Scaling Optimal Transport to High-Dimensional Gaussian Distributions

Charles Bouveyron, Marco Corneli

Scaling Optimal Transport to High-Dimensional Gaussian Distributions

Charles Bouveyron & Marco Corneli
Université Côte d'Azur, INRIA, CNRS, Maasai, Nice, France

February 5, 2025

Abstract

Although optimal transport (OT) has recently become very popular in machine learning, it faces challenges when dealing with high-dimensional data, such as images or omics data. Current OT approaches for high-dimensional situations rely on projections of the data or measures onto low-dimensional spaces, which inevitably results in information loss. In this work, we consider the case of high-dimensional Gaussian distributions with parsimonious covariance structures and lower intrinsic dimension. We exhibit a simplified closed-form expression of the 2-Wasserstein distance with an efficient and robust calculation procedure based on a low-dimensional decomposition of empirical covariance matrices, without relying on data projections. Furthermore, we provide a closed-form expression for the Monge map, which involves the exact calculation of the square-root and inverse square-root of the source distribution covariance matrix. This approach offers analytical and computational advantages, as demonstrated by our numerical experiments, which quantitatively evaluate these benefits in comparison to existing methods. In addition to being able to compute both the W_2^2-distance and the transport map, our method outperforms model-free methods, in high dimension, even in the case of non-Gaussian distributions.

1 Introduction

Due to its proven versatility, optimal transport (OT) is becoming more and more popular within the machine learning community (Peyré et al., 2019). Basically, once the observed data is identified with a probability distribution (possibly the empirical mass function), optimal transport allows to consistently assess the similarity between complex instances such as point clouds, images or graphs. However, as the modern data are increasingly high-dimensional, OT is also now facing an old problem in optimization and statistical learning: the curse of dimensionality (Bellman, 1957). Among the OT problems that have to face the high dimensionality of the data, we can mention as a popular example the calculation of the Frechet inception distance (FID, Heusel et al., 2017) for comparing the distribution of generated images with the distribution of a set of ground-truth images, using the Wasserstein distance between two full Gaussian distributions.

1.1 Statistical learning in high-dimensional spaces

In many application domains of machine learning, such as image analysis, genomics, chemometrics or personalized medicine, the observed data are frequently high-dimensional and learning from such

data is a challenging problem. Indeed, statistical learning in such high-dimensional spaces is made difficult both because of estimation biases and numerical problems (Giraud, 2021; Wainwright, 2019). In particular, when considering the generative (model-based) framework, most learning methods show a disappointing behavior in high-dimensional spaces. They suffer from the well-known curse of dimensionality which is mainly due to the fact that generative methods turn to be dramatically over-parametrized in high-dimensional spaces (Bouveyron et al., 2019). Moreover, even though many variables are measured to describe the studied phenomenon, only a small subset of these original variables is in fact relevant for both modeling and learning. In recent years, several works tried to reduce the data dimensionality or select relevant variables while building a generative predictor, showing excellent results. In this context, there are two main approaches. On the one hand, some works assume that the data of each class live in different low-dimensional subspaces. On the other hand, some other works assume that the classes differ only with respect to some of the original features. Both approaches present two practical advantages: results are improved by the removing of non informative features and the result interpretation is eased by the visualization in the subspaces or the meaning of retained variables. We may recommend to refer to Bouveyron et al. (2019, Chap. 8) and Bouveyron and Brunet-Saumard (2014) for a full overview in the contexts of classification, clustering and dimension reduction. As we focus here on the question of an efficient modeling of high-dimensional distributions, a key work in this context is due to Tipping and Bishop (1999b) who have shown that the subspace of principal component analysis (PCA) could be retrieved from the maximum-likelihood estimator of a parameter, in a particular factor analysis model called probabilistic PCA (PPCA). This probabilistic framework led to diverse Bayesian analysis of PCA (Bishop, 1999; Minka, 2000) and extensions in various ML situations such as classification Bouveyron et al. (2007b) and clustering Tipping and Bishop (1999a); Bouveyron et al. (2007a); McNicholas and Murphy (2008). As it will be shown in this paper, this model will be once again a game-changer tool, here for the optimal transport between high-dimensional Gaussian distributions.

1.2 Optimal Transport with Wasserstein distance

Based on the modern formulation of Kantorovich (1942), standard optimal transport generally relies on the Wasserstein distance. Given two random variables X_1 and X_2 supported on $\mathbb{R}^p$, with finite second moments and whose marginal cumulative distribution functions are denoted by μ_1 and μ_2, respectively, the squared 2-Wasserstein distance is defined as:

$$W_2^2(\mu_1, \mu_2) := \min_{\pi \in \Pi(\mu_1, \mu_2)} \mathbb{E}_{(X_1, X_2) \sim \pi} ||X_1 - X_2||_2^2, \tag{1}$$

where $\Pi(\mu_1, \mu_2)$ denotes the set of *joint* distributions with marginals μ_1 and μ_2, respectively and $\| \cdot \|_2$ denotes the standard Euclidean norm. The joint distribution π^* minimizing the expectation on the r.h.s. of Eq. (1) is known as optimal coupling or optimal transport plan. As it can be understood from the above equation, OT lifts a metric defined on some ground space (here $\mathbb{R}^p$ with Euclidean metric) to a metric on the probability distributions supported on that space. The above definition extends to probability measures with support on more general separable metric spaces and higher order Wasserstein distances. However, in this paper we only focus on the 2-Wasserstein distance between measures supported on $\mathbb{R}^p$, for some integer p. For an in depth inspection of Wasserstein distances and their properties the reader is referred to Villani et al. (2009); Santambrogio (2015); Peyré et al. (2019).

In the particular where case the random variables X_1 and X_2 are Gaussian, it was shown that the Wasserstein distance can be computed in closed form (Dowson and Landau, 1982; Takatsu, 2011). Moreover, in force of the Brenier's theorem (see for instance Peyré et al., 2019, Theorem 2.1) there exists a unique transport or *Monge* map $T^* : \mathbb{R}^p \to \mathbb{R}^p$ linked to the optimal transport plan π^* by the following relation[1]

$$\mathbb{E}_{(X_1,X_2)\sim\pi^*}\left[h(X_1,X_2)\right] = \mathbb{E}_{X_1\sim\mu_1}\left[h(X_1,T^*(X_1))\right],$$

holding for any continuous function $h : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Also T^* has closed form in the Gaussian case. If the Gaussian distributions of X_1 and X_2 must be inferred from the data, i.e. two point clouds in dimensions $\mathbb{R}^p$, the closed formulas for the Wasserstein distance and the Monge map T^* between μ_1 and μ_2 can always been computed. However, due to the difficulties in the estimation of the covariance matrices in high dimension, those formulas lead to poor estimates (shown in Section 3). The alternative approach, seeking to compute the Wasserstein distance between μ_1 and μ_2 via the empirical distributions, is also doomed to failure due to the known instability of OT in high dimension (Dudley, 1969; Fournier and Guillin, 2015). Among the existing solutions, we can cite Sliced Wasserstein distances (SWD, see Nguyen and Ho, 2024, and the references therein), which attack the high-dimensional problem by averaging the optimal OT costs between 1D measures, obtained by projecting the original measures onto several random directions. Another approach builds Subspace Robust Wasserstein distances (SRW, Paty and Cuturi, 2019), which are defined by modifying the 2-Wasserstein cost in such a way to find an optimal matching between projections of the original measures onto a k-dimensional subspace. However, both SWD and SRW do not allow to estimate the Monge map T^*. Still based on the empirical distributions, Muzellec and Cuturi (2019) introduced in the literature two methods to extend a Monge map which is optimal on a subspace to one that is *nearly* optimal on the entire space. However, all those approaches are non-parametric, meaning that the Gaussianity of the input data is never used (nor the closed formulas mentioned above).

1.3 Contributions of the paper

This work focuses on the use of the high-dimensional Gaussian (HD-Gaussian) distributions, induced by the probabilistic PCA (PPCA) model, for the optimal transport between high-dimensional data distributions. In particular, this paper features the three main contributions:

(i) exhibition of a closed-form expression of the 2-Wasserstein distance between two HD-Gaussian distributions, with an efficient and robust calculation procedure based on a low-dimensional subspace decomposition.

(ii) generalization to a less restrictive framework of previous state-of-the-art results, which considered Gaussian distributions with similar covariance orientations or structures.

(iii) exhibition of a closed-form expression of the Monge map for the transport of a HD-Gaussian distribution on another one, involving an exact calculation of both the square-root and the inverse square-root of the covariance matrix of the source distribution, avoiding in turn many numerical drawbacks in high-dimensional practical situations.

[1] In a short-hand notation one writes $\pi^* = (\mathrm{Id}, T^*)_\sharp \mu_1$, where $\sharp$ is the push-forward operator.

Interestingly, these results remain valid in the case of HD-Gaussian distributions with different intrinsic dimensions. It is also worth underlying that the proposed approach, named hereafter OT-HDGauss, is able to compute both the W_2^2-distance and the transport map for HD-Gaussian distributions. Furthermore, the analytical and numerical advantages of our approach in high dimensions allow it to outperform model-free methods in the case of non-Gaussian distributions. These contributions are supported by numerical experiments that highlight the performance and robustness of the proposed OT-HDGauss method to both the dimensionality and the sample size, and this in comparison with the most recent OT approaches.

