Unsupervised Data Denoising through Variance Maximization under Kantorovich Domination

Summer School on Optimal Transport, Stochastic Analysis and Application to Machine Learning

June 2024

Tongseok Lim (Purdue University's Daniels School of Business)

Joint work with **Brendan Pass**, **Marcelo Souza**, **Joshua Hiew**

Assumption on Data distribution

$\mathcal{P} = \mathcal{P}(\mathbb{R}^d) =$ probabilities over $\mathbb{R}^d$ with finite second moments

$\mathcal{P}_0 =$ centered probabilities ($\int x \, \mathrm{d}\mu(x) = 0$) with finite 2nd moments

$X \sim \rho, Y \sim \nu, R \sim \epsilon$ s.t. $Y = X + R$ and $\mathbb{E}[R|X] = 0$ $(\rho, \nu, \epsilon \in \mathcal{P}_0)$

Want to recover ρ from the observed data ν which is disturbed by ϵ.

Assumption on Data distribution

$\mathcal{P} = \mathcal{P}(\mathbb{R}^d) =$ probabilities over $\mathbb{R}^d$ with finite second moments

$\mathcal{P}_0 =$ centered probabilities ($\int x \, \mathrm{d}\mu(x) = 0$) with finite 2nd moments

$X \sim \rho, Y \sim \nu, R \sim \epsilon$ s.t. $Y = X + R$ and $\mathbb{E}[R|X] = 0$ $(\rho, \nu, \epsilon \in \mathcal{P}_0)$

Want to recover ρ from the observed data ν which is disturbed by ϵ.

We suppose that ρ belongs to a particular domain $\mathcal{D}$ $(\subseteq \mathcal{P}_0)$.

Joint distribution $\pi = \mathcal{L}(X, Y)$ of ρ, ν is a martingale : $\mathbb{E}_\pi[Y|X] = X$

$\mathcal{M}(\mu, \nu) =$ set of all martingale couplings of μ, ν

$\Pi(\mu, \nu) =$ set of all couplings of μ, ν

Assumption on Data distribution

$\mathcal{P} = \mathcal{P}(\mathbb{R}^d) =$ probabilities over $\mathbb{R}^d$ with finite second moments

$\mathcal{P}_0 =$ centered probabilities ($\int x \,\mathrm{d}\mu(x) = 0$) with finite 2nd moments

$X \sim \rho, Y \sim \nu, R \sim \epsilon$ s.t. $Y = X + R$ and $\mathbb{E}[R|X] = 0$ $(\rho, \nu, \epsilon \in \mathcal{P}_0)$

Want to recover ρ from the observed data ν which is disturbed by ϵ.

We suppose that ρ belongs to a particular domain $\mathcal{D}$ ($\subseteq \mathcal{P}_0$).

Joint distribution $\pi = \mathcal{L}(X, Y)$ of ρ, ν is a martingale : $\mathbb{E}_\pi[Y|X] = X$

$\mathcal{M}(\mu, \nu) =$ set of all martingale couplings of μ, ν

$\Pi(\mu, \nu) =$ set of all couplings of μ, ν

Given data ν and search domain $\mathcal{D}$, we look for $\mu \in \mathcal{D}$ solving

$$\min_{\mu \in \mathcal{D}} \min_{\pi \in \mathcal{M}(\mu,\nu)} \mathbb{E}_\pi[|X - Y|^2]. \tag{1}$$

$\Rightarrow$ We try to summarize ν by an optimal μ^* within the domain $\mathcal{D}$.

Domain examples

Probabilities on curves and surfaces. Let $\Omega \subseteq \mathbb{R}^d$ be compact.

$$\mathcal{C}_{k,L} = \left\{\alpha : [0, T] \to \Omega \,\middle|\, \alpha \in C^k, |\alpha^{(k)}| \leq M, |\alpha^{(k)}(t) - \alpha^{(k)}(s)| \leq L|t - s|\right\},$$
$$\mathcal{D} = \left\{\mu \in \mathcal{P}(\mathbb{R}^d) \,\middle|\, \mathrm{spt}(\mu) \subseteq \mathrm{Im}(\alpha) \text{ for some } \alpha \in \mathcal{C}_k\right\}.$$

$\Longrightarrow$ $\mathcal{D}$ is closed under the Wasserstein metric W_2

Domain examples

Probabilities on curves and surfaces. Let $\Omega \subseteq \mathbb{R}^d$ be compact.

