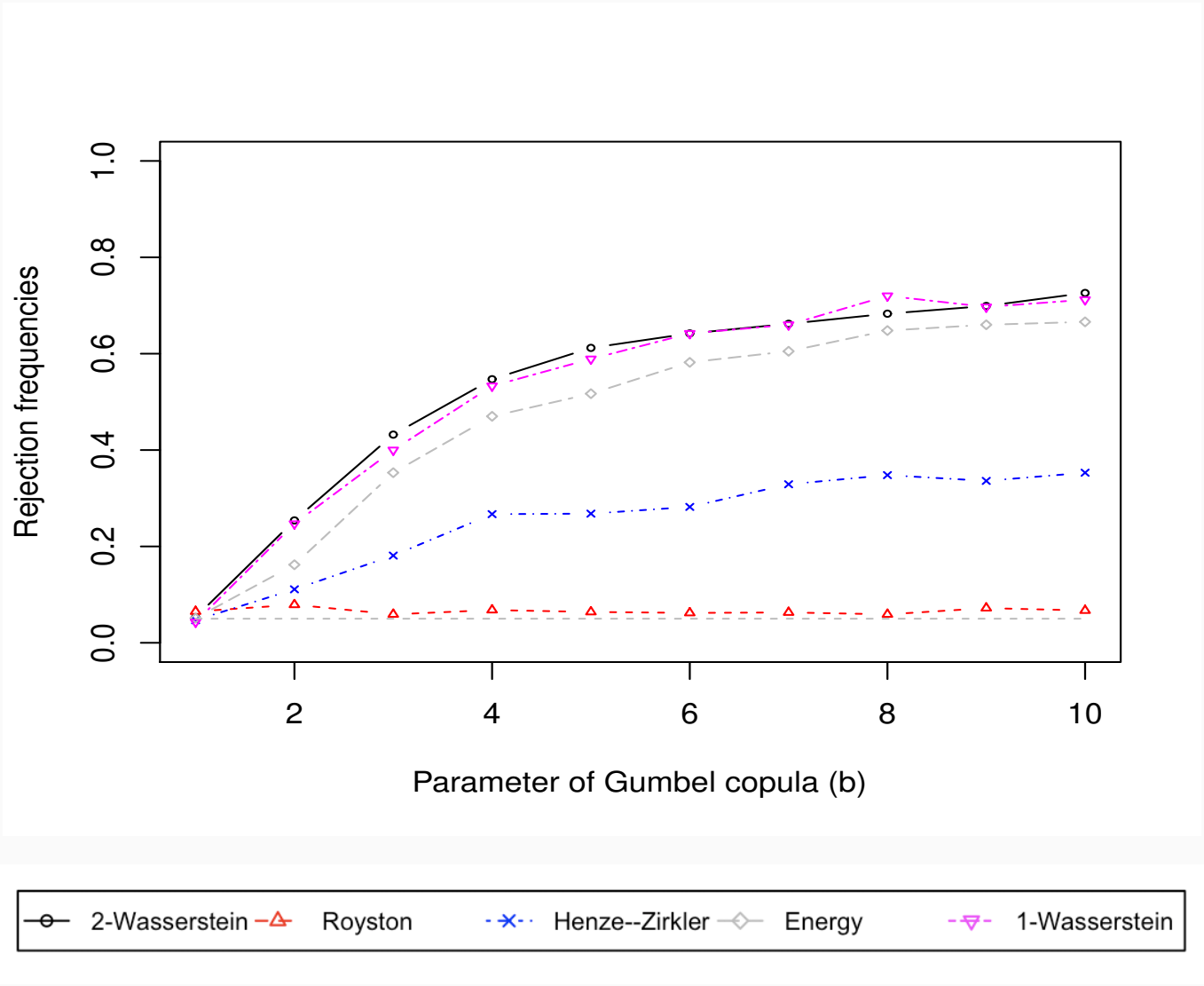
# Bivariate Gaussian family



# Pentavariate Gaussian family



# Pentavariate Gaussian family



# Measuring dependence

# Objective

Given a random vector $(X_1, \ldots, X_d)$, consider two complementary subsets of $\{1, \ldots, d\}$, $D_1$ and $D_2$.