2 Optimal Transport between HD-Gaussian Distributions

2.1 The HD-Gaussian distribution

To overcome the well-known "curse of the dimensionality" in statistical learning, Tipping and Bishop (1999b) have proposed a parsimonious Gaussian distribution, induced by a probabilistic view of PCA, that splits the modelling between a low-dimensional subspace where the data actually live and a noise component. This HD-Gaussian distribution can be defined as follows.

Definition 2.1. A p-dimensional random vector $X \in \mathbb{R}^p$ follows a HD-Gaussian distribution $\mathcal{N}_{HD}(m, U, \Lambda, \sigma^2, d)$ if it exists a low-dimensional latent random vector $Y \in \mathbb{R}^d$, of intrinsic dimensionality $d < p$, and a p-dimensional noise random vector $\varepsilon \in \mathbb{R}^p$ such that:

$$\begin{aligned} \mathrm{X} &= UY + m + \varepsilon, \\ Y &\sim \mathcal{N}(0, \Lambda), \\ \varepsilon &\sim \mathcal{N}(0, \sigma^2 \mathrm{I}_p), \end{aligned}$$

where U is a $p \times d$ transformation matrix whose columns are orthonormal vectors, $m \in \mathbb{R}^p$ is the mean vector, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ and $\sigma^2 > 0$.

Under these assumptions, it can be shown that the HD-Gaussian distribution $\mathcal{N}_{HD}(m, U, \Lambda, \sigma^2, d)$ is a specific Gaussian distribution with a structured covariance matrix.

Proposition 2.2. *A p-dimensional random vector $X \in \mathbb{R}^p$ following a HD-Gaussian distribution $\mathcal{N}_{HD}(m, U, \Lambda, \sigma^2, d)$ is distributed as:*

$$X \sim \mathcal{N}(m, Q\Delta Q^t),$$

where $Q = [U, R]$, the $p \times p$ matrix made of U and an orthonormal complementary R, and Δ is a block-diagonal matrix:

$$\Delta = \begin{pmatrix} \boxed{\begin{matrix} \delta_1 & & 0 \\ & \ddots & \\ 0 & & \delta_d \end{matrix}} & & \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} d \\ & 0 & \\ & \boxed{\begin{matrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{matrix}} & \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} (p-d) \end{pmatrix}$$

with $\delta_j = \lambda_j + \sigma^2$ and $\delta_j > \sigma^2$, for $j = 1, ..., d$.

Proof. Assuming that $\mathrm{X} = UY + m + \varepsilon$, where $Y \sim \mathcal{N}(0, \Lambda)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_p)$, the conditional distribution of Y is therefore Gaussian:

$$X \mid \mathrm{Y} \sim \mathcal{N}(U\mathrm{Y} + m, \sigma^2 \mathrm{I}_p),$$

and the marginal distribution of X is a Gaussian distribution with a specific structured covariance structure:

$$X \sim \mathcal{N}(m, \Sigma),$$

where $\Sigma = U\Lambda U^t + \sigma^2 \mathrm{I}_p$. Introducing $Q = [U, R]$, the $p \times p$ matrix made of U and an orthonormal complementary R, the covariance matrix Σ can be easily rewritten $\Sigma = Q\Delta Q^t$ where $\Delta = \mathrm{diag}\left(\delta_1, ..., \delta_d, \sigma^2, ..., , for \text{j=1,...,d}. This allow stoconclude.\right.$ $\square$

Therefore, the HD-Gaussian distribution is fully parametrized by the set of parameters $\theta = \{m, U, \lambda_j, \sigma^2, d; \forall j = 1, ..., d\}$.

2.2 Calculation of the 2-Wasserstein distance

Let us now consider two HD-Gaussian probability distributions $\mu_1 \sim \mathcal{N}_{HD}(m_1, U_1, \Lambda_1, \sigma_1^2, d_1)$ and $\mu_2 \sim \mathcal{N}_{HD}(m_2, U_2, \Lambda_2, \sigma_2^2, d_2)$ on $\mathbb{R}^p$ for which we would like to compute the 2-Wasserstein distance. The following proposition exhibits a closed-form expression of $W_2(\mu_1, \mu_2)$, which in turn yields to numerically efficient and stable calculations.

Proposition 2.3. *The 2-Wasserstein distance between two HD-Gaussian distributions* $\mu_1 \sim \mathcal{N}_{HD}(m_1, U_1, \Lambda_1, \sigma_1^2, d_1)$ *and* $\mu_2 \sim \mathcal{N}_{HD}(m_2, U_2, \Lambda_2, \sigma_2^2, d_2)$ *is*

$$\begin{aligned} W_2^2(\mu_1, \mu_2) = \|m_1 - m_2\|_2^2 + \mathrm{trace}(\Lambda_1) + \mathrm{trace}(\Lambda_2) \\ + p(\sigma_1^2 + \sigma_2^2) - 2\mathrm{trace}(A^{\frac{1}{2}}), \end{aligned}$$

where A *can be expressed as:*

$$A = U_1\Lambda_1 U_1^t U_2 \Lambda_2 U_2^t + \sigma_1^2 U_2 \Lambda_2 U_2^t + \sigma_2^2 U_1 \Lambda_1 U_1^t + \sigma_1^2 \sigma_2^2 I_p.$$

Proof. In the case when $c(x, y) = \|x - y\|_2^2$, the 2-Wasserstein distance between two Gaussian distributions $\mu_1 \sim \mathcal{N}(m_1, \Sigma_1)$ and $\mu_2 \sim \mathcal{N}(m_1, \Sigma_1)$, is known to have the following explicit form (Dowson and Landau, 1982; Takatsu, 2011)

$$\begin{aligned} W_2(\mu_1, \mu_2)^2 = \|m_1 - m_2\|_2^2 + \mathrm{trace}(\Sigma_1) + \mathrm{trace}(\Sigma_2) \\ - 2\mathrm{trace}\left[\left(\Sigma_1^{\frac{1}{2}} \Sigma_2 \Sigma_1^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]. \end{aligned} \tag{2}$$

Thanks to Proposition 2.2, this result can be extended to two HD-Gaussian distributions $\mu_1 \sim \mathcal{N}_{HD}(m_1, U_1, \Lambda_1, \sigma_1^2, d_1)$ and $\mu_2 \sim \mathcal{N}_{HD}(m_2, U_2, \Lambda_2, \sigma_2^2, d_2)$, by considering that Σ_1 and Σ_2 have specific parsimonious structures, i.e. $\Sigma_i = U_i \Lambda_i U_i^t + \sigma_i^2 I_p$, with $\Lambda_i = \mathrm{diag}(\delta_{i1} - \sigma_i^2, ..., \delta_{id} - \sigma_i^2)$, for $i = 1, 2$. It is first straightforward to establish that $\mathrm{trace}(\Sigma_1) = \mathrm{trace}(\Lambda_1) + p\sigma_1^2$, and similarly for Σ_2.

Let's now consider the computation of $\mathrm{trace}\left[\left(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]$. Reminding that trace of matrix M is equal to the sum of its eigenvalues $\omega_1(M), ..., \omega_p(M)$, we can write:

$$\begin{aligned}\mathrm{trace}\left[\left(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] &= \sum_{j=1}^{p}\omega_j\left[\left(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \\ &= \sum_{j=1}^{p}\omega_j^{\frac{1}{2}}\left[\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}}\right] \\ &= \sum_{j=1}^{p}\omega_j^{\frac{1}{2}}\left[\Sigma_1\Sigma_2\right].\end{aligned}$$

This allows us to conclude that $\mathrm{trace}\left[\left(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] = \mathrm{trace}\left[(\Sigma_1\Sigma_2)^{\frac{1}{2}}\right]$. Then, exploiting the parsimonious structure of both Σ_1 and Σ_2, one can get a form of $\Sigma_1\Sigma_2$ that depends only on low rank matrix calculations:

$$\begin{aligned}\Sigma_1\Sigma_2 &= (U_1\Lambda_1U_1^t + \sigma_1^2I_p)(U_2\Lambda_2U_2^t + \sigma_2^2I_p) \\ &= U_1\Lambda_1U_1^tU_2\Lambda_2U_2^t + \sigma_1^2U_2\Lambda_2U_2^t + \sigma_2^2U_1\Lambda_1U_1^t \\ &+ \sigma_1^2\sigma_2^2I_p.\end{aligned}$$

Combining the different parts above allows us to conclude. □

Remark 2.4. It is first important to notice that Proposition 2.3 provides a numerically efficient ways to compute the 2-Wasserstein distance in high-dimensional spaces. Indeed, the formulae exhibited above involves the computing of the trace of the square root of a matrix which is expressed only with low-rank matrix calculations. This will even be more determinant when the different elements involved need to be estimated from the data, as discussed later in this paper.

Remark 2.5. Let us also notice that Proposition 2.3 is valid even when the intrinsic dimensions d_1 and d_2 of the two HD-Gaussian distributions are different.

In the specific case where the two distributions share the same subspace, i.e. $U_1 = U_2$, the previous result reduces to an even simpler form of the 2-Wasserstein distance, as stated in the next proposition.

