On the Convergence of Maximum Variance Unfolding

Ery Arias-Castro

Department of Mathematics,
University of California, San Diego

Joint work with **Bruno Pelletier**, Université Rennes II, France

Partial support from the US National Science Foundation

Dimensionality reduction

The problem of reducing the dimension of the data is a basic problem in statistics and machine (unsupervised) learning.

Dimensionality reduction

The problem of reducing the dimension of the data is a basic problem in statistics and machine (unsupervised) learning.
Suppose we observe points $x_1, \ldots, x_n \in \mathbb{R}^p$. The goal is to reduce

the dimension from p to $d < p$.

By this we mean to find a *meaningful* embedding

$$x_i \in \mathbb{R}^p \to y_i \in \mathbb{R}^d$$

Nonlinear setting

The most classical methods, PCA/MDS, are tailored to the case where $x_1, \dots, x_n$ lie in the vicinity of an affine surface.

Nonlinear setting

The most classical methods, PCA/MDS, are tailored to the case where $x_1, \ldots, x_n$ lie in the vicinity of an affine surface.

In modern settings, we suppose that the data points $x_1, \ldots, x_n$ lie in the vicinity of a curved submanifold, M.

Nonlinear setting

The most classical methods, PCA/MDS, are tailored to the case where $x_1, \ldots, x_n$ lie in the vicinity of an affine surface.

In modern settings, we suppose that the data points $x_1, \ldots, x_n$ lie in the vicinity of a curved submanifold, M.

The **intrinsic distance** on M is defined as:

$$\delta_M(x, x') = \inf\{t : \exists \gamma : [0,t] \to M, \text{ 1-Lipschitz} \\ \text{with } \gamma(0) = x \text{ and } \gamma(t) = x'\}.$$

Nonlinear setting

The most classical methods, PCA/MDS, are tailored to the case where $x_1, \dots, x_n$ lie in the vicinity of an affine surface.

In modern settings, we suppose that the data points $x_1, \dots, x_n$ lie in the vicinity of a curved submanifold, M.

The **intrinsic distance** on M is defined as:

$$\delta_M(x, x') = \inf\{t : \exists \gamma : [0, t] \to M, \text{ 1-Lipschitz} \\ \text{with } \gamma(0) = x \text{ and } \gamma(t) = x'\}.$$

The **goal** is to find an embedding, $x_i \to y_i$, that preserves the intrinsic pairwise distances as much as possible:

$$\delta_M(x_i, x_j) \approx \|y_i - y_j\|, \quad \forall i, j,$$

where $\|\cdot\|$ denotes the Euclidean norm.

Literature

- Self-Organazing Maps, Kohonen (1984)
- Principal Surfaces, Hastie (1984)
- Kernel PCA, Schölkopf, Smola and Muller (1999)
- Isomap, Tenebaum, de Silva and Langford (2000)
- Local Linear Embedding, Roweis and Saul (2000)
- Laplacian Eigenmaps, Belkin and Niyogi (2003)
- Manifold Charting, Brand (2003)
- Hessian Eigenmaps, Donoho and Grimes (2003)
- Local Tangent Space Alignment, Zhang and Zha (2004)
- Max Variance Unfolding, Weinberger and Saul (2004)
- Diffusion Maps, Coifman and Lafon (2006)
- ... and many more!

Convergence guarantees

A number of these methods have some convergence guarantees of some sort, for example:

- Isomap (Bernstein et al, 2000)
- Laplacian Eigenmaps (with some implications for LLE) (Belkin and Niyogi, 2005; Giné and Koltchinskii, 2006; Hein et al., 2005; Singer, 2006; von Luxburg et al., 2008)
- LTSA (Smith, Huo and Zha, 2008)
- A discrete HLLE was analyzed by Ye and Zhi (2012)
- LLE was analyzed by Goldberg and Ritov (2012) for the problem of conformal embedding

Convergence guarantees

A number of these methods have some convergence guarantees of some sort, for example:

- Isomap (Bernstein et al, 2000)
- Laplacian Eigenmaps (with some implications for LLE) (Belkin and Niyogi, 2005; Giné and Koltchinskii, 2006; Hein et al., 2005; Singer, 2006; von Luxburg et al., 2008)
- LTSA (Smith, Huo and Zha, 2008)
- A discrete HLLE was analyzed by Ye and Zhi (2012)
- LLE was analyzed by Goldberg and Ritov (2012) for the problem of conformal embedding

Goldberg et al (2008) prove *negative* results for output-normalized methods such as LLE, Laplacian Eigenmaps, HLLE, LTSA, Diffusion Maps.