What is the dependence between $\left(X_j\right)_{j \in D_1}$ and $\left(X_k\right)_{k \in D_2}$ ?

# G-copulas

Given a random vector $(X_1, \ldots, X_d)$ from a measure $\mu$ with continuous marginal cumulative distribution functions $F_j(x)$ for $1 \leq j \leq d$, define the G-copula as

$$(\Phi^{-1} \circ F_1(X_1), \ldots, \Phi^{-1} \circ F_d(X_d)),$$

where $\Phi$ is the gaussian c.d.f. We denote the law of the G-copula of $\mu$ by $G_\mu$.

# New dependence measures

Set, $\gamma_d = \mathcal{N}(0, I_d)$.

# New dependence measures

Set, $\gamma_d = \mathcal{N}(0, I_d)$.

$$\mathfrak{D}_1(\mu, D_1, D_2) := \frac{W_2^2(G_\mu, \gamma_d) - W_2^2(G_{\mu,D_1} \otimes G_{\mu,D_2}, \gamma_d)}{\sup_{\nu \in \Gamma(G_{\mu,D_1}, G_{\mu,D_2})} W_2^2(\nu, \gamma_d) - W_2^2(G_{\mu,D_1} \otimes G_{\mu,D_2}, \gamma_d)}.$$

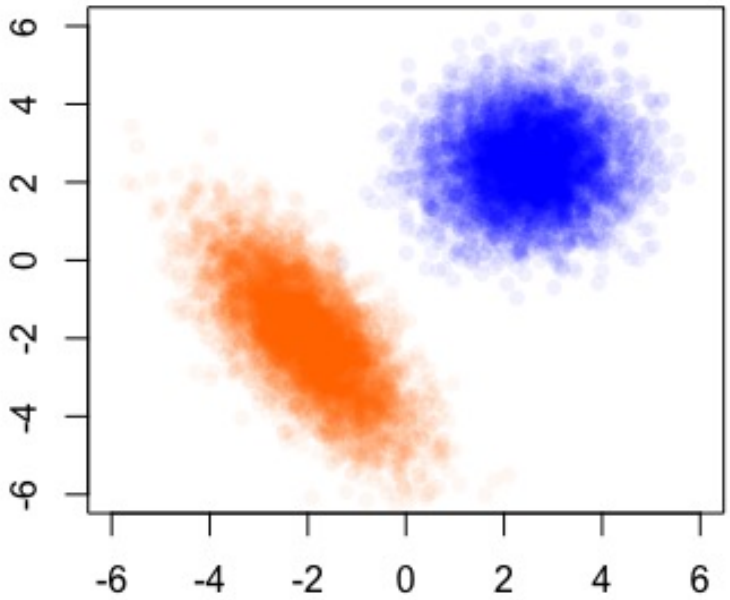
# New dependence measures

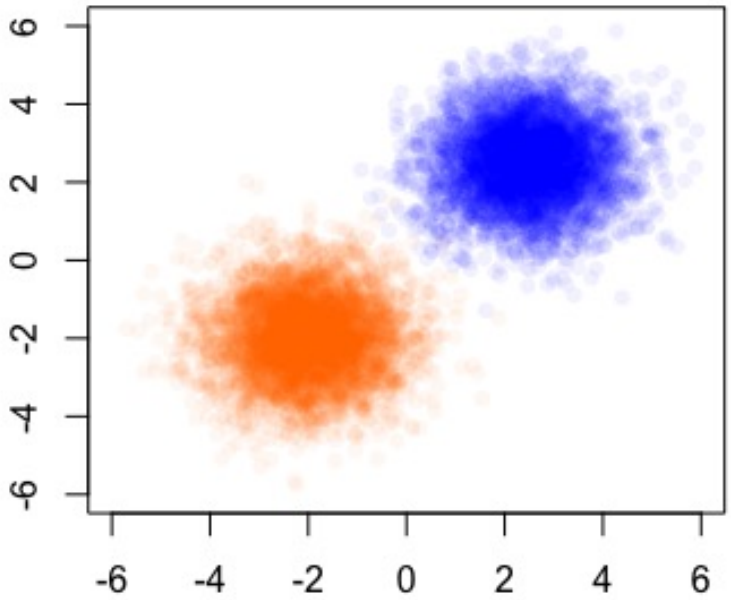
Set, $\gamma_d = \mathcal{N}(0, I_d)$.

$$\mathfrak{D}_1(\mu, D_1, D_2) := \frac{W_2^2(G_\mu, \gamma_d) - W_2^2(G_{\mu,D_1} \otimes G_{\mu,D_2}, \gamma_d)}{\sup_{\nu \in \Gamma(G_{\mu,D_1}, G_{\mu,D_2})} W_2^2(\nu, \gamma_d) - W_2^2(G_{\mu,D_1} \otimes G_{\mu,D_2}, \gamma_d)}.$$

$$\mathfrak{D}_2(\mu, D_1, D_2) := \frac{W_2^2(G_\mu, G_{\mu,D_1} \otimes G_{\mu,D_2})}{\sup_{\kappa \in \Gamma(G_{\mu,D_1}, G_{\mu,D_2})} W_2^2(\kappa, G_{\mu,D_1} \otimes G_{\mu,D_2})}.$$

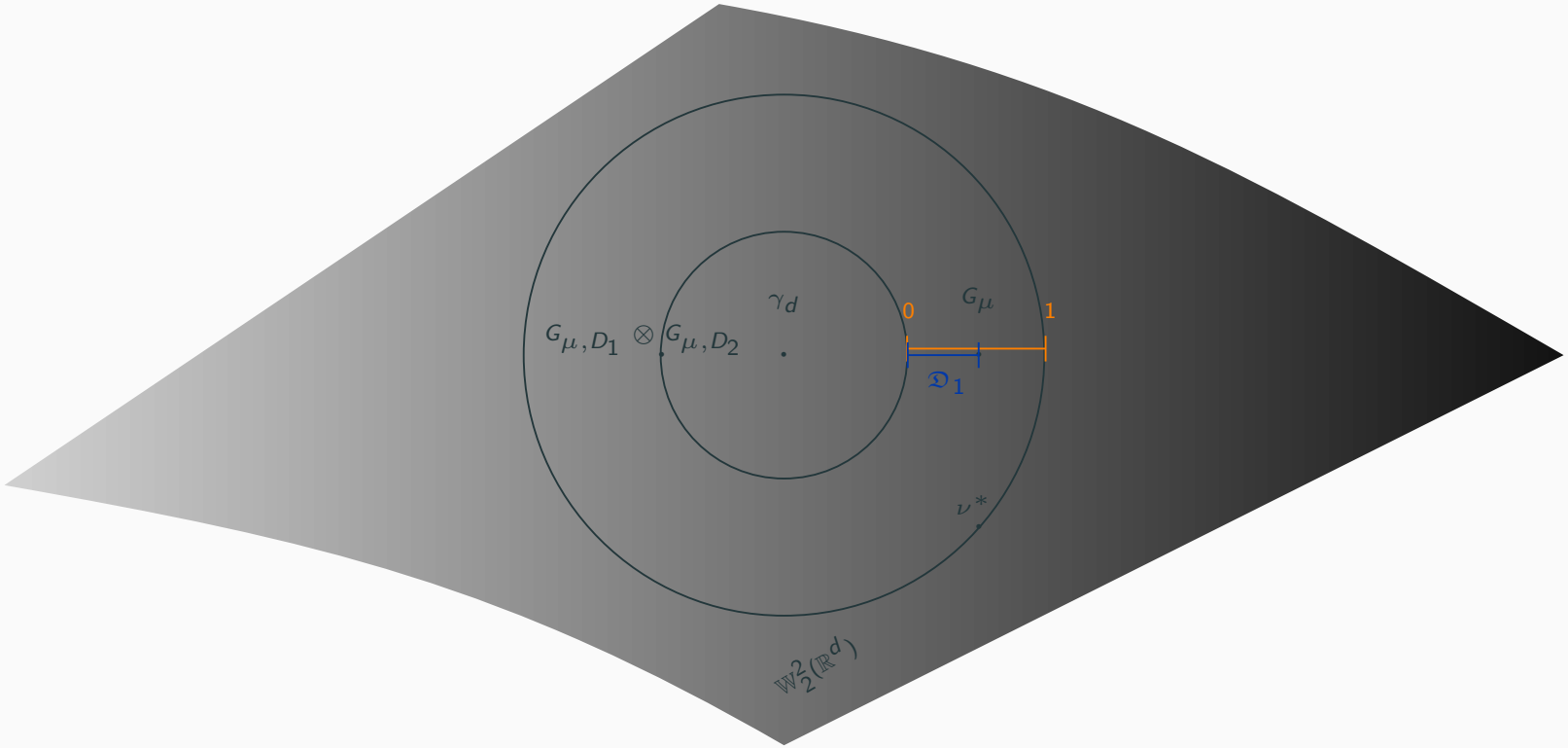
# Wasserstein distance (rappel)