Proposition 2.6. *The 2-Wasserstein distance between two HD-Gaussian distributions* $\mu_1 \sim \mathcal{N}_{HD}(m_1, U, \Lambda_1, \sigma_1^2, d)$ *and* $\mu_2 \sim \mathcal{N}_{HD}(m_2, U, \Lambda_2, \sigma_2^2, d)$ *is*

$$\begin{aligned}W_2^2(\mu_1, \mu_2) = \|m_1 - m_2\|_2^2 &+ \sum_{j=1}^{d}\left(\sqrt{\delta_{1j}} - \sqrt{\delta_{2j}}\right)^2 \\ &+ p\left(\sigma_1 - \sigma_2\right)^2.\end{aligned}$$

Proof. Starting with the result of Proposition 2.3 and assuming now that $U_1 = U_2 = U$, we can first rewrite $(\Sigma_1\Sigma_2)^{\frac{1}{2}}$ as:

$$\begin{aligned}
[\Sigma_1\Sigma_2]^{\frac{1}{2}} &= \left[(U\Lambda_1U^t + \sigma_1^2 I_p)(U\Lambda_2U^t + \sigma_2^2 I_p)\right]^{\frac{1}{2}} \\
&= \left[(Q\Delta_1Q^t)(Q\Delta_2Q^t)\right]^{\frac{1}{2}} \\
&= \left[Q(\Delta_1\Delta_2)Q^t\right]^{\frac{1}{2}} = Q\left[\Delta_1\Delta_2\right]^{\frac{1}{2}}Q^t \\
&= Q\mathrm{diag}(\sqrt{\delta_{11}\delta_{21}}, ..., \sqrt{\delta_{1d}\delta_{2d}}, \sigma_1\sigma_2, ..., \sigma_1\sigma_2)Q^t.
\end{aligned}$$

Therefore, as Q is an orthonormal $p \times p$ matrix, $\mathrm{trace}([\Sigma_1\Sigma_2]^{\frac{1}{2}})$ becomes:

$$\mathrm{trace}([\Sigma_1\Sigma_2]^{\frac{1}{2}}) = \sum_{j=1}^{d}\sqrt{\delta_{1j}}\sqrt{\delta_{2j}} + (p-d)\sigma_1\sigma_2.$$

Reporting this quantity in the final formulation of the 2-Wasserstein distance, we get

$$\begin{aligned}
W_2^2(\mu_1,\mu_2) &= \|m_1 - m_2\|_2^2 + \mathrm{trace}(\Lambda_1) + \mathrm{trace}(\Lambda_2) \\
&\quad + p(\sigma_1^2 + \sigma_2^2) - 2\mathrm{trace}([\Sigma_1\Sigma_2]^{\frac{1}{2}}) \\
&= \|m_1 - m_2\|_2^2 + \sum_{j=1}^{d}(\lambda_{1j} + \lambda_{2j}) + p(\sigma_1^2 + \sigma_2^2) \\
&\quad - 2\left(\sum_{j=1}^{d}\sqrt{\delta_{1j}}\sqrt{\delta_{2j}} + (p-d)\sigma_1\sigma_2\right).
\end{aligned}$$

Finally, recalling that $\delta_{ij} = \lambda_{ij} + \sigma_i^2$, we get

$$\begin{aligned}
W_2^2(\mu_1,\mu_2) &= \|m_1 - m_2\|_2^2 + \sum_{j=1}^{d}\left(\delta_{1j} + \delta_{2j} - 2\sqrt{\delta_{1j}}\sqrt{\delta_{2j}}\right) \\
&\quad + (p-d)(\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2) \\
&= \|m_1 - m_2\|_2^2 + \sum_{j=1}^{d}\left(\sqrt{\delta_{1j}} - \sqrt{\delta_{2j}}\right)^2 \\
&\quad + (p-d)(\sigma_1 - \sigma_2)^2
\end{aligned}$$

This concludes the proof. □

Remark 2.7. The above proposition recovers and generalizes results established by several previous works, including Dowson and Landau (1982), Takatsu (2011) and Peyré et al. (2019). Indeed, if we set $d = d_1 = d_2 = p - 1$, we recover exactly those well-known results. If $d < p$, and in particular if d is small compared to p, this new formula is proposing a sort of regularization of the general expression, which may have an interesting numerical behavior in practical high-dimensional situations.

Remark 2.8. Despite the elegant form of the 2-Wasserstein distance in the case $U_1 = U_2$, exploiting this formula is far from being trivial in practice and this is rarely highlighted in the literature. Indeed, the (statistical) estimation of a common subspace of dimension d of two sets of data distributed as two different HD-Gaussian distributions is a quite complex problem, that requires the use of iterative algorithms, such as the Flury-Gautschi algorithm Flury and Gautschi (1986), to solve this problem.

2.3 Calculation of the optimal transport plan

Let us now consider the calculation of the optimal transport plan between two HD-Gaussian distributions μ_1 and μ_2. The following proposition exhibits a closed-form expression of the Monge map for the transport of μ_1 toward μ_2, involving an exact calculation of the inverse square-root of the covariance matrix of the source distribution.

Theorem 2.9. *The optimal transport map T^* between two HD-Gaussian distributions $\mu_1 \sim \mathcal{N}_{HD}(m_1, U_1, \Lambda_1, \sigma_1^2, d_1)$ and $\mu_2 \sim \mathcal{N}_{HD}(m_2, U_2, \Lambda_2, \sigma_2^2, d_2)$ is*

$$\forall x \in \mathbb{R}^p,\ T^*(x) = m_2 + \Sigma_1^{-\frac{1}{2}} \left[\Sigma_1^{\frac{1}{2}} \Sigma_2 \Sigma_1^{\frac{1}{2}}\right]^{\frac{1}{2}} \Sigma_1^{-\frac{1}{2}} (x - m_1),$$

where both $\Sigma_1^{\frac{1}{2}}$ and $\Sigma_1^{-\frac{1}{2}}$ have the explicit closed-form formulations

$$\Sigma_1^{\frac{1}{2}} = \sigma_1 I_p + U_1 C_1 U_1^t,$$

with $C_1 = \mathrm{diag}(\sqrt{\delta_{11}} - \sigma_1, ..., \sqrt{\delta_{1d}} - \sigma_1) > 0$ and

$$\Sigma_1^{-\frac{1}{2}} = \frac{1}{\sigma_1} \left(I_p - U_1 D_1 U_1^t\right)$$

with $D_1 = \mathrm{diag}\left(\frac{\sqrt{\delta_{11}} - \sigma_1}{\sqrt{\delta_{11}}}, ..., \frac{\sqrt{\delta_{1d}} - \sigma_1}{\sqrt{\delta_{1d}}}\right)$.

Proof. The optimal transport map T^* between two Gaussian distributions $\mu_1 \sim \mathcal{N}(m_1, \Sigma_1)$ and $\mu_2 \sim \mathcal{N}(m_2, \Sigma_2)$ is affine and is given by (Dowson and Landau, 1982; Takatsu, 2011):

$$\forall x \in \mathbb{R}^p,\ T^*(x) = m_2 + A^{-1}(x - m_1), \tag{3}$$

where $A^{-1} = \Sigma_1^{-\frac{1}{2}} \left[\Sigma_1^{\frac{1}{2}} \Sigma_2 \Sigma_1^{\frac{1}{2}}\right]^{\frac{1}{2}} \Sigma_1^{-\frac{1}{2}}$ involves difficult computations in high-dimensional spaces. Assuming that μ_1 and μ_2 have structured covariance structures of the form of the HD-Gaussian distribution, i.e. $\Sigma_i = U_i \Lambda_i U_i^t + \sigma_i^2 I_p$, for $i = 1, 2$, let us first focus on the computation of $\Sigma_1^{\frac{1}{2}}$. To do so, we start with Theorem 1.35 of Higham (2008) which expresses the form of $f(M)$ where $M = AB + \alpha I_p$ and f is defined on the spectrum of $AB + \alpha I_p$:

$$f(M) = f(\alpha) I_p + A(BA)^{-1} \left(f(BA + \alpha I_d) - f(\alpha) I_d\right) B.$$

Applying this result to the function $f(x) = x^{\frac{1}{2}}$, we get:

$$M^{\frac{1}{2}} = \alpha^{\frac{1}{2}} I_p + A(BA)^{-1} \left((BA + \alpha I_d)^{\frac{1}{2}} - \alpha^{\frac{1}{2}} I_d\right) B.$$

Working now with $M = \Sigma_1 = U_1(\Lambda_1 U_1^t) + \sigma_1^2 I_p$, we get:

$$\begin{aligned}\Sigma_1^{\frac{1}{2}} =& \sigma_1 I_p + U_1 \left((\Lambda_1 U_1^t) U_1\right)^{-1} \\ & \left((\sigma_1^2 I_d + (\Lambda_1 U_1^t) U_1)^{\frac{1}{2}} - \sigma_1 I_d\right)(\Lambda_1 U_1^t).\end{aligned}$$

Since $U_1^t U_1 = I_d$, the equation reduces to:

$$\Sigma_1^{\frac{1}{2}} = \sigma_1 I_p + U_1 \Lambda_1^{-1} \left((\sigma_1^2 I_d + \Lambda_1)^{\frac{1}{2}} - \sigma_1 I_d\right) \Lambda_1 U_1^t.$$