$$\mathcal{C}_{k,L} = \left\{\alpha : [0, T] \to \Omega \,\middle|\, \alpha \in C^k, |\alpha^{(k)}| \leq M, |\alpha^{(k)}(t) - \alpha^{(k)}(s)| \leq L|t-s|\right\},$$

$$\mathcal{D} = \left\{\mu \in \mathcal{P}(\mathbb{R}^d) \,\middle|\, \mathrm{spt}(\mu) \subseteq \mathrm{Im}(\alpha) \text{ for some } \alpha \in \mathcal{C}_k\right\}.$$

$\implies$ $\mathcal{D}$ is closed under the Wasserstein metric W_2

Probabilities on k points (with free weights). Given $k \in \mathbb{N}$, define

$$\mathcal{D}^k = \{\mu \,|\, |\mathrm{spt}(\mu)| \leq k\} = \left\{\mu = \sum_{i=1}^k u_i \delta_{x_i} \,\middle|\, x_i \in \mathbb{R}^d, u_i \geq 0, \sum_{i=1}^k u_i = 1\right\}$$

$\implies$ $\mathcal{D}^k$ relates our problem (1) to the k-means clustering problem.

Domain examples

Probabilities on curves and surfaces. Let $\Omega \subseteq \mathbb{R}^d$ be compact.

$$\mathcal{C}_{k,L} = \left\{\alpha : [0,T] \to \Omega \,\middle|\, \alpha \in C^k, |\alpha^{(k)}| \leq M, |\alpha^{(k)}(t) - \alpha^{(k)}(s)| \leq L|t-s|\right\},$$
$$\mathcal{D} = \left\{\mu \in \mathcal{P}(\mathbb{R}^d) \,\middle|\, \mathrm{spt}(\mu) \subseteq \mathrm{Im}(\alpha) \text{ for some } \alpha \in \mathcal{C}_k\right\}.$$

$\implies$ $\mathcal{D}$ is closed under the Wasserstein metric W_2

Probabilities on k points (with free weights). Given $k \in \mathbb{N}$, define

$$\mathcal{D}^k = \{\mu \,|\, |\mathrm{spt}(\mu)| \leq k\} = \left\{\mu = \sum_{i=1}^k u_i \delta_{x_i} \,\middle|\, x_i \in \mathbb{R}^d, u_i \geq 0, \sum_{i=1}^k u_i = 1\right\}$$

$\implies$ $\mathcal{D}^k$ relates our problem (1) to the k-means clustering problem.

Probabilities on k points with uniform weights.

$$\mathcal{D}^k_{\mathrm{Uni}} = \left\{\mu = \frac{1}{k}\sum_{i=1}^k \delta_{x_i} \,\middle|\, x_i \in \mathbb{R}^d\right\}$$

Probabilities on monotone increasing curves.

... Domains are non-convex in general.

Variance maximization s.t. Convex order constraint

For any $\pi \in \mathcal{M}(\mu,\nu)$, since $\mathbb{E}_\pi[XY] = \mathbb{E}_\mu[X\mathbb{E}_\pi[Y|X]] = \mathbb{E}_\mu[|X|^2]$,

$$\begin{aligned}\mathbb{E}_\pi[|X-Y|^2] &= \mathbb{E}_\nu[|Y|^2] - \mathbb{E}_\mu[|X|^2] \\ &= \mathrm{Var}(\nu) - \mathrm{Var}(\mu) \quad \text{if} \;\; \mu,\nu \in \mathcal{P}_0.\end{aligned}$$

Since the (empirical) data ν is given and fixed, the problem

$$\min_{\mu\in\mathcal{D}} \min_{\pi\in\mathcal{M}(\mu,\nu)} \mathbb{E}_\pi[|X-Y|^2]$$

can be equivalently formulated as

$$\max_{\mu\in\mathcal{D},\, \mu\preceq_{\mathsf{c}}\nu} \mathrm{Var}(\mu) \tag{2}$$

μ,ν are in convex order $\Leftrightarrow \mu \preceq_{\mathsf{c}} \nu \Leftrightarrow \int f d\mu \le \int f d\nu \;\forall$convex function f

$\Leftrightarrow \mathcal{M}(\mu,\nu)$ is nonempty (Strassen's theorem)

Existence of solutions and Convergence as noise vanishes

Theorem 1. i) If $\mathcal{D} \subseteq \mathcal{P}(\mathbb{R}^d)$ is W_2-closed, then (2) attains a solution.
ii) Let μ^* be a solution to (2). Then for any $\rho \in \mathcal{D}$ with $\rho \preceq_{\mathsf{c}} \nu$, we have

$$W_2(\mu^*, \rho) \leq \sqrt{\mathrm{Var}(\nu) - \mathrm{Var}(\rho)} + W_2(\nu, \rho).$$

Consequently, $W_2(\mu^*, \rho) \to 0$ as $W_2(\nu, \rho) \to 0$, i.e., as the noise vanishes.