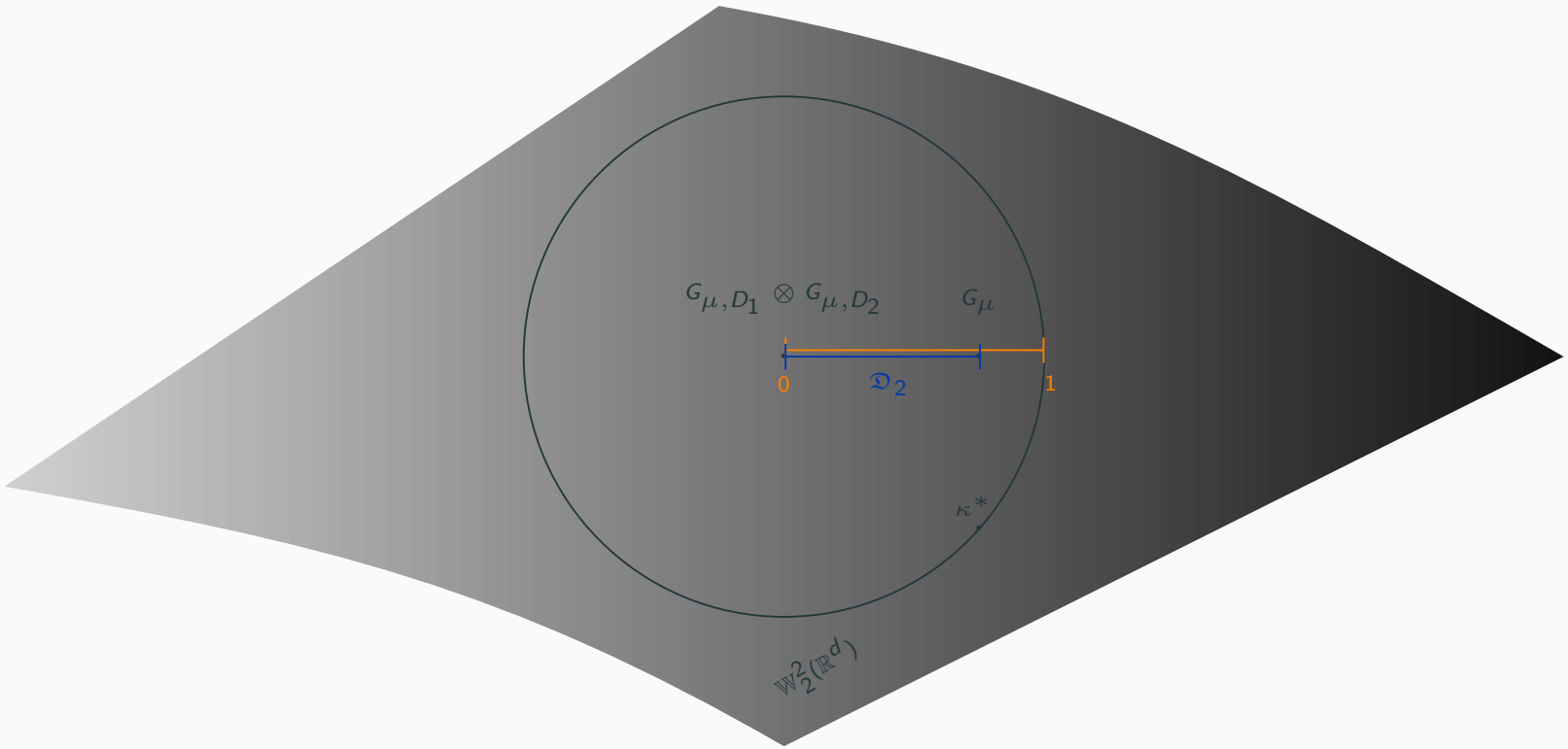



# Representation and properties

# Representation - $\mathfrak{D}_1$



# Representation - $\mathfrak{D}_2$



# Properties

- Invariance to increasing transformations of the margins.
- $0 \leq \mathfrak{D}(\mu, D_1, D_2) \leq 1$.
- $\mathfrak{D}(\mu, D_1, D_2) = 0$ if and only if $\mu = \mu_{D_1} \otimes \mu_{D_2}$.
- If $d_1 = d_2 = 1$, both measures are equal and equal to

$$\frac{W_2^2(G_\mu, \gamma_2)}{W_2^2(\gamma_{\text{co}}, \gamma_2)} = \frac{W_2^2(G_\mu, \gamma_2)}{2\left(2 - \sqrt{2}\right)}.$$

# Gaussian copulas

# Bivariate Gaussian copula

For a bivariate Gaussian with correlation coefficient $\rho$,

$$\mathfrak{D} = \frac{2 - \sqrt{1+\rho} - \sqrt{1-\rho}}{2 - \sqrt{2}}.$$

# Gaussian copula - $\mathfrak{D}_1$

$$\Sigma = \begin{bmatrix} \Sigma_1 & F \\ F^\top & \Sigma_2 \end{bmatrix}$$

For a MVN with correlation matrix $\Sigma$,

$$\mathfrak{D}_1 = \frac{\operatorname{tr}\left[\Sigma_1^{1/2}\right] + \operatorname{tr}\left[\Sigma_2^{1/2}\right] - \operatorname{tr}\left[\Sigma^{1/2}\right]}{\operatorname{tr}\left[\Sigma_1^{1/2}\right] + \operatorname{tr}\left[\Sigma_2^{1/2}\right] - \min_\Psi \operatorname{tr}\left[\Sigma_\Psi^{1/2}\right]},$$

with

$$\Sigma_\Psi = \begin{bmatrix} \Sigma_1 & \Psi \\ \Psi^\top & \Sigma_2 \end{bmatrix}.$$

# Gaussian copula - $\mathfrak{D}_2$

For a MVN with correlation matrix $\Sigma$,

$$\mathfrak{D}_2 = \frac{d - \text{tr}\left[\left(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2}\right)^{1/2}\right]}{d - \min_\psi \text{tr}\left[\left(\Sigma_0^{1/2}\Sigma_\psi\Sigma_0^{1/2}\right)^{1/2}\right]}.$$