Furthermore, as $\Lambda_1 = \mathrm{diag}(\delta_{i1} - \sigma_i^2, ..., \delta_{id} - \sigma_i^2)$, we get $(\sigma_1^2 I_d + \Lambda_1)^{\frac{1}{2}} = \mathrm{diag}(\sqrt{\delta_{11}}, ..., \sqrt{\delta_{1d}})$, which can be reinjected in the above fomula:

$$\begin{aligned}\Sigma_1^{\frac{1}{2}} &= \sigma_1 I_p + U_1 \Lambda_1^{-1} C_1 \Lambda_1 U_1^t \\ &= \sigma_1 I_p + U_1 C_1 U_1^t,\end{aligned}$$

where $C_1 = \mathrm{diag}(\sqrt{\delta_{11}} - \sigma_1, ..., \sqrt{\delta_{1d}} - \sigma_1)$. Let's now consider the computation of $\Sigma_1^{-\frac{1}{2}}$:

$$\Sigma_1^{-\frac{1}{2}} = \left(\sigma_1 I_p + U_1 \Lambda_1^{-1} C_1 \Lambda_1 U_1^t\right)^{-1}.$$

Using now the Woodbury formula, we get:

$$\begin{aligned}\Sigma_1^{-\frac{1}{2}} &= \frac{1}{\sigma_1} I_p - \frac{1}{\sigma_1} U_1 \left(C_1^{-1} + U_1^t \frac{1}{\sigma_1} I_d U_1\right)^{-1} U_1^t \frac{1}{\sigma_1} I_p \\ &= \frac{1}{\sigma_1}\left(I_p - \frac{1}{\sigma_1} U_1 \left(C_1^{-1} + \frac{1}{\sigma_1} I_d\right)^{-1} U_1^t\right) \\ &= \frac{1}{\sigma_1}\left(I_p - \frac{1}{\sigma_1} U_1 \tilde{D}_1 U_1\right),\end{aligned}$$

where $\tilde{D}_1 = \left(C_1^{-1} + \frac{1}{\sigma_1} I_d\right)^{-1}$. Taking into account the diagonal structure of C_1, we can write:

$$\begin{aligned}\tilde{D}_1 &= \left(C_1^{-1} + \frac{1}{\sigma_1} I_d\right)^{-1} \\ &= \left[\mathrm{diag}(\sqrt{\delta_{1j}} - \sigma_1)_{j=1,...d}^{-1} + \frac{1}{\sigma_1} I_d\right]^{-1} \\ &= \left[\mathrm{diag}(\frac{1}{\sqrt{\delta_{1j}} - \sigma_1} + \frac{1}{\sigma_1})_{j=1,...d}\right]^{-1} \\ &= \sigma_1 \mathrm{diag}\left(\frac{\sqrt{\delta_{1j}} - \sigma_1}{\sqrt{\delta_{1j}}}\right)_{j=1,...d}.\end{aligned}$$

We finally get:

$$\Sigma_1^{-\frac{1}{2}} = \frac{1}{\sigma_1}\left(I_p - U_1 D_1 U_1\right),$$

with $D_1 = \mathrm{diag}\left(\frac{\sqrt{\delta_{1j}} - \sigma_1}{\sqrt{\delta_{1j}}}\right)_{j=1,...d}$. □

Remark 2.10. The computation of the transport map T^* usually requires the inversion of a covariance matrix, which will be rarely of full rank in high-dimensional spaces. In our case, Proposition 2.9 provides an explicit and stable inverse of the square-root of the covariance matrix Σ_1 and consequently an efficient and numerically stable way of computing the transport plan T, even in situations where Σ_1 and Σ_2 are not of full rank. In addition, Proposition 2.9 also provides an explicit form of the square-root of Σ_1.

Remark 2.11. Once again, the result of Proposition 2.9 is valid even when d_1 and d_2 are different. This is naturally a key point in practical situations where there is no reason to have distributions with identical intrinsic dimensions.

2.4 Inference and intrinsic dimension estimation

Inference Assuming that two point clouds $X^{(1)}$ and $X^{(2)}$ sampled from HD-Gaussisan distributions are given and that their intrinsic dimensions d_1 and d_2 are known, the computation of both the 2-Wasserstein distance and the associated transport map requires the estimation of the parameters $\mu_i, \lambda_{ij}, \sigma_i$ and U_i, for $i = 1, 2$ and $j = 1, ..., d$. Following Tipping and Bishop (1999b), the maximum likelihood estimates of those parameters are, for $i = 1, 2$:

$$\hat{\mu}_i = \sum_{\ell=1}^{n_i} x_\ell^{(i)}/n_i, \qquad \hat{\lambda}_{ij} = \omega_j(S_i) - \hat{\sigma}_i$$
$$\hat{\sigma}_i = \left(\mathrm{trace}(S_i) - \sum_{j=1}^{d} \omega_j(S_i) \right) /(p-d),$$

and $\hat{U}_i$ is formed by the d_i leading eigenvectors (i.e. associated with the d_i largest eigenvalues $\omega_j(S_i)$) of S_i. Finally, $S_i = (X^{(i)} - \hat{\mu}_i)^t (X^{(i)} - \hat{\mu}_i)$ is the empirical covariance matrix.

Estimation of the intrinsic dimensions As in practical situations the intrinsic dimensionality of the data is not known, we also need to estimate d_1 and d_2 from the data. This question has been intensively studied in the last two decades and remains a difficult question in general. Among the possible solutions, we can cite the works of Cattell (1966), Bouveyron et al. (2011), Josse and Husson (2012) and Bouveyron et al. (2020). Even though intrinsic dimensionality estimation is a challenging task in general, the effect of some variation on the estimation of the actual dimensions of the source and target distributions will be limited in our case since we model the data in the whole high-dimensional space with a parsimonious approach, without effective dimension reduction. As illustrated in Appendix B, the cross-validation approach of Josse and Husson (2012) for PCA performs well in a variety of situations and we recommend to use it in practice. This technique will be used in the following for estimating the intrinsic dimensions of the source and target distributions.

3 Numerical experiments

3.1 Experimental setups and methods

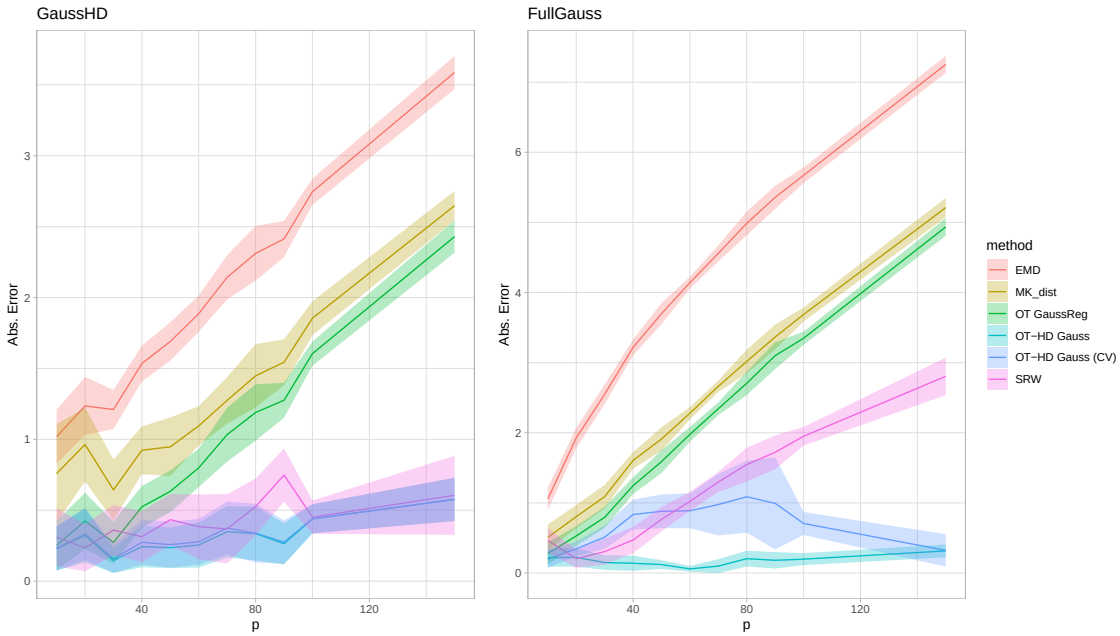
Simulated scenarios For evaluating the proposed method performance in calculating both the Wasserstein distance and the Monge map, we designed 3 simulation scenarios:

i) The first scenario, hereafter referred to as `GaussHD`, consists in drawing n observations in dimension p from two HD-Gaussian distributions, as defined by Definition 2.1. While n and p are allowed to vary in order to both test the effects "sample size" and "high dimension", we assumed that both distributions share $d_1 = d_2 = 5$ and that are centered (these requirements, i.e. same intrinsic dimension and centrality, are needed by some competitor approaches). Moreover, we set $\sigma_1^2 = 0.4$ and $\sigma_2^2 = 0.2$ whereas $\text{diag}(\Lambda) = \{\lambda_i, \lambda_i, \lambda_i, \lambda_i, \lambda_i\}$, with $i = 1, 2$, $\lambda_1 = 3.6$ and $\lambda_2 = 1.8$.
ii) The second scenario, called `FullGauss`, assumes the data are sampled form two centered Gaussian distributions ($d_1 = d_2 = p$). For the source distribution (respectively the destination distribution) we created a decreasing sequence of p eigenvalues ranging from 3.6 to 0.4 (1.8 to 0.2) representing the spectrum of the covariance matrix.
iii) The last simulation scenario considers two non Gaussian distributions: the skew-Normal (Azzalini, 2013) and Student distributions. In this case, data are sampled from multivariate (p-dimensional) skew-Normal and Student distributions where the scaling / correlation matrices of the two distributions are simulated as for the `FullGauss` scenario.