Existence of solutions and Convergence as noise vanishes

Theorem 1. i) If $\mathcal{D} \subseteq \mathcal{P}(\mathbb{R}^d)$ is W_2-closed, then (2) attains a solution.
ii) Let μ^* be a solution to (2). Then for any $\rho \in \mathcal{D}$ with $\rho \preceq_{\mathsf{c}} \nu$, we have

$$W_2(\mu^*, \rho) \leq \sqrt{\mathrm{Var}(\nu) - \mathrm{Var}(\rho)} + W_2(\nu, \rho).$$

Consequently, $W_2(\mu^*, \rho) \to 0$ as $W_2(\nu, \rho) \to 0$, i.e., as the noise vanishes.

Proof. i) The set $\mathcal{M}_\nu = \{\mu \,|\, \mu \preceq_{\mathsf{c}} \nu\}$ is W_2-compact, so is $\mathcal{D} \cap \mathcal{M}_\nu$.

ii) Recall $W_2(\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} \sqrt{\mathbb{E}_\pi[|X - Y|^2]}$.

$$\begin{aligned} W_2(\mu^*, \rho) &\leq W_2(\mu^*, \nu) + W_2(\nu, \rho) \\ &\leq \sqrt{\mathbb{E}_\pi[|X - Y|^2]} + W_2(\nu, \rho) \quad \text{for any } \pi \in \mathcal{M}(\mu^*, \nu) \\ &= \sqrt{\mathrm{Var}(\nu) - \mathrm{Var}(\mu^*)} + W_2(\nu, \rho) \\ &\leq \sqrt{\mathrm{Var}(\nu) - \mathrm{Var}(\rho)} + W_2(\nu, \rho). \end{aligned}$$

□

A weak version of the convex order

In the problem $\max_{\mu \in \mathcal{D},\, \mu \preceq_{\mathsf{c}} \nu} \mathrm{Var}(\mu)$, it is difficult to check $\mu \preceq_{\mathsf{c}} \nu$ if $d \geq 2$.

$\Longrightarrow$ We introduce a weaker version of convex order.

A weak version of the convex order

In the problem $\max_{\mu\in\mathcal{D},\, \mu\preceq_{\mathsf{C}}\nu} \mathrm{Var}(\mu)$, it is difficult to check $\mu \preceq_{\mathsf{C}} \nu$ if $d \geq 2$.

$\implies$ We introduce a weaker version of convex order.

Definition. We say $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ are in Kantorovich order, $\mu \preceq_{\mathsf{K}} \nu$, if

$$\max_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_\pi\langle X, Y - X\rangle \geq 0 \iff K(\mu,\nu) \geq \mathbb{E}_\mu|X|^2$$

$$\iff W_2(\mu,\nu)^2 \leq \mathbb{E}_\nu|Y|^2 - \mathbb{E}_\mu|X|^2$$

where $K(\mu,\nu) := \max_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_\pi\langle X, Y\rangle = \max_{\pi\in\Pi(\mu,\nu)} \int \langle x, y\rangle d\pi(x,y).$

Note. $\mu \preceq_{\mathsf{C}} \nu \implies \mu \preceq_{\mathsf{K}} \nu$.

$\implies$ We consider the problem $\max_{\mu\in\mathcal{D},\, \mu\preceq_{\mathsf{K}}\nu} \mathrm{Var}(\mu)$.

Properties of the Kantorovich order

Set $\mathcal{M}_\nu^K = \{\mu \,|\, \mu \preceq_\mathsf{K} \nu\}$. Given $\mathcal{D} \cup \{\nu\} \subseteq \mathcal{P}_0$, we consider the problem

$$\max_{\mu \in \mathcal{D} \cap \mathcal{M}_\nu^K} \mathrm{Var}(\mu). \tag{3}$$

$\mathcal{M}_\nu^K$ is convex, weakly compact, W_2-closed, but not W_2-compact

$\implies$ existence of solution is assured if e.g. $\mathcal{D}$ is W_2-compact.