# Gaussian copula - Dealing with denominators

Set $\Sigma_j = O_j \Lambda_j O_j^\top$ with $O_j$ orthogonal and $\Lambda_j$ diagonal.

$$\Sigma_\psi = \begin{bmatrix} O_1 & 0 \\ 0 & O_2 \end{bmatrix} \begin{bmatrix} \Lambda_1 & K \\ K^\top & \Lambda_2 \end{bmatrix} \begin{bmatrix} O_1^\top & 0 \\ 0 & O_2^\top \end{bmatrix}.$$

**Theorem**

*To minimise* $\mathrm{tr}\left[\Sigma_{\psi(K)}^{1/2}\right]$, *take*

$$K = \left(\Lambda_1^{1/2}(\Lambda_2^{1/2})_{d_1 \times d_1} \quad 0_{d_1 \times (d_2 - d_1)}\right).$$

# A link with entropy, K-L and more

This construct

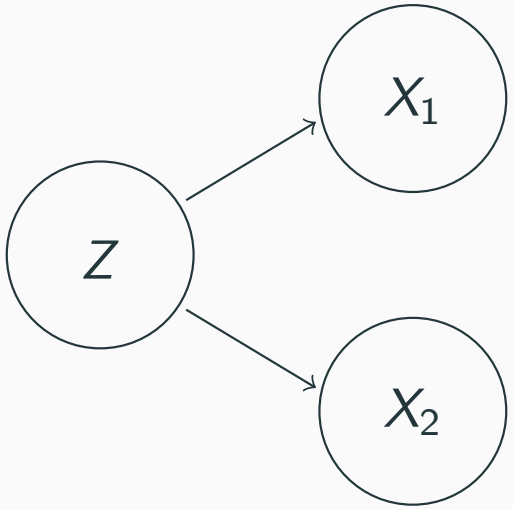
- maximizes $\mathrm{tr}(\Psi^\top \Psi)$ and therefore maximizes the RV coefficient;
- maximizes $W_2(\mathcal{N}_d(0, \Sigma_\Psi), \mathcal{N}_d(0, \Sigma_0))$;
- maximizes $W_2(\mathcal{N}_d(0, \Sigma_\Psi), \mathcal{N}_d(0, I_d))$;

# A link with entropy, K-L and more

This construct

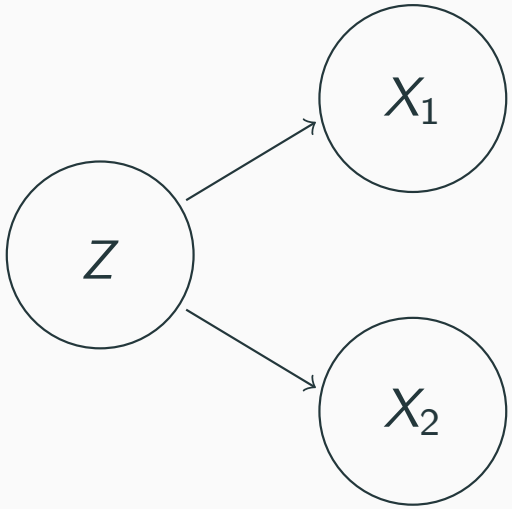
- maximizes $\mathrm{tr}(\Psi^\top \Psi)$ and therefore maximizes the RV coefficient;
- maximizes $W_2(\mathcal{N}_d(0, \Sigma_\Psi), \mathcal{N}_d(0, \Sigma_0))$;
- maximizes $W_2(\mathcal{N}_d(0, \Sigma_\Psi), \mathcal{N}_d(0, I_d))$;
- minimizes $\mathrm{Ent}(X)$ for $X \sim \mathcal{N}_d(0, \Sigma_\Psi)$;
- maximizes $D_{\mathrm{KL}}(\mathcal{N}_d(0, \Sigma_\Psi) \,\|\, \mathcal{N}_d(0, \Sigma_0))$;
- maximizes $D_{\mathrm{KL}}(\mathcal{N}_d(0, \Sigma_\Psi) \,\|\, \mathcal{N}_d(0, I_d))$;
- minimises the operator entropy, $-\sum_{i \leq d} \lambda_i \log(\lambda_i)$.