State-of-the-art methods Once the source and target point clouds are sampled, in order to compute the Wasserstein distance (and possibly the Monge map) between the generating distributions, two global strategies exist, depending the considered approach. Either the parameters of the distributions are learned from the data and *then* the Wasserstein distance and Monge map are computed via the Gaussian closed formulas, otherwise each point is equipped with mass $1/n$ and the Wasserstein distance (or another OT distance) is learned *directly*, numerically. We will compare hereafter the following 9 methods, adopting one or the other strategy: `OT-Gauss`, the classical W_2^2-distance and Monge map computations between 2 Gaussians, `OT-GaussReg`, its ridge-regularization, the Earth movers distance `EMD`, `Sinkhorn` (Cuturi, 2013), `SRW` (Paty and Cuturi, 2019), `SWD` (Bonneel et al., 2015), `MK-dist` and `MI-dist` Muzellec and Cuturi (2019), and `OT-HDGauss`, the approach proposed in this work. More details about these approaches are given in Appendix A. Let us notice that `EMD`, `Sinkhorn`, `SWD` and `SRW` are limited to the calculation of OT distances and they cannot compute the Monge map. Consequently, they won't be used for comparisons about transport maps.

3.2 Computation of the W_2^2-distance

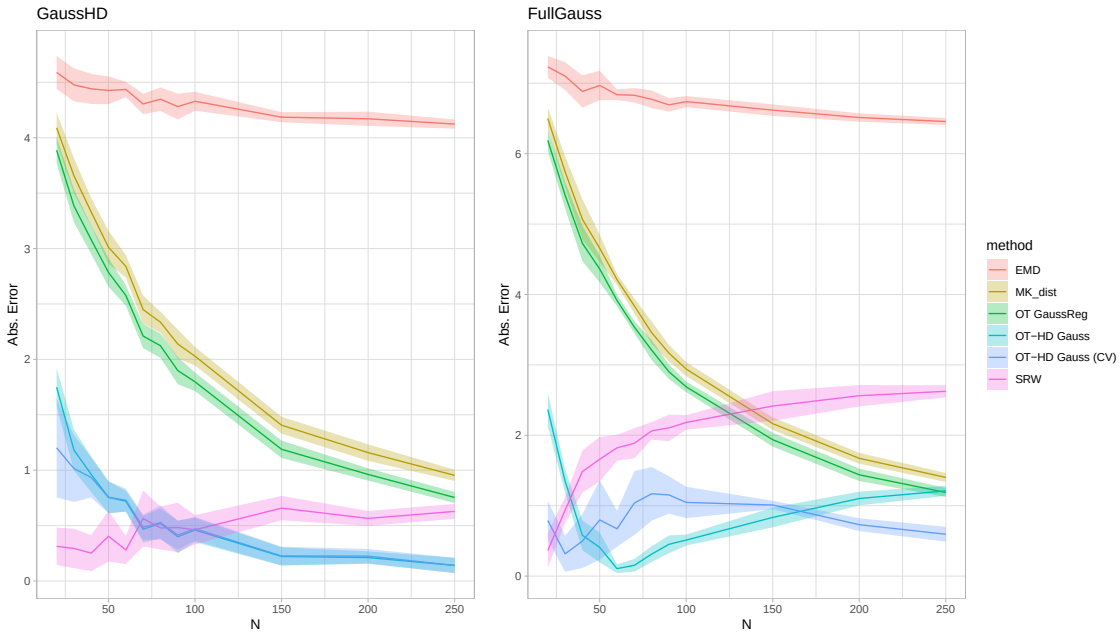
In this first experiment, we focus on the numerical computation of the W_2^2-distance between two Gaussian distributions (`HDGauss` and `FullGauss` scenarios). In particular, we aim to study the robustness of the considered OT approaches against the data dimensionality p and the sample size n. To this end, we first simulated data from source and target distributions according to the `HDGauss` and `FullGauss` scenarios, with a fixed sample size $n = 50$ and varying dimensions of the observations space $p \in [10, 150]$. We applied the OT methods listed above on these simulated data to calculate the W_2^2-distance between the source and target distributions. For methods working on subspaces, i.e. `MK-dist`, `MI-dist` and `SRW`, we always provided them with the actual intrinsic dimension $d = 5$. Same for `OT-HDGauss`. The performance of these approaches in computing the W_2^2-distance between μ_s and μ_t is assessed by the absolute-value difference with the exact Gaussian W_2^2-distance computed between the true distributions, whose parameters are known. All results are averaged over 25 replications. Figure 1 presents the performance in computing the W_2^2-distance for the different methods, according to the space dimensionality p, and this for both HD-Gaussian and full Gaussian distributions. In order to keep the exposition uncluttered, we did not report results for



(a) HD-Gaussian, $n = 50$ (b) Full Gaussian, $n = 50$

Figure 1: Absolute value difference between the actual W_2^2-distance (computed from the true distributions) and their estimations using the compared methods, with a fixed dimension $n = 50$ of the observation space and where the dimension p of the observation space varies: (a) with simulated HD-Gaussian distributions, (b) with simulated full Gaussian distributions. Results are averaged over 25 replications.

all the methods listed in the previous section. In particular: the behaviour of `OT-Gauss` is almost indistinguishable from `OT-GaussReg` in low-dimension, and stops working for $p > n$. Same remark for `Sinkhorn` and `EMD` (not reported). Instead, `SWD` and `MI-dist` are systematically outperformed by (for instance) `MK-dist` and this is reported in Appendix C. From Figure 1, it clearly appears that `EMD`, `MK-dist` and `OT-Gauss` have a high sensibility to the data dimensionality in both scenarios and make important errors in the computation of the Wasserstein distance in high-dimensional spaces. In the case of the `HDGauss` scenario, `SRW` and the two `OT-HDGauss` approaches show a good robustness to the dimensionality and see their estimations of the Wasserstein distance are little impacted by the increase of the dimensionality. Not surprisingly, the two `OT-HDGauss` approaches only slightly outperform `SRW` here since the simulation scenario is favorable. This is however not the case for the `FullGauss` scenario (Figure 1-b) where the `OT-HDGauss` approaches outperform all approaches, including `SRW`, even though the data are not simulated according to their model. We also studied the robustness of the considered OT approaches against the sample size n in high dimensions. For this, we simulated data from source and target distributions according to the `HDGauss` and `FullGauss` scenarios, with a fixed dimension $p = 100$ and varying sample sizes $n \in [20, 250]$. Figure 2 presents the performance in computing the W_2^2-distance for the different methods, according to the sample size n for both simulation scenarios. One can first notice that `EMD` performs badly whatever the sample size. Conversely, `MK-dist` and `OT-Gauss` benefit from the increase of the sample size and significantly improve their performance when the sample size is clearly larger than the space dimensionality. This experiment also reveals a surprising behavior of



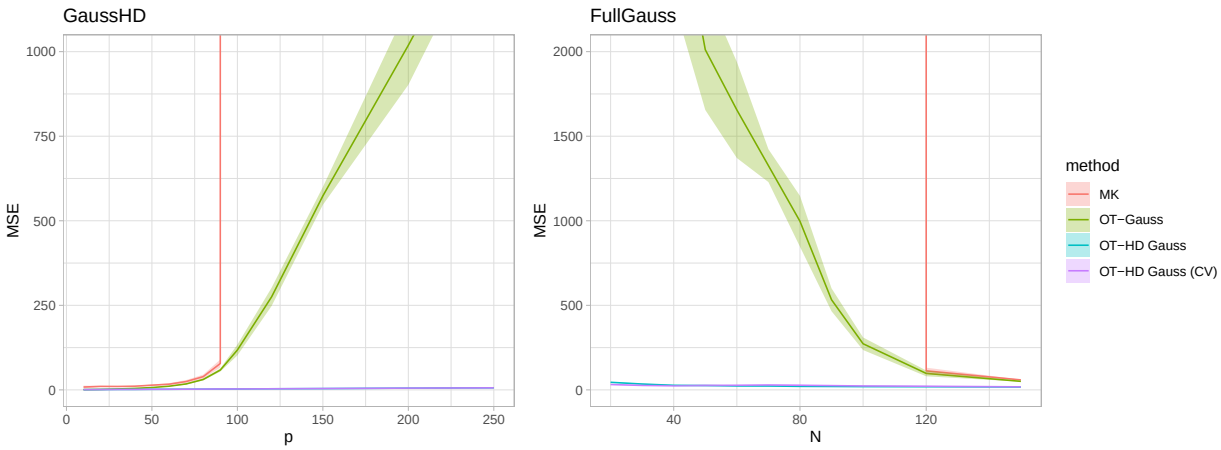
(a) HD-Gaussian, $p = 100$ (b) Full Gaussian, $p = 100$

Figure 2: Absolute value difference between the actual W_2^2-distance (computed from the true distributions) and their estimations using the compared methods, with a fixed dimension $p = 100$ of the observation space and where the sample size n varies: (a) with simulated HD-Gaussian distributions, (b) with simulated full Gaussian distributions. Results are averaged over 25 replications.

`SRW`, which was not possible to see in the previous experiment: the performance of `SRW` decreases with the increase of the sample size. In the `FullGauss` scenario `OT-HDGauss` exhibits the same behavior and this can be explained by the fact that generative model it is based on is no longer the true one, thing that emerges for large n. However in the `HDGauss` scenario the data *almost* live in a subspace of dimension d, thing that should favour SRW but apparently it does not. This is rather counter intuitive and probably linked to the fact that `SRW` does not compute the exact Wasserstein distance, but a lower-bound of it. Finally, `OT-HDGauss` demonstrates here again a clear robustness to the sample size in high dimensions, in both scenarios. In the `FullGauss` case, the `OT-HDGauss` with the CV procedure to select the intrinsic dimension has to be recommended since it better adapts to the data.