Properties of the Kantorovich order

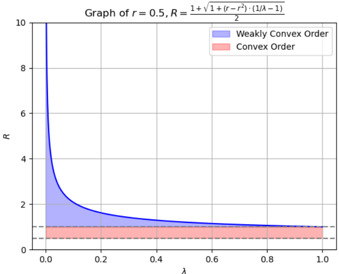
Set $\mathcal{M}_\nu^K = \{\mu \,|\, \mu \preceq_\kappa \nu\}$. Given $\mathcal{D} \cup \{\nu\} \subseteq \mathcal{P}_0$, we consider the problem

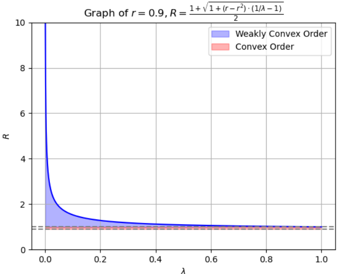
$$\max_{\mu \in \mathcal{D} \cap \mathcal{M}_\nu^K} \mathrm{Var}(\mu). \tag{3}$$

$\mathcal{M}_\nu^K$ is convex, weakly compact, W_2-closed, but not W_2-compact
$\implies$ existence of solution is assured if e.g. $\mathcal{D}$ is W_2-compact.

Ex. $\sigma_r =$ uniform probability over a centered sphere in $\mathbb{R}^d$ with radius r.
Let $\nu = \sigma_1$, and $\mu = (1-\lambda)\sigma_r + \lambda\sigma_R$ for $0 \le r \le R$ and $\lambda \in [0,1]$. Then

$$\mu \preceq_\kappa \nu \iff R \le \frac{1+\sqrt{1+4(r-r^2)(\frac{1}{\lambda}-1)}}{2} \text{ for each } r \in [0,1],\ \lambda \in (0,1).$$





Relationship with principal component analysis (PCA)

V_m = set of m-dimensional subspaces of $\mathbb{R}^d$. Consider the domain

$$\mathcal{D}_m = \{\mu \in \mathcal{P}_0 \mid \mu(L) = 1 \text{ for some } L \in V_m\}.$$

Theorem 2. For $L \in V_m$, the solution to the problem $\max\limits_{\mu \preceq_{\mathsf{K}} \nu,\, \mu(L)=1} \mathrm{Var}(\mu)$ is uniquely given by the orthogonal projection (push-forward) of ν onto L.

Relationship with principal component analysis (PCA)

V_m = set of m-dimensional subspaces of $\mathbb{R}^d$. Consider the domain

$$\mathcal{D}_m = \{\mu \in \mathcal{P}_0 \,|\, \mu(L) = 1 \text{ for some } L \in V_m\}.$$

Theorem 2. For $L \in V_m$, the solution to the problem $\max\limits_{\mu \preceq_{\mathsf{K}} \nu,\, \mu(L)=1} \mathrm{Var}(\mu)$ is uniquely given by the orthogonal projection (push-forward) of ν onto L.

$\Rightarrow$ PCA is a special case of the weak formulation (3) wrt the domain $\mathcal{D}_1$. To see this, we recall that the first principal component is defined as a direction that maximizes the variance of the projected data. Theorem 2 shows that the first principal component can be given by any $L_1 \in V_1$ satisfying $\mu_1(L_1) = 1$ for some μ_1 solving the problem $\max\limits_{\mu \in \mathcal{D}_1,\, \mu \preceq_{\mathsf{K}} \nu} \mathrm{Var}(\mu)$, in which case μ_1 is the orthogonal projection of the data ν onto L_1.

Relationship with principal component analysis (PCA)

V_m = set of m-dimensional subspaces of $\mathbb{R}^d$. Consider the domain

$$\mathcal{D}_m = \{\mu \in \mathcal{P}_0 \,|\, \mu(L) = 1 \text{ for some } L \in V_m\}.$$

Theorem 2. For $L \in V_m$, the solution to the problem $\max\limits_{\mu \preceq_{\mathsf{K}} \nu,\, \mu(L)=1} \mathrm{Var}(\mu)$ is uniquely given by the orthogonal projection (push-forward) of ν onto L.

$\Rightarrow$ PCA is a special case of the weak formulation (3) wrt the domain $\mathcal{D}_1$. To see this, we recall that the first principal component is defined as a direction that maximizes the variance of the projected data. Theorem 2 shows that the first principal component can be given by any $L_1 \in V_1$ satisfying $\mu_1(L_1) = 1$ for some μ_1 solving the problem $\max\limits_{\mu \in \mathcal{D}_1,\, \mu \preceq_{\mathsf{K}} \nu} \mathrm{Var}(\mu)$, in which case μ_1 is the orthogonal projection of the data ν onto L_1.