# Perfect dependence for Gaussian copula




Dimension of $Z$ is $\max(d_1, d_2)$

# Perfect dependence for Gaussian copula




Dimension of $Z$ is $\max(d_1, d_2)$

In the perfect dependence case, the correlation matrix's principal components of the first subvector are perfectly correlated with those of the second subvector, according to the sizes of the corresponding eigenvalues.

# Semi-parametric correlation matrix

Given an i.i.d. $d$-dimensional random sample from $\mu$, $(X_i)_{i \leq n}$, let us define the Gaussian *pseudo-observations*

$$\hat{Z}_i := (\Phi^{-1} \circ \tfrac{n}{n+1} \hat{F}_{1,n}(X_{i,1}), \cdots, \Phi^{-1} \circ \tfrac{n}{n+1} \hat{F}_{d,n}(X_{i,d})),$$

where $\hat{F}_{j,n}$ is the empirical distribution function of the $j$th margin.

# Semi-parametric correlation matrix

Given an i.i.d. $d$-dimensional random sample from $\mu$, $(X_i)_{i \leq n}$, let us define the Gaussian *pseudo-observations*

$$\hat{Z}_i := (\Phi^{-1} \circ \tfrac{n}{n+1} \hat{F}_{1,n}(X_{i,1}), \cdots, \Phi^{-1} \circ \tfrac{n}{n+1} \hat{F}_{d,n}(X_{i,d})),$$

where $\hat{F}_{j,n}$ is the empirical distribution function of the $j$th margin.

$$\hat{\Sigma}_n := c_n \sum_{i=1}^{n} \hat{Z}_i \hat{Z}_i^\top,$$

where $c_n^{-1} = \sum_{i=1}^{n} [\Phi^{-1}(i/(n+1))]^2$.

# (Semi-)parametric estimator

$$\hat{\mathfrak{D}}_1 := \frac{\mathrm{tr}[\hat{\Sigma}_1^{1/2}] + \mathrm{tr}[\hat{\Sigma}_2^{1/2}] - \mathrm{tr}[\hat{\Sigma}^{1/2}]}{\mathrm{tr}[\hat{\Sigma}_1^{1/2}] + \mathrm{tr}[\hat{\Sigma}_2^{1/2}] - \mathrm{tr}[\hat{\Sigma}_{\psi^*}^{1/2}]}$$

# (Semi-)parametric estimator

$$\hat{\mathfrak{D}}_1 := \frac{\mathrm{tr}[\hat{\Sigma}_1^{1/2}] + \mathrm{tr}[\hat{\Sigma}_2^{1/2}] - \mathrm{tr}[\hat{\Sigma}^{1/2}]}{\mathrm{tr}[\hat{\Sigma}_1^{1/2}] + \mathrm{tr}[\hat{\Sigma}_2^{1/2}] - \mathrm{tr}[\hat{\Sigma}_{\psi^*}^{1/2}]}$$

$$\hat{\mathfrak{D}}_2 := \frac{d - \mathrm{tr}\left[\left(\hat{\Sigma}_0^{1/2}\hat{\Sigma}\hat{\Sigma}_0^{1/2}\right)^{1/2}\right]}{d - \mathrm{tr}\left[\left(\hat{\Sigma}_0^{1/2}\hat{\Sigma}_{\tilde{\psi}}\hat{\Sigma}_0^{1/2}\right)^{1/2}\right]}.$$