Maximum Variance Unfolding (MVU)

Suggested by (Weinberger and Saul, 2004), who also used the name Semidefinite Embedding.

Maximum Variance Unfolding (MVU)

Suggested by (Weinberger and Saul, 2004), who also used the name Semidefinite Embedding.

For $r > 0$, define the (random) set

$$\mathcal{Y}_{n,r} = \Big\{ y_1, \dots, y_n \in \mathbb{R}^p \ : \\ \|y_i - y_j\| \le \|x_i - x_j\| \text{ when } \|x_i - x_j\| \le r \Big\}.$$

Maximum Variance Unfolding (MVU)

Suggested by (Weinberger and Saul, 2004), who also used the name Semidefinite Embedding.

For $r > 0$, define the (random) set

$$\mathcal{Y}_{n,r} = \{y_1, \dots, y_n \in \mathbb{R}^p : \\ \|y_i - y_j\| \leq \|x_i - x_j\| \text{ when } \|x_i - x_j\| \leq r\}.$$

Maximum Variance Unfolding (MVU)

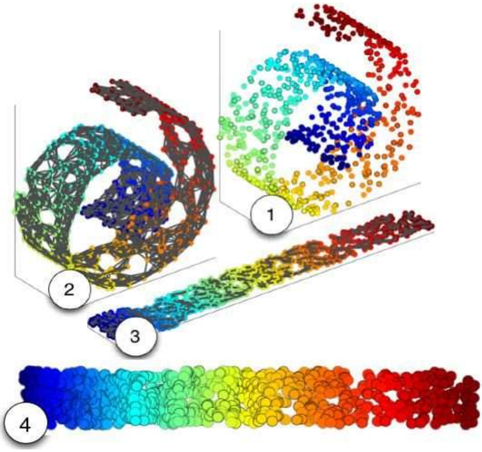
Given a neighborhood radius $r > 0$, maximize

$$\mathcal{E}(Y) := \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \|y_i - y_j\|^2,$$

over

$$Y = (y_1, \dots, y_n) \in \mathcal{Y}_{n,r}.$$

1. Dataset (point cloud)
2. Neighborhood graph
3. Unfolding (variance maximization)
4. Apply PCA (or MDS) to flatten the surface



Taken from (Weinberger and Saul, 2004)

Contribution

Our main contribution is establishing convergence guarantees for Maximum Variance Unfolding (MVU).

We also derive some implications in terms of actual embedding performance.

Contribution

Our main contribution is establishing convergence guarantees for Maximum Variance Unfolding (MVU).

We also derive some implications in terms of actual embedding performance.

We mention the work of Paprotny and Garcke (2012). They establish that, under the assumption that M is geodesically convex, MVU recovers a distance matrix that approximates

$$(\delta_M(x_i, x_j) : i, j = 1, \ldots, n).$$

Continuum Maximum Variance Unfolding

We define a continuum analog to MVU, which will be shown to be the large-sample ($n \to \infty$) limit of the MVU optimization problem defined earlier.

Continuum Maximum Variance Unfolding

We define a continuum analog to MVU, which will be shown to be the large-sample ($n \to \infty$) limit of the MVU optimization problem defined earlier. In the **noiseless** case, assume that $x_1, \ldots, x_n$ are

sampled independently from μ, a (Borel) probability distribution with support a connected and compact set $M \subset \mathbb{R}^p$.