Inductively, given the first $i-1$ principal components $L_1, \dots, L_{i-1}$, the ith principal component L_i is defined as a direction orthogonal to $L_1, \dots, L_{i-1}$ that maximizes the variance of the projected data. Again by Theorem 2, the ith principal component can be given by any $L_i \in V_1$ satisfying $\mu_i(L_i) = 1$ for some μ_i solving the problem $\max\limits_{\mu \in \mathcal{D}_{1,i},\, \mu \preceq_{\mathsf{K}} \nu} \mathrm{Var}(\mu)$, where $\mathcal{D}_{1,i} := \{\mu \in \mathcal{D}_1 \,|\, \exists L \in V_1 \text{ s.t. } L \perp L_j \ \forall j = 1, \dots, i-1 \text{ and } \mu(L) = 1\}$.

Relationship with PCA — nonvanishing noise case

Following Yuxin Chen, Yuejie Chi, Jianqing Fan and Cong Ma (2021), "Spectral Methods for Data Science: A Statistical Perspective", consider

$$Y = L^* W + R$$

where $W \sim \mathcal{N}(0, I_m)$ is an m-dimensional vector of latent factors, $L^* \in \mathbb{R}^{d\times m}$ represents a factor loading matrix that is not known a priori, and $R \sim \mathcal{N}(0, \sigma^2 I_d)$ stands for random noise not explained by W. Assume $L^* = U^* {\Lambda^*}^{1/2}$ and W and R are independent. Let $\nu = \mathcal{L}(Y)$.

Goal) Estimate the subspace spanned by L^* and latent factors W.

$\Rightarrow$ In PCA literature, $\mathrm{Im}(L^*)$ is referred to as the principal subspace.

Relationship with PCA — nonvanishing noise case

Following Yuxin Chen, Yuejie Chi, Jianqing Fan and Cong Ma (2021), "Spectral Methods for Data Science: A Statistical Perspective", consider

$$Y = L^* W + R$$

where $W \sim \mathcal{N}(0, I_m)$ is an m-dimensional vector of latent factors, $L^* \in \mathbb{R}^{d\times m}$ represents a factor loading matrix that is not known a priori, and $R \sim \mathcal{N}(0, \sigma^2 I_d)$ stands for random noise not explained by W. Assume $L^* = U^* {\Lambda^*}^{1/2}$ and W and R are independent. Let $\nu = \mathcal{L}(Y)$.

Goal) Estimate the subspace spanned by L^* and latent factors W.

$\Rightarrow$ In PCA literature, $\mathrm{Im}(L^*)$ is referred to as the principal subspace.

We define the domain $\mathcal{D} = \{\mu_L = \mathcal{L}(LW) \,|\, L = U\Lambda^{1/2}\}$.

Theorem 3. If $\nu_n \xrightarrow{W_2} \nu$, $\exists N$ s.t. $\mathcal{D} \cap \mathcal{M}_{\nu_n}^K \neq \emptyset$ for all $n \geq N$, and for any $\mu_{L_n} \in \underset{\mu \in \mathcal{D}\cap\mathcal{M}_{\nu_n}^K}{\operatorname{argmax}} \mathrm{Var}(\mu)$ with $L_n = U_n \Lambda_n^{1/2}$, we have

$$L_n L_n^T - \sigma_n^2 U_n U_n^T \to L^* {L^*}^T \quad \text{and} \quad \sigma_n^2 \to \sigma^2 \quad \text{as} \quad n \to \infty,$$

where $\sigma_n^2 = \int |y - p_{L_n}(y)|^2 \nu_n(dy)$ is an estimator of noise variance σ^2.

Relationship with k-means clustering

The variance maximization problem s.t. the Kantorovich order represents a different problem than the problem s.t. the convex order, because the set $\{\mu \mid \mu \preceq_{\mathsf{K}} \nu\}$ can be potentially much bigger than $\{\mu \mid \mu \preceq_{\mathsf{c}} \nu\}$.

Relationship with k-means clustering

The variance maximization problem s.t. the Kantorovich order represents a different problem than the problem s.t. the convex order, because the set $\{\mu \,|\, \mu \preceq_{\mathsf{K}} \nu\}$ can be potentially much bigger than $\{\mu \,|\, \mu \preceq_{\mathsf{c}} \nu\}$.