3.3 Computation of the Monge map

This experiment now focuses on the computation of the transport map, that our approach is also able to compute. Here again, we aim at studying the the robustness against the data dimensionality p and the sample size n of the OT approaches allowing the Monge map computation. To this end, we simulated data from source and target distributions according to the `HDGauss` and `FullGauss` scenarios, first with a fixed sample size $n = 50$ and varying dimensions of the observations space $p \in [10, 150]$, and second with a fixed dimension $p = 100$ and varying sample sizes $n \in [20, 250]$. On these simulated data set, we then applied the 3 methods able to compute the transport map. In order to measure the performance of the transport undertaken, we simulated two point clouds, one



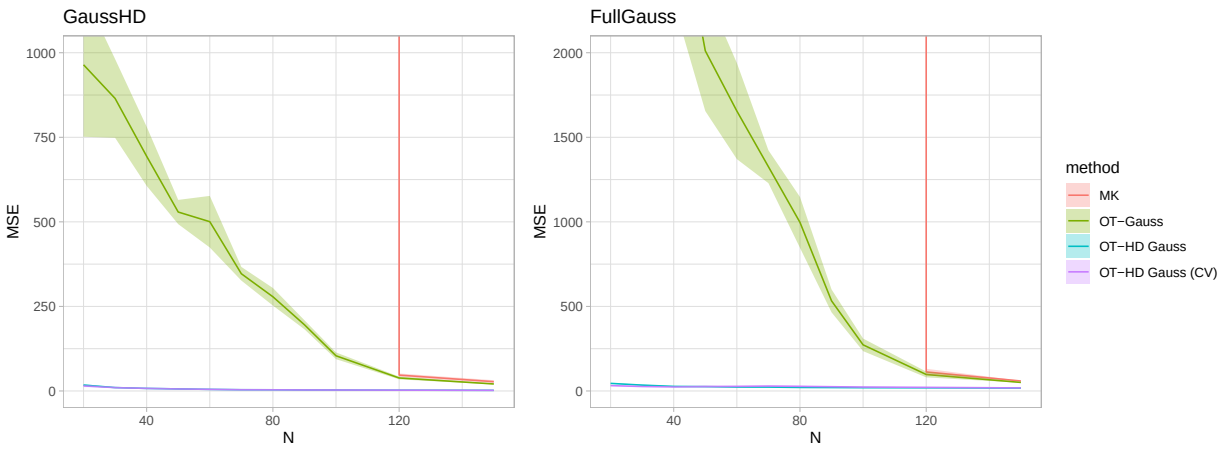
(a) HD-Gaussian, $n = 50$ (b) Full Gaussian, $n = 50$

Figure 3: Mean squared error between the Monge map estimated by the compared methods and the actual transport on a test data set, with a fixed sample size $n = 50$ and where the dimension p of the observation space varies: (a) with simulated HD-Gaussian distributions and (b) with simulated full Gaussian distributions. Results are averaged over 25 replications.

from a source measure μ_1 and the other from a destination measure μ_2, both being either Gaussian or HD-Gaussian distributions. Then, each cloud was split in train and test. We used the source-train and the destination-train to estimate the Monge map between μ_1 and μ_2, then we transported the source-test with the oracle Monge map as well as with the Monge maps estimated by all methods. Finally the mean squared error (MSE) between the test-transported points (oracle vs. estimated) was computed. The panels (a) and (b) of Figure 3 present the performance evolution of the best OT methods according to the space dimensionality (the sample size is fixed to $n = 50$), for both the HD-Gaussian and full Gaussian scenarios. In both scenarios, one can first notice that `MK-dist` fails to compute the transport map in dimension higher than 80. After this dimension $p = 80$, one can also observe a rapid deterioration in performance of `OT-Gauss`, even with a numerical regularization. Conversely, the `OT-HDGauss` and `OT-HDGauss (CV)` approaches show once again a good robustness in performance when the space dimensionality increases. Figure 4 presents the results of the same 4 OT methods when the sample size n varies and for a fixed dimensionality $p = 100$ of the observation space. One can observe similar results here, and this for both simulated scenarios: `MK-dist` fails to compute the transport map for n smaller than 120 and this sample size is also a breakpoint for the performance of `OT-Gauss`. Here again, `OT-HDGauss` and `OT-HDGauss (CV)` turn out to be robust in performance against the sample size, even when $n << p$.

3.4 Transport of non Gaussian distributions

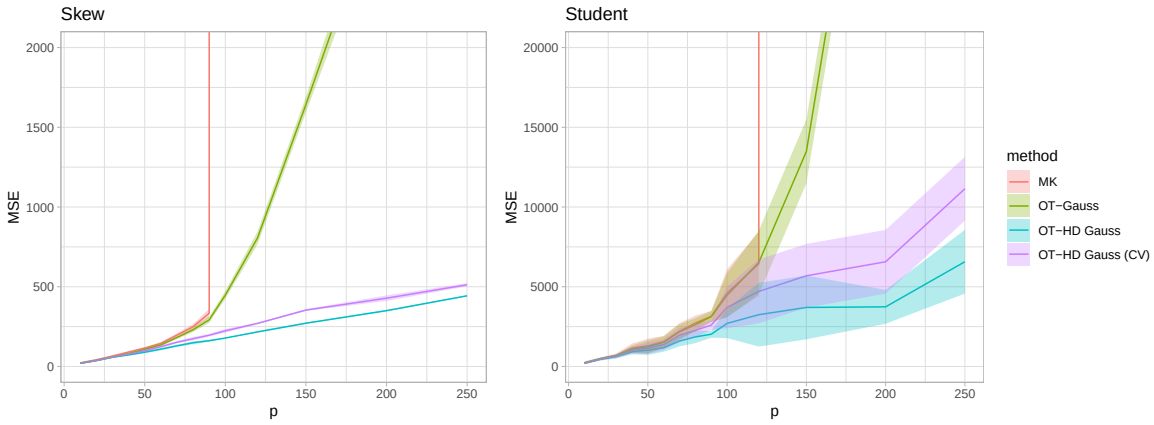
This last experiment focuses on the transport of non Gaussian distributions. The aim here is to study the robustness of our approach to deviation from the Gaussian assumption and to compare



(a) HD-Gaussian, $p = 100$ (b) Full Gaussian, $p = 100$

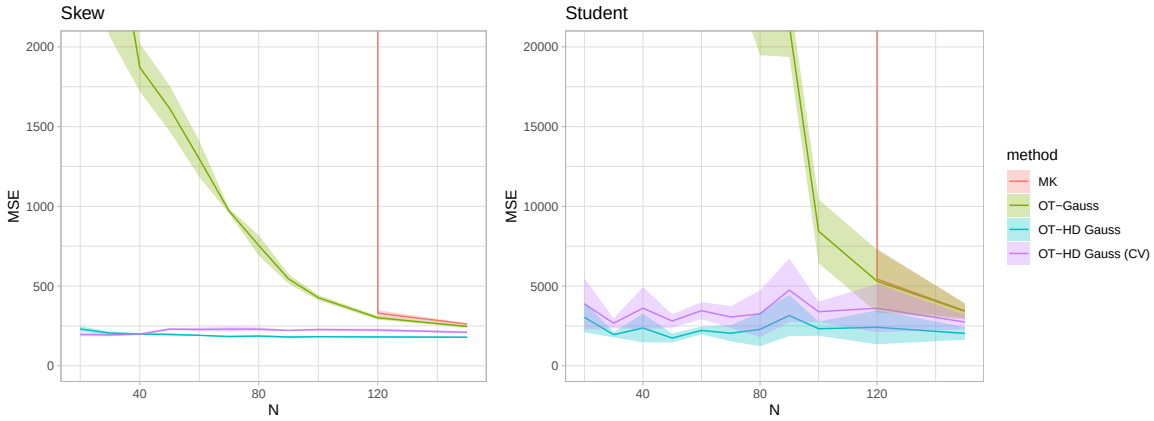
Figure 4: Mean squared error between the Monge map estimated by the compared methods and the actual transport on a test data set, with a fixed dimension $p = 100$ of the observation space and where the sample size n varies: (a) with simulated HD-Gaussian distributions and (b) with simulated full Gaussian distributions. Results are averaged over 25 replications.

with other transport approaches (`MK-dist` and `MI-dist`) that are model-free, in the context of high-dimensional spaces. To this end, we simulated a point cloud from a source distribution either multivariate skew Normal or Student distribution for different sample sizes and varying dimensions of the observation space. Figures 10 and 11 of Appendix D present pairs plots of simulated data from these non Gaussian distributions. After splitting the source cloud into train and test, we transported the source-train with a fixed linear map in such a way to leave the transported points centered at the origin and used the source-train and transported source-train in order to estimate the Monge maps with the three methods. As before, the MSE between the transported source-test (oracle vs. estimated) was computed. Figures 5 and 6 present the performance evaluations of the same OT methods for multivariate Skew Normal and Student distributions respectively according to the space dimensionality p and the sample size n. The performances of `MK-dist` and `OT-Gauss` are similar to the Gaussian case (previous experiment). Even though its robustness is less impressive than in the Gaussian case, `OT-HDGauss` performs here also quite well in general even though the distributions clearly differ from the Gaussian one, and in any case clearly outperforms all tested OT methods.