# Inference for (Semi-)parametric estimator

$$\hat{\mathfrak{D}}_1 := \frac{\mathrm{tr}[\hat{\Sigma}_1^{1/2}] + \mathrm{tr}[\hat{\Sigma}_2^{1/2}] - \mathrm{tr}[\hat{\Sigma}^{1/2}]}{\mathrm{tr}[\hat{\Sigma}_1^{1/2}] + \mathrm{tr}[\hat{\Sigma}_2^{1/2}] - \mathrm{tr}[\hat{\Sigma}_{\psi^*}^{1/2}]}$$

**Theorem**

*If the eigenvalues of $\Sigma_1$ and $\Sigma_2$ are distinct and $\Sigma$ is positive definite,*

$$\sqrt{n}(\hat{\mathfrak{D}}_j - \mathfrak{D}_j) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_j^2(\Sigma)),$$

*where $\sigma_j^2(\Sigma) \geq 0$ is some function of the correlation matrix $\Sigma$ and $j = 1, 2$.*

# To go further

- A fully nonparametric estimator exists.
- Provably convergent.
- Optimisation on the manifold of bistochastic matrices needed for the denominator.

# Summary

- Wasserstein distance measure cost to “reshape” distributions
- Useful for GoF
- Helps measure dependence
- Nice performance and properties

# Questions ?

# Selected references


- del Barrio, E., Cuesta-Albertos, J. A., Hallin, M., & Matrán, C. (2018). Center-Outward Distribution Functions, Quantiles, Ranks, and Signs in $\mathbb{R}^d$. arXiv preprint arXiv:1806.01238.
- Henze, N., & Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. Communications in Statistics-Theory and Methods, 19(10), 3595-3617.
- Khmaladze, E. (2016). Unitary transformations, empirical processes and distribution free testing. Bernoulli, 22(1), 563-588.
- Mérigot, Q. (2011, August). A multiscale approach to optimal transport. In Computer Graphics Forum (Vol. 30, No. 5, pp. 1583-1592). Oxford, UK: Blackwell Publishing Ltd.
- Ramdas, A., Trillos, N., & Cuturi, M. (2017). On wasserstein two-sample testing and related families of nonparametric tests. Entropy, 19(2), 47.
- Rippl, T., Munk, A., & Sturm, A. (2016). Limit laws of the empirical Wasserstein distance: Gaussian distributions. Journal of Multivariate Analysis, 151, 90-109.
- Rizzo, M. L., & Székely, G. J. (2016). Energy distance. Wiley Interdisciplinary Reviews: Computational Statistics, 8(1), 27-38.
- Royston, J. P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. Journal of the Royal Statistical Society: Series C (Applied Statistics), 31(2), 115-124.
- Varadarajan, V. S. (1958). On the convergence of sample probability distributions. Sankhyā: The Indian Journal of Statistics (1933-1960), 19(1/2), 23-26.

# In a nutshell

**Goal**

Propose and investigate Goodness-of-Fit tests based on the Wasserstein distance $W_p$.

- Simple null

$$T_{n,p} := W_p^p(\hat{P}_n, P_0)$$

- Parametric family

$$T_{n,p}^* := W_p^p(\hat{P}_n, P_{\hat{\theta}_n})$$

$$W_p^p(\mu, \nu) := \min_{X \sim \nu, Y \sim \mu} \mathbb{E}\|X - Y\|^p$$

**Key points**

- Tests proved consistent against fixed alternatives
- Good performance in various simulation settings

# Bootstrap

Conditions for $H_n(\hat{\theta}_n) \xrightarrow{\mathcal{L}} H_n(\theta_0)$

1. $\forall \theta_n \to \theta_0, H_n(\theta_n) \xrightarrow{\mathcal{L}} H_n(\theta_0)$
2. $\hat{\theta}_n \to \theta_0$ in $\mathrm{P}_{\theta_0,n}$- probability.

Currently hindered by lack of distributional results.