However, we question if we are essentially addressing a different problem.

$$\mathcal{D}^k = \{\mu \,|\, |\operatorname{spt}(\mu)| \le k\} = \left\{\mu = \sum_{i=1}^{k} u_i \delta_{x_i} \,\middle|\, x_i \in \mathbb{R}^d, u_i \ge 0, \sum_{i=1}^{k} u_i = 1\right\}.$$

Relationship with k-means clustering

The variance maximization problem s.t. the Kantorovich order represents a different problem than the problem s.t. the convex order, because the set $\{\mu \mid \mu \preceq_{\mathsf{K}} \nu\}$ can be potentially much bigger than $\{\mu \mid \mu \preceq_{\mathsf{c}} \nu\}$.

However, we question if we are essentially addressing a different problem.

$$\mathcal{D}^k = \left\{\mu \,\middle|\, |\operatorname{spt}(\mu)| \le k\right\} = \left\{\mu = \sum_{i=1}^{k} u_i \delta_{x_i} \,\middle|\, x_i \in \mathbb{R}^d, u_i \ge 0, \sum_{i=1}^{k} u_i = 1\right\}.$$

Theorem 4. If $\mathcal{D} = \mathcal{D}^k$ or $\mathcal{D}^k_{\mathrm{Uni}}$, every optimizer μ for $\max\limits_{\mu \in \mathcal{D},\, \mu \preceq_{\mathsf{K}} \nu} \operatorname{Var}(\mu)$ satisfies $\mu \preceq_{\mathsf{c}} \nu$, and hence, solves the original problem $\max\limits_{\mu \in \mathcal{D},\, \mu \preceq_{\mathsf{c}} \nu} \operatorname{Var}(\mu)$.

Relationship with k-means clustering

The variance maximization problem s.t. the Kantorovich order represents a different problem than the problem s.t. the convex order, because the set $\{\mu \mid \mu \preceq_{\mathsf{K}} \nu\}$ can be potentially much bigger than $\{\mu \mid \mu \preceq_{\mathsf{c}} \nu\}$.

However, we question if we are essentially addressing a different problem.

$$\mathcal{D}^k = \{\mu \,|\, |\mathrm{spt}(\mu)| \le k\} = \left\{\mu = \sum_{i=1}^{k} u_i \delta_{x_i} \,\middle|\, x_i \in \mathbb{R}^d, u_i \ge 0, \sum_{i=1}^{k} u_i = 1\right\}.$$

Theorem 4. If $\mathcal{D} = \mathcal{D}^k$ or $\mathcal{D}^k_{\mathrm{Uni}}$, every optimizer μ for $\max_{\mu \in \mathcal{D},\, \mu \preceq_{\mathsf{K}} \nu} \mathrm{Var}(\mu)$ satisfies $\mu \preceq_{\mathsf{c}} \nu$, and hence, solves the original problem $\max_{\mu \in \mathcal{D},\, \mu \preceq_{\mathsf{c}} \nu} \mathrm{Var}(\mu)$.

Corollary. $\max_{\mu \in \mathcal{D}^k,\, \mu \preceq_{\mathsf{K}} \nu} \mathrm{Var}(\mu)$ is equivalent to the k-means problem

$$\min_{\mu \in \mathcal{D}^k} \min_{\pi \in \Pi(\mu,\nu)} \int |x-y|^2 d\pi(x,y).$$

Proof. $\min_{\mu \in \mathcal{D}^k} \min_{\pi \in \Pi(\mu,\nu)} \int |x-y|^2 d\pi = \min_{\mu \in \mathcal{D}^k} \min_{\pi \in \mathcal{M}(\mu,\nu)} \int |x-y|^2 d\pi.$ $\square$

Reformulation into bivariate optimization problem

The formulation $\max_{\mu \in \mathcal{D},\, \mu \preceq_{\mathsf{K}} \nu} \mathrm{Var}(\mu)$ incorporates important data reduction approaches, such as PCA and k-means, by selecting an appropriate $\mathcal{D}$.

$\Rightarrow$ How to solve the problem effectively? Recall that for $\mu, \nu \in \mathcal{P}_0(\mathbb{R}^d)$:

$$\mu \preceq_{\mathsf{K}} \nu \Leftrightarrow K(\mu, \nu) \geq \mathrm{Var}(\mu) \quad \text{where} \quad K(\mu, \nu) = \max_{\pi \in \Pi(\mu, \nu)} \mathbb{E}_\pi \langle X, Y \rangle.$$

Reformulation into bivariate optimization problem

The formulation $\max_{\mu\in\mathcal{D},\, \mu\preceq_{\mathsf{K}}\nu} \mathsf{Var}(\mu)$ incorporates important data reduction approaches, such as PCA and k-means, by selecting an appropriate $\mathcal{D}$.