(a) Skew Normal, $n = 50$ (b) Student, $n = 50$

Figure 5: Mean squared error between the Monge map estimated by the compared methods and the actual transport on a test data set, with a fixed sample size $n = 50$ and where the dimension p of the observation space varies: a) with simulated Skew Normal distributions and (b) with Student distributions. Results are averaged over 25 replications.



(a) Skew Normal $p = 100$ (b) Student, $p = 100$

Figure 6: Mean squared error between the Monge map estimated by the compared methods and the actual transport on a test data set, with a fixed dimension $p = 100$ of the observation space and where the sample size n varies: a) with simulated Skew Normal distributions and (b) with Student distributions. Results are averaged over 25 replications.

4 Conclusion and discussion

This work has focused on the optimal transport of high-dimensional Gaussian (HD-Gaussian) distributions, induced by the probabilistic PCA (PPCA) model. In particular, we exhibited of a closed-form expression of the Wasserstein distance between two HD-Gaussian distributions, with an efficient and robust calculation procedure based on a low-dimensional subspace decomposition, and this without relying on data projections. This result also generalizes previous state-of-the-art results which considered Gaussian distributions with similar covariance orientations or structures. Furthermore, we provided a closed-form expression of the Monge map for the transport of a HD-Gaussian distribution on another one, involving an exact calculation of both the square-root and the inverse square-root of the covariance matrix of the source distribution. This result avoids in turn many numerical drawbacks in high-dimensional practical situations and remain valid in the case of HD-Gaussian distributions with different intrinsic dimensions. These contributions are supported by numerical experiments that highlight the performance and robustness of the proposed OT-HDGauss procedure to both the dimensionality and the sample size, and this in comparison with the most recent OT approaches. The numerical experiments also showed that the analytical and numerical advantages of our approach in high dimensions allow it to also outperform model-free methods in the case of non-Gaussian distributions. Among the possible further work, it would be interesting to consider the extension of this approach to mixture models, in particular Gaussian mixture models.

References

Azzalini, A. (2013). *The skew-normal and related families*, volume 3. Cambridge University Press.

Bellman, R. (1957). *Dynamic Programming*. Rand Corporation research study. Princeton University Press.

Bishop, C. M. (1999). Variational principal components.

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. (2015). Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45.

Bouveyron, C. and Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, 71:52–78.

Bouveyron, C., Celeux, G., and Girard, S. (2011). Intrinsic dimension estimation by maximum likelihood in isotropic probabilistic pca. *Pattern Recognition Letters*, 32(14):1706–1713.

Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). *Model-based clustering and classification for data science: with applications in R*, volume 50. Cambridge University Press.

Bouveyron, C., Girard, S., and Schmid, C. (2007a). High-dimensional data clustering. *Computational statistics & data analysis*, 52(1):502–519.

Bouveyron, C., Girard, S., and Schmid, C. (2007b). High-dimensional discriminant analysis. *Communications in Statisticsâ Theory and Methods*, 36(14):2607–2623.

Bouveyron, C., Latouche, P., and Mattei, P.-A. (2020). Exact dimensionality selection for bayesian pca. *Scandinavian Journal of Statistics*, 47(1):196–211.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, 1(2):245–276.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Dowson, D. and Landau, B. (1982). The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455.

Dudley, R. M. (1969). The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. (2021). Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8.

Flury, B. N. and Gautschi, W. (1986). An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. *SIAM Journal on Scientific and Statistical Computing*, 7(1):169–184.

Fournier, N. and Guillin, A. (2015). On the rate of convergence in wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3):707–738.

Giraud, C. (2021). *Introduction to high-dimensional statistics*. Chapman and Hall/CRC.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Higham, N. J. (2008). *Functions of Matrices*. Society for Industrial and Applied Mathematics.

Josse, J. and Husson, F. (2012). Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics & Data Analysis*, 56(6):1869–1879.

Kantorovich, L. V. (1942). On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201.

McNicholas, P. D. and Murphy, T. B. (2008). Parsimonious gaussian mixture models. *Statistics and Computing*, 18:285–296.

Minka, T. (2000). Automatic choice of dimensionality for pca. *Advances in neural information processing systems*, 13.

Muzellec, B. and Cuturi, M. (2019). Subspace detours: Building transport plans that are optimal on subspace projections. *Advances in Neural Information Processing Systems*, 32.

Nguyen, K. and Ho, N. (2024). Energy-based sliced wasserstein distance. *Advances in Neural Information Processing Systems*, 36.

Paty, F.-P. and Cuturi, M. (2019). Subspace robust wasserstein distances. In *International conference on machine learning*, pages 5072–5081. PMLR.

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Santambrogio, F. (2015). Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94.

Takatsu, A. (2011). Wasserstein geometry of gaussian measures.

Tipping, M. E. and Bishop, C. M. (1999a). Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2):443–482.

Tipping, M. E. and Bishop, C. M. (1999b). Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(3):611–622.

Villani, C. et al. (2009). *Optimal transport: old and new*, volume 338. Springer.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

Appendix

A Details about the competitors used in numerical experiments

We provide below more details about the methods used as competitors in the numerical experiments:
- `OT-Gauss`: classical W_2^2-distance and Monge map computations between 2 full Gaussians, using respectively Eq. (2) and Eq. (3). The computation of the square-root matrices is based on a Schur decomposition (`sqrtm` function in R) and the inverse of covariance matrices is performed using the Moore-Penrose generalized inverse (`ginv` function).
- `OT-GaussReg`: classical W_2^2-distance and Monge map computations (as above), with an additional regularization of the rank of the covariance matrices ($\tilde{\Sigma}_i = \Sigma_i + \gamma I_p$, where $\gamma = 1e^{-3}$ in the experiments, $i = 1, 2$),
- `EMD`: Earth movers distance as implemented in the Python Optimal Transport (POT Flamary et al., 2021) library (`ot.emd2` function),
- `Sinkhorn`: Solution of the entropic regularized optimal transport problem, as described in Cuturi (2013) and implemented in the POT library (`ot.sinkhorn2` function),
- `SRW`: Subspace robust Wasserstein distance of Paty and Cuturi (2019), GitHub code[2],
- `SWD`: Sliced Wasserstein distance as implemented in POT and following Bonneel et al. (2015),
- `MK-dist`: Monge-Knothe transport plan and relative distance as described in Muzellec and Cuturi (2019),
- `MI-dist`: Monge Independent distance as described in Muzellec and Cuturi (2019),
- `OT-HDGauss`: the approach proposed in this work that implements the computations of the W_2^2-distance with Theorem 2.3 and the Monge map with Theorem 2.9. Additionally we denote by `OT-HDGauss (CV)` the version of our approach where the intrinsic dimension is selected by cross-validation as in Josse and Husson (2012).

B Intrinsic dimension estimation

This section aims to compare the performance of methods proposed repectively by Bouveyron et al. (2011) (hereafter PPCA-ds), Josse and Husson (2012) (PCA-CV) and Cattell (1966) (Catttell) for estimating the intrinsic dimension of HD-Gaussian distributions. In order to evaluate the effect of the estimation of the intrinsic dimensions using these techniques, we measured the error made in computing the Monge map between two HD-Gaussian distributions using our approach (based on the Theorem 2.9). For this comparison, we simulated two (isotropic) HD-Gaussian distributions with intrinsic dimensions $d_1 = d_2 = 5$, a signal-to-noise ratio of $\delta/\sigma^2 = 5$, in dimensions $p = 100$ and with varying number of observations ($n \in 50, 100, 250$). The OT-HD approach was used to compute the optimal transport plan between the two simulated distributions. Figure 7 presents the mean squared errors (MSE) measured between the two distributions with OT-HD on test data for different methods for the intrinsic dimension estimation and for different sample sizes of the data used for learning the transport map. The results are averaged over 25 simulated datasets. The results clearly show that the PCA-CV approach of Josse and Husson (2012) is the most efficient one for this task and should be recommended.