Continuum Maximum Variance Unfolding

We define a continuum analog to MVU, which will be shown to be the large-sample ($n \to \infty$) limit of the MVU optimization problem defined earlier. In the **noiseless** case, assume that $x_1, \ldots, x_n$ are

sampled independently from μ, a (Borel) probability distribution with support a connected and compact set $M \subset \mathbb{R}^p$. Recall the

intrinsic distance on M:

$$\delta_M(x, x') = \inf\{t : \exists \gamma : [0, t] \to M, \text{ 1-Lipschitz with } \gamma(0) = x \text{ and } \gamma(t) = x'\}.$$

For $L > 0$, define the class of L-Lipschitz functions:

$$\mathcal{F}_L = \{ f : M \to \mathbb{R}^p \; : \\ \|f(x) - f(x')\| \le L\, \delta_M(x, x'),\ \forall x, x' \in M \}.$$

For $L > 0$, define the class of L-Lipschitz functions:

$$\mathcal{F}_L = \{ f : M \to \mathbb{R}^p \ : \\ \|f(x) - f(x')\| \le L\,\delta_M(x,x'),\ \forall x, x' \in M \}.$$

Continuum MVU

Maximize

$$\mathcal{E}(f) := \int_{M \times M} \|f(x) - f(x')\|^2 \mu(\mathrm{d}x)\mu(\mathrm{d}x'),$$

over

$$f \in \mathcal{F}_1.$$

Central condition on M

Regularity assumption

There is a non-decreasing function $c : [0, \infty) \to [0, \infty)$ with $c(r) \to 0$ when $r \to 0$, such that, for all $x, x' \in M$,

$$\delta_M(x, x') \leq \big(1 + c(\|x - x'\|)\big)\|x - x'\|.$$

Central condition on M

Regularity assumption

There is a non-decreasing function $c : [0, \infty) \to [0, \infty)$ with $c(r) \to 0$ when $r \to 0$, such that, for all $x, x' \in M$,

$$\delta_M(x, x') \leq \big(1 + c(\|x - x'\|)\big)\|x - x'\|.$$

In a nutshell

$$\delta_M(x, x') \sim \|x - x'\|, \quad \|x - x'\| \to 0.$$

Central condition on M

Regularity assumption

There is a non-decreasing function $c : [0, \infty) \to [0, \infty)$ with $c(r) \to 0$ when $r \to 0$, such that, for all $x, x' \in M$,

$$\delta_M(x, x') \leq \big(1 + c(\|x - x'\|)\big)\|x - x'\|.$$

In a nutshell

$$\delta_M(x, x') \sim \|x - x'\|, \quad \|x - x'\| \to 0.$$

This assumption is also crucial for Isomap to work. Bernstein et al (2000) prove that it holds when M is a compact, smooth and geodesically convex submanifold (e.g., without boundary).

Central condition on M

Regularity assumption

There is a non-decreasing function $c : [0, \infty) \to [0, \infty)$ with $c(r) \to 0$ when $r \to 0$, such that, for all $x, x' \in M$,

$$\delta_M(x, x') \leq \big(1 + c(\|x - x'\|)\big) \|x - x'\|.$$

In a nutshell

$$\delta_M(x, x') \sim \|x - x'\|, \quad \|x - x'\| \to 0.$$

This assumption is also crucial for Isomap to work. Bernstein et al (2000) prove that it holds when M is a compact, smooth and geodesically convex submanifold (e.g., without boundary).

We extend this to compact, smooth submanifolds with smooth boundary, and to tubular neighborhoods of such sets.
(The latter allows us to study noisy settings.)

Convergence of Discrete MVU to Continuum MVU

Let $\mathcal{S}_1$ be the set of solutions to Continuum MVU.

Convergence of Discrete MVU to Continuum MVU

Let $\mathcal{S}_1$ be the set of solutions to Continuum MVU.

Theorem

Let μ be a (Borel) probability distribution with support $M \subset \mathbb{R}^p$, which is connected, compact and satisfies the regularity condition, and assume that $x_1, \ldots, x_n$ are sampled independently from μ.