$\Rightarrow$ How to solve the problem effectively? Recall that for $\mu, \nu \in \mathcal{P}_0(\mathbb{R}^d)$:

$$\mu \preceq_{\mathsf{K}} \nu \Leftrightarrow K(\mu,\nu) \geq \mathsf{Var}(\mu) \quad \text{where} \quad K(\mu,\nu) = \max_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_\pi \langle X, Y\rangle.$$

$\lambda_\# \mu =$ dilation of μ by $\lambda \in \mathbb{R}$. That is, $\mathcal{L}(X) = \mu \implies \lambda_\#\mu = \mathcal{L}(\lambda X)$.

Theorem 5. Assume $\mathcal{D}$ is a cone: $\lambda_\#\mu \in \mathcal{D}$ for any $\mu \in \mathcal{D}$ and $\lambda \geq 0$. Then the problem $\max_{\mu\in\mathcal{D},\, \mu\preceq_{\mathsf{K}}\nu} \mathsf{Var}(\mu)$ is equivalent to

$$\max_{\substack{\xi\in\mathcal{D},\, \mathsf{Var}(\xi)\leq 1\\ \pi\in\Pi(\xi,\nu)}} \mathbb{E}_\pi \langle X, Y\rangle \tag{4}$$

in the sense that for any solution (ξ^*, π^*) to (4), $K(\xi^*,\nu)_\#\xi^*$ solves (3). Conversely, for any solution μ^* of (3), $\left(\frac{1}{\sqrt{\mathsf{Var}(\mu^*)}}_\#\mu^*, \pi^*\right)$ solves (4) with any OT π^* between $\frac{1}{\sqrt{\mathsf{Var}(\mu^*)}}_\#\mu^*$ and ν.

Iterative linear optimization

The constraint $\mu \preceq_\kappa \nu$ has been removed from $\max_{\substack{\xi \in \mathcal{D},\, \mathrm{Var}(\xi) \leq 1 \\ \pi \in \Pi(\xi, \nu)}} \mathbb{E}_\pi \langle X, Y \rangle$

$\Rightarrow$ enables iterative linear optimization in (ξ, π).

$$\text{Set} \quad \nu = \sum_{j=1}^{n} v_j \delta_{y_j}, \quad \mathcal{D}^k = \left\{ \mu = \sum_{i=1}^{k} u_i \delta_{x_i} \,\middle|\, x_i \in \mathbb{R}^d, u_i \geq 0, \sum_{i=1}^{k} u_i = 1 \right\},$$

$$\Pi(\cdot, \nu) := \left\{ \pi = (\pi_{ij})_{\substack{i=1,\ldots,k \\ j=1,\ldots,n}} \,\middle|\, \pi \text{ is a proby matrix with } \sum_{i=1}^{k} \pi_{ij} = v_j \ \forall j \right\}.$$

Iterative linear optimization

If $\mathcal{D} = \mathcal{D}^k$ for example, we may iterate:

Step 1. Given $\pi \in \Pi(\cdot, \nu)$, write $u_i = \sum_j \pi_{ij}$ for $i = 1, ..., k$. Solve

$$\max_{(x_1,\dots,x_k)} \sum_{i,j} \pi_{ij} \langle x_i, y_j \rangle \quad s.t. \quad \sum_i u_i |x_i|^2 = 1, \ \sum_i u_i x_i = 0.$$

Step 2. Given $(x_1, ..., x_k) \in (\mathbb{R}^d)^k$, solve

$$\max_{\pi \in \Pi(\cdot,\nu)} \sum_{i,j} \pi_{ij} \langle x_i, y_j \rangle \quad s.t. \quad u_i = \sum_j \pi_{ij}, \ \sum_i u_i |x_i|^2 = 1, \ \sum_i u_i x_i = 0.$$

$\Rightarrow$ Steps 1 and 2 monotonically increase the objective $\sum_{i,j} \pi_{ij} \langle x_i, y_j \rangle$.

Numeric examples

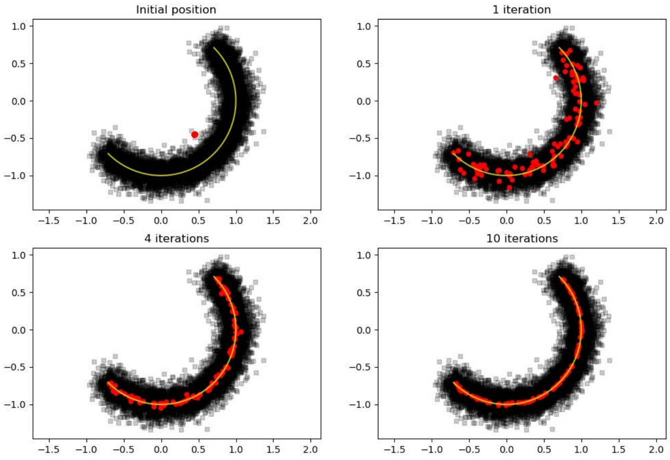


Figure: Convergence towards the prior distribution. $n = 10000$, $k = 100$, $d = 2$.