[2] `https://github.com/francoispierrepaty/SubspaceRobustWasserstein`

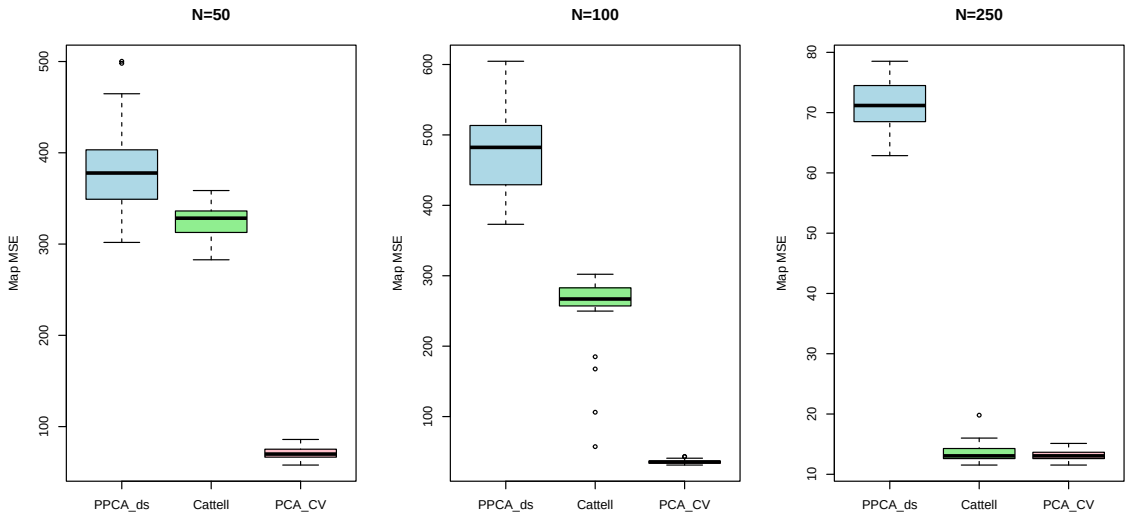


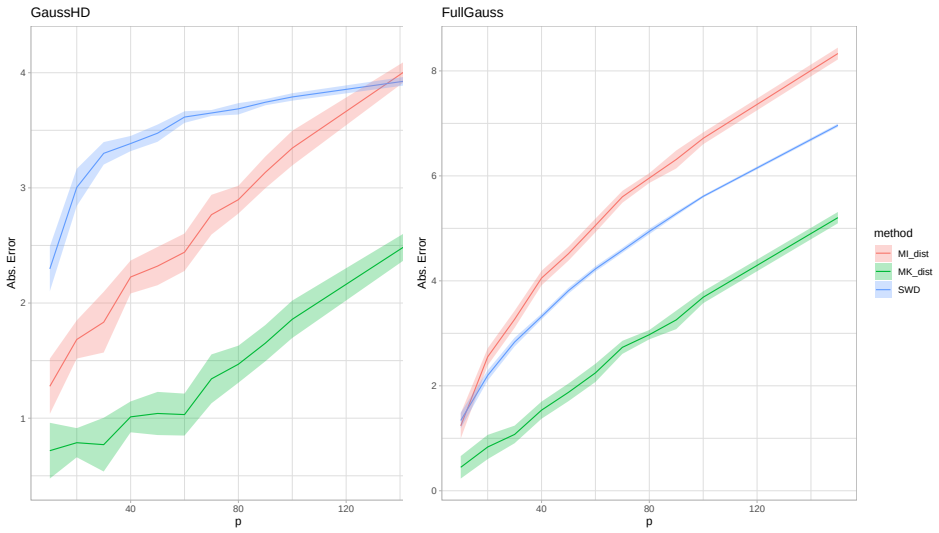
Figure 7: Effect of the choice of the intrinsic dimension estimation method on the mean squared errors of the transport of test data using OT-HD, for different sample sizes of the source distribution.

C Computation of the W_2^2-distance: additional results

We report in this section some additional results, visible in Figures 8 and 9, and comparing `MI-dist` and `SWD` with `MK-dist`. We are in the very same simulated scenarios described in Section 3.3 and, as it can be seen, `MK-dist` outperforms the two competitors.

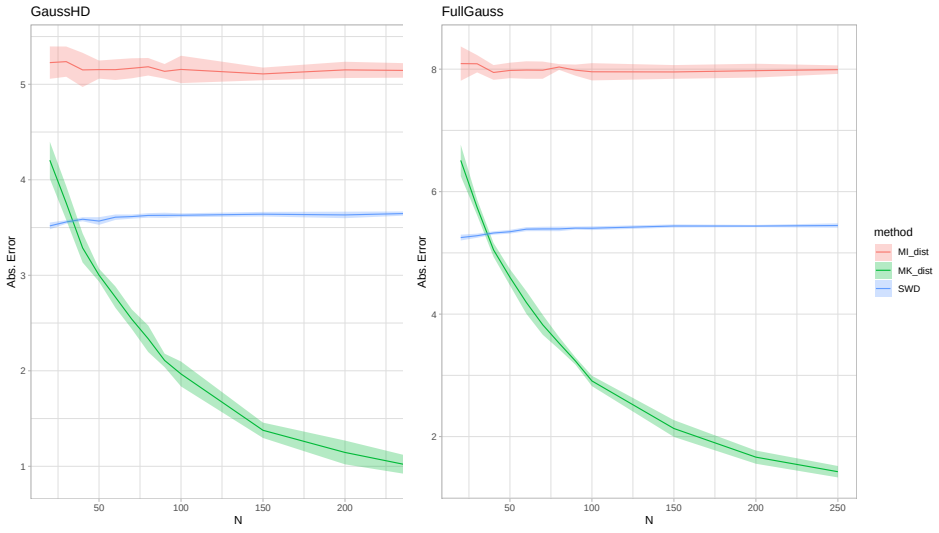
D Transport of Non Gaussian distributions

Figures 10 and 11 present pairs plots of simulated non Gaussian distributions used as source and target distributions in the experiment on non Gaussian data (Section 3.4).



(a) HD-Gaussian, $n = 50$ (b) Full Gaussian, $n = 50$

Figure 8: Absolute value difference between the actual W_2^2-distance (computed from the true distributions) and their estimations using the compared methods, with a fixed sample size $n = 50$ and where the dimension p of the observation space varies: (a) with simulated HD-Gaussian distributions, (b) with simulated full Gaussian distributions. Results are averaged over 25 replications.



(a) HD-Gaussian, $p = 100$ (b) Full Gaussian, $p = 100$

Figure 9: Absolute value difference between the actual W_2^2-distance (computed from the true distributions) and their estimations using the compared methods, with a fixed dimension $p = 100$ of the observation space and where the sample size n varies: (a) with simulated HD-Gaussian distributions, (b) with simulated full Gaussian distributions. Results are averaged over 25 replications.

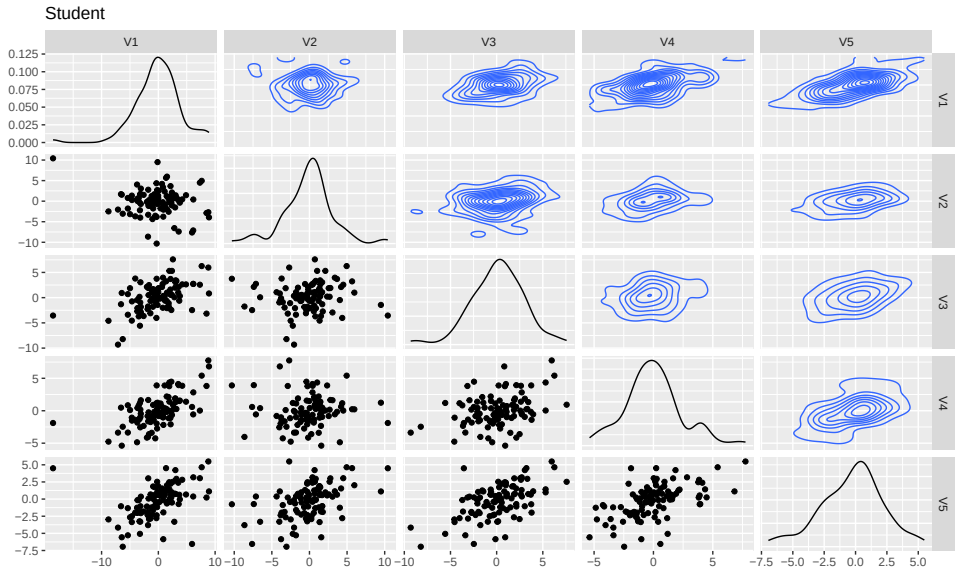


Figure 10: Pairs plot of the simulated Student distribution ($p = 5$) used for the experiment on non Gaussian data.

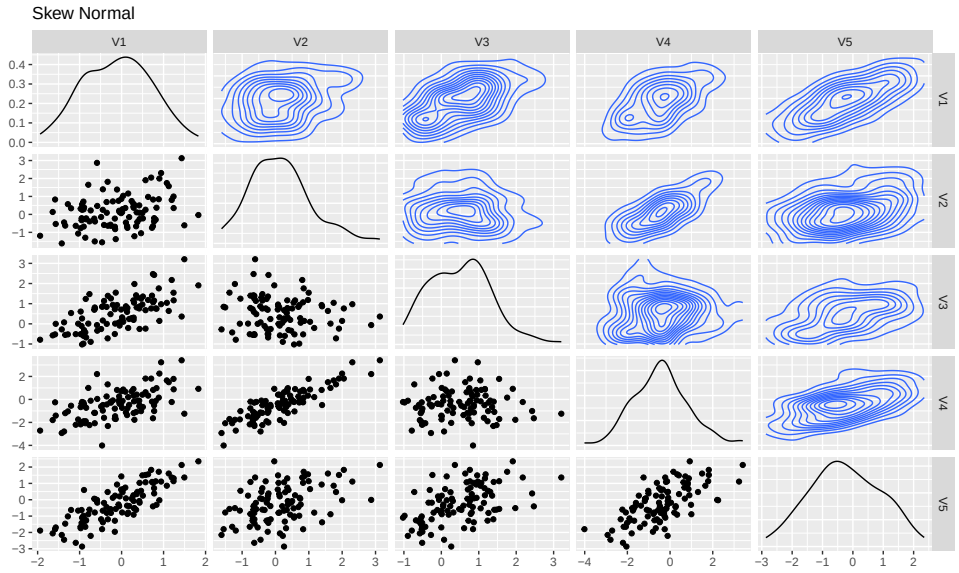


Figure 11: Pairs plot of the simulated Skew Normal distribution ($p = 5$) used for the experiment on non Gaussian data.