Convergence of Discrete MVU to Continuum MVU

Let $\mathcal{S}_1$ be the set of solutions to Continuum MVU.

Theorem

Let μ be a (Borel) probability distribution with support $M \subset \mathbb{R}^p$, which is connected, compact and satisfies the regularity condition, and assume that $x_1, \ldots, x_n$ are sampled independently from μ.

Then, for $r_n \to 0$ sufficiently slowly, we have

$$\sup\{\mathcal{E}(Y) : Y \in \mathcal{Y}_{n,r_n}\} \to \sup\{\mathcal{E}(f) : f \in \mathcal{F}_1\},$$

Convergence of Discrete MVU to Continuum MVU

Let $\mathcal{S}_1$ be the set of solutions to Continuum MVU.

Theorem

Let μ be a (Borel) probability distribution with support $M \subset \mathbb{R}^p$, which is connected, compact and satisfies the regularity condition, and assume that $x_1, \ldots, x_n$ are sampled independently from μ.

Then, for $r_n \to 0$ sufficiently slowly, we have

$$\sup\{\mathcal{E}(Y) : Y \in \mathcal{Y}_{n,r_n}\} \to \sup\{\mathcal{E}(f) : f \in \mathcal{F}_1\},$$

and for any solution $\hat{Y}_n = (\hat{y}_1, \ldots, \hat{y}_n)$ of Discrete MVU,

$$\inf_{f \in \mathcal{S}_1} \max_{1 \leq i \leq n} \|\hat{y}_i - f(x_i)\| \to 0,$$

almost surely as $n \to \infty$.

Details: condition on r_n

For $\eta > 0$, define

$$\Lambda(\eta) = \{\forall x \in M, \exists i = 1, \ldots, n : \|x - x_i\| \leq \eta\},$$

which is the event that $x_1, \ldots, x_n$ forms an η-covering of M.

Details: condition on r_n

For $\eta > 0$, define

$$\Lambda(\eta) = \{\forall x \in M, \exists i = 1, \ldots, n : \|x - x_i\| \leq \eta\},$$

which is the event that $x_1, \ldots, x_n$ forms an η-covering of M.

Connectivity requirement

Let $r_n \to 0$ in such a way that

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\Lambda(\lambda_n r_n)^{\mathsf{c}}\right) < \infty, \text{ for some sequence } \lambda_n \to 0.$$

(Such a sequence exists.)

Details: from point sets to functions

Lemma

Assume that $\Lambda(\eta)$ holds for some $\eta \leq r/4$. Then any vector $Y = (y_1, \ldots, y_n) \in \mathcal{Y}_{n,r}$ is of the form $Y = (f(x_1), \ldots, f(x_n))$ for some $f \in \mathcal{F}_{1+6\eta/r}$.

Details: from point sets to functions

Lemma

Assume that $\Lambda(\eta)$ holds for some $\eta \leq r/4$. Then any vector $Y = (y_1, \ldots, y_n) \in \mathcal{Y}_{n,r}$ is of the form $Y = (f(x_1), \ldots, f(x_n))$ for some $f \in \mathcal{F}_{1+6\eta/r}$.

Indeed, we first prove that

$$\|y_i - y_j\| \leq \big(1 + 6\eta/r\big)\delta_M(x_i, x_j), \ \forall i, j.$$

Details: from point sets to functions

Lemma

Assume that $\Lambda(\eta)$ holds for some $\eta \leq r/4$. Then any vector $Y = (y_1, \ldots, y_n) \in \mathcal{Y}_{n,r}$ is of the form $Y = (f(x_1), \ldots, f(x_n))$ for some $f \in \mathcal{F}_{1+6\eta/r}$.

Indeed, we first prove that

$$\|y_i - y_j\| \leq \big(1 + 6\eta/r\big)\delta_M(x_i, x_j), \ \forall i, j.$$

We then apply Kirszbraun's Extension to the $(1 + 6\eta/r)$-Lipschitz function $g : \{y_1, \ldots, y_n\} \to \mathbb{R}^d$, extending it into $f : M \to \mathbb{R}^d$, also $(1 + 6\eta/r)$-Lipschitz.

Hence,

$$\sup_{Y \in \mathcal{Y}_{n,r}} \mathcal{E}(Y) \leq \sup_{f \in \mathcal{F}_{1+6\eta/r}} \mathcal{E}(Y_n(f)).$$

where $Y_n(f) := (f(x_1), \ldots, f(x_n))$.

Hence,

$$\sup_{Y \in \mathcal{Y}_{n,r}} \mathcal{E}(Y) \leq \sup_{f \in \mathcal{F}_{1+6\eta/r}} \mathcal{E}(Y_n(f)).$$

where $Y_n(f) := (f(x_1), \ldots, f(x_n))$.

(Recall the function c from the regularity condition.)

Hence,

$$\sup_{Y \in \mathcal{Y}_{n,r}} \mathcal{E}(Y) \leq \sup_{f \in \mathcal{F}_{1+6\eta/r}} \mathcal{E}(Y_n(f)).$$

where $Y_n(f) := (f(x_1), \ldots, f(x_n))$.

(Recall the function c from the regularity condition.)

Take $f \in \mathcal{F}_{1-c(r)}$. For any i, j such that $\|x_i - x_j\| \leq r$,

$$\begin{aligned} \|f(x_i) - f(x_j)\| &\leq (1 - c(r))\delta_M(x_i, x_j) \\ &\leq (1 - c(r))(1 + c(r))\|x_i - x_j\| \\ &\leq \|x_i - x_j\| \end{aligned}$$

Hence,

$$\sup_{Y \in \mathcal{Y}_{n,r}} \mathcal{E}(Y) \leq \sup_{f \in \mathcal{F}_{1+6\eta/r}} \mathcal{E}(Y_n(f)).$$

where $Y_n(f) := (f(x_1), \ldots, f(x_n))$.

(Recall the function c from the regularity condition.)

Take $f \in \mathcal{F}_{1-c(r)}$. For any i, j such that $\|x_i - x_j\| \leq r$,

$$\begin{aligned} \|f(x_i) - f(x_j)\| &\leq (1 - c(r))\delta_M(x_i, x_j) \\ &\leq (1 - c(r))(1 + c(r))\|x_i - x_j\| \\ &\leq \|x_i - x_j\| \end{aligned}$$

Consequently, $Y_n(f) \in \mathcal{Y}_{n,r}$, and we conclude that

$$\sup_{Y \in \mathcal{Y}_{n,r}} \mathcal{E}(Y) \geq \sup_{f \in \mathcal{F}_{1-c(r)}} \mathcal{E}(Y_n(f)).$$

After some elementary steps:

$$\left|\sup_{Y\in\mathcal{Y}_{n,r}}\mathcal{E}(Y)-\sup_{f\in\mathcal{F}_1}\mathcal{E}(f)\right| \leq (1+\beta(r,\eta))\sup_{f\in\mathcal{F}_1}\left|\mathcal{E}(Y_n(f))-\mathcal{E}(f)\right| + 3\beta(r,\eta)\,\mathrm{diam}(M)^2,$$

when $\beta(r,\eta):=\max(c(r),6\eta/r)$ is small enough.

Regularity of the energy

For any f and g in $\mathcal{F}_1$, and any x and x' in M:

$$\begin{aligned}
& \left| \|f(x) - f(x')\|^2 - \|g(x) - g(x')\|^2 \right| \\
\leq\ & \|f(x) - f(x') - g(x) + g(x')\| \\
& \times \|f(x) - f(x') + g(x) - g(x')\| \\
\leq\ & \left[\|f(x) - g(x)\| + \|f(x') - g(x')\| \right] \\
& \times \left[\|f(x) - f(x')\| + \|g(x) - g(x')\| \right] \\
\leq\ & 4\|f - g\|_\infty \operatorname{diam}(M).
\end{aligned}$$

Regularity of the energy

For any f and g in $\mathcal{F}_1$, and any x and x' in M:

$$\begin{aligned}
& \left| \|f(x) - f(x')\|^2 - \|g(x) - g(x')\|^2 \right| \\
\leq \quad & \|f(x) - f(x') - g(x) + g(x')\| \\
& \times \|f(x) - f(x') + g(x) - g(x')\| \\
\leq \quad & \left[\|f(x) - g(x)\| + \|f(x') - g(x')\| \right] \\
& \times \left[\|f(x) - f(x')\| + \|g(x) - g(x')\| \right] \\
\leq \quad & 4\|f - g\|_\infty \operatorname{diam}(M).
\end{aligned}$$

Hence,

$$\left| \mathcal{E}(f) - \mathcal{E}(g) \right| \leq 4\|f - g\|_\infty \operatorname{diam}(M), \tag{1}$$

and

$$\left| \mathcal{E}(Y_n(f)) - \mathcal{E}(Y_n(g)) \right| \leq 4\|f - g\|_\infty \operatorname{diam}(M). \tag{2}$$

Finite ε-entropy

Fix $x_0 \in M$ and define

$$\mathcal{F}_1^0 = \{f \in \mathcal{F}_1 : f(x_0) = 0\}.$$

We have

$$\sup_{f \in \mathcal{F}_1} \left| \mathcal{E}(Y_n(f)) - \mathcal{E}(f) \right| = \sup_{f \in \mathcal{F}_1^0} \left| \mathcal{E}(Y_n(f)) - \mathcal{E}(f) \right|.$$

$\mathcal{F}_1^0$ has finite ε-entropy for $\| \cdot \|_\infty$ by the Arzelà-Ascoli Theorem.

$\mathcal{F}_1^0$ has finite ε-entropy for $\|\cdot\|_\infty$ by the Arzelà-Ascoli Theorem.

This and the Law of Large Numbers imply that

$$\sup_{f\in\mathcal{F}_1} \left|\mathcal{E}(Y_n(f)) - \mathcal{E}(f)\right| \to^P 0, \quad n \to \infty. \tag{3}$$

$\mathcal{F}_1^0$ has finite ε-entropy for $\|\cdot\|_\infty$ by the Arzelà-Ascoli Theorem.

This and the Law of Large Numbers imply that

$$\sup_{f\in\mathcal{F}_1}\left|\mathcal{E}(Y_n(f))-\mathcal{E}(f)\right| \to^P 0, \quad n\to\infty. \tag{3}$$

For an almost sure convergence, we use a concentration bound of U-processes (e.g., Hoeffding's) and the Borel-Cantelli Lemma.

Convergence in solution

Let $(\hat{Y}_n) \subset \mathcal{Y}_{n,r_n}$ be a sequence of solutions to Discrete MVU.

Convergence in solution

Let $(\hat{Y}_n) \subset \mathcal{Y}_{n,r_n}$ be a sequence of solutions to Discrete MVU.

There is $\hat{f}_n \in \mathcal{F}^0_{(1+6\eta_n/r_n)}$ such that $\hat{Y}_n = Y_n(\hat{f}_n)$.

Note that $\hat{f}_n \in \mathcal{F}^0_2$ eventually as $n \to \infty$.

Convergence in solution

Let $(\hat{Y}_n) \subset \mathcal{Y}_{n,r_n}$ be a sequence of solutions to Discrete MVU.

There is $\hat{f}_n \in \mathcal{F}^0_{(1+6\eta_n/r_n)}$ such that $\hat{Y}_n = Y_n(\hat{f}_n)$.

Note that $\hat{f}_n \in \mathcal{F}^0_2$ eventually as $n \to \infty$.

Since $\mathcal{F}^0_2$ is compact for $\| \cdot \|_\infty$, it suffices to prove that any accumulation point of $(\hat{f}_n)$ is in $\mathcal{S}^0_1 = \mathcal{F}^0_1 \cap \mathcal{S}_1$.

This is due to the Lipschitz property of $\mathcal{E}$, (1)-(2), and the uniform convergence (3).

Recovering an isometric embedding

Isometry assumption

We assume that M is isometric to a compact, connected domain $D \subset \mathbb{R}^d$, so that there is a bijection $\psi : M \to D$ satisfying

$$\delta_D(\psi(x), \psi(x')) = \delta_M(x, x'),\ \forall x, x' \in M.$$

Recovering an isometric embedding

Isometry assumption

We assume that M is isometric to a compact, connected domain $D \subset \mathbb{R}^d$, so that there is a bijection $\psi : M \to D$ satisfying

$$\delta_D(\psi(x), \psi(x')) = \delta_M(x, x'),\ \forall x, x' \in M.$$

The goal becomes to recover ψ up to a rigid transformation.

Consistency under convexity

Theorem

Suppose that M is isometric to a convex subset $D \subset \mathbb{R}^d$ with isometry mapping $\psi : M \to D$, and that μ is comparable to the uniform distribution on M, in that there is $\alpha > 0$ such that

$$\mu(B(x,\eta)) \geq \alpha \,\mathrm{vol}_d(B(x,\eta) \cap M), \quad \forall x \in M, \forall \eta > 0.$$

Consistency under convexity

Theorem

Suppose that M is isometric to a convex subset $D \subset \mathbb{R}^d$ with isometry mapping $\psi : M \to D$, and that μ is comparable to the uniform distribution on M, in that there is $\alpha > 0$ such that

$$\mu(B(x,\eta)) \geq \alpha \, \mathrm{vol}_d(B(x,\eta) \cap M), \quad \forall x \in M, \forall \eta > 0.$$

Then

$$\mathcal{S}_1 = \{\zeta \circ \psi \,:\, \zeta \in \mathrm{Isom}(\mathbb{R}^p)\},$$

meaning Continuum MVU recovers ψ up to a rigid transformation.

Consistency under convexity

Theorem

Suppose that M is isometric to a convex subset $D \subset \mathbb{R}^d$ with isometry mapping $\psi : M \to D$, and that μ is comparable to the uniform distribution on M, in that there is $\alpha > 0$ such that

$$\mu(B(x,\eta)) \geq \alpha \, \mathrm{vol}_d(B(x,\eta) \cap M), \quad \forall x \in M, \forall \eta > 0.$$

Then

$$\mathcal{S}_1 = \{\zeta \circ \psi \,:\, \zeta \in \mathrm{Isom}(\mathbb{R}^p)\},$$

meaning Continuum MVU recovers ψ up to a rigid transformation.

Note that Isomap is also successful under similar conditions (Bernstein et al, 2000).

Proof. Since D is convex,

$$\delta_D(z, z') = \|z - z'\|, \quad \forall z, z' \in D.$$

Proof. Since D is convex,

$$\delta_D(z, z') = \|z - z'\|, \quad \forall z, z' \in D.$$

Hence, for all f in $\mathcal{F}_1$, we have

$$\begin{aligned}
\mathcal{E}(f) &= \int_{M\times M} \|f(x) - f(x')\|^2 \mu(\mathrm{d}x)\mu(\mathrm{d}x') \\
&\leq \int_{M\times M} \delta_M(x, x')^2 \mu(\mathrm{d}x)\mu(\mathrm{d}x') \\
&= \int_{M\times M} \delta_D(\psi(x), \psi(x'))^2 \mu(\mathrm{d}x)\mu(\mathrm{d}x') \\
&= \int_{M\times M} \|\psi(x) - \psi(x')\|^2 \mu(\mathrm{d}x)\mu(\mathrm{d}x') \\
&= \mathcal{E}(\psi).
\end{aligned}$$

Since $\psi \in \mathcal{F}_1$, we have $\psi \in \mathcal{S}_1$, meaning,

$$\sup_{f \in \mathcal{F}_1} \mathcal{E}(f) = \mathcal{E}(\psi),$$

And since $\mathcal{E}(\zeta \circ \psi) = \mathcal{E}(\psi)$ for any isometry $\zeta : \mathbb{R}^d \to \mathbb{R}^d$,

$$\{\zeta \circ \psi \, : \, \zeta \in \mathrm{Isom}