Numeric examples

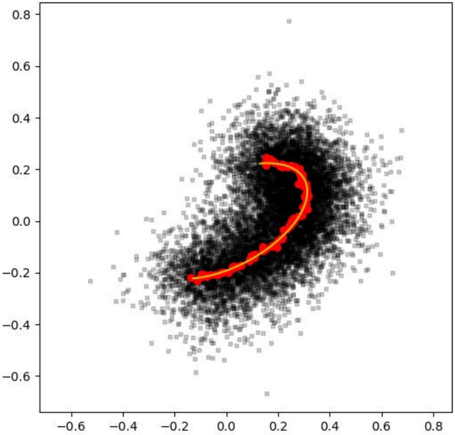


Figure: High-dimension arc example: $n = 10000$, $k = 100$, $d = 30$.

Numeric examples

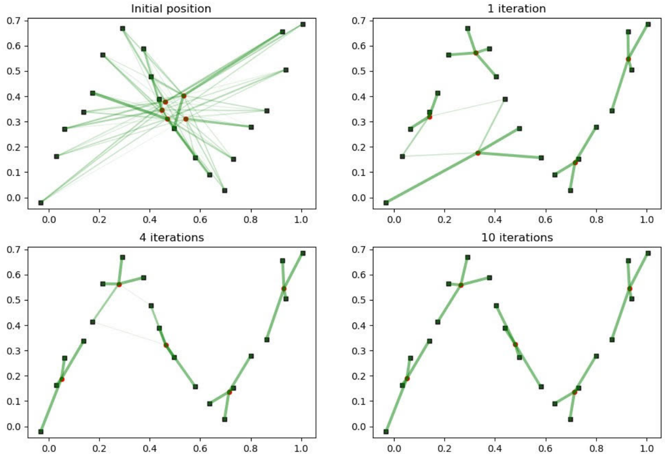


Figure: Zigzag example with transport lines. $n = 20$, $k = 5$, $d = 2$

Numeric examples

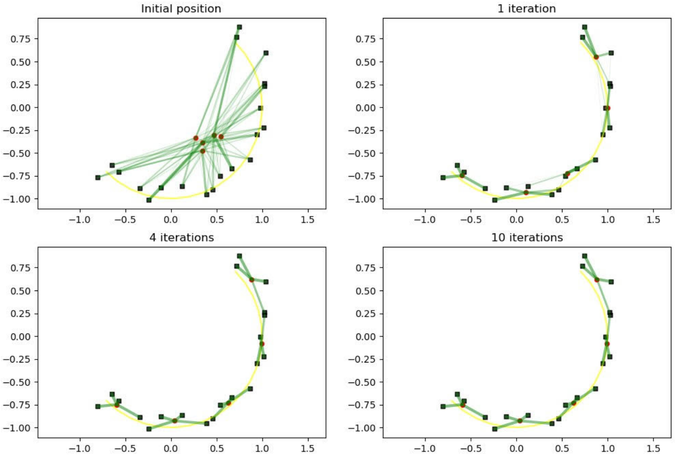


Figure: Arc example with transport lines. $n = 20$, $k = 5$, $d = 2$

Summary

· We propose a denoising approach of data ν in which we maximize variance of the first marginal μ of a martingale coupling $\pi \in \mathcal{M}(\mu, \nu)$

$\iff$ maximize $\mathsf{Var}(\mu)$ for μ dominated by data ν in convex order.

· The approach is adaptable and versatile

$\implies$ Changing the domain $\mathcal{D}$ yields different problems.

· Due to the computational complexity and inflexibility of the convex order, we propose using a weaker domination, the Kantorovich order.

$\implies$ For some domains $\mathcal{D}$, solutions μ under $\mu \preceq_{\mathsf{K}} \nu$ satisfies $\mu \preceq_{\mathsf{c}} \nu$.

$\implies$ $\preceq_{\mathsf{K}}$ allows us to reformulate into a bivariate optimization problem.

· Effective numerical schemes tailored to each domain $\mathcal{D}$ are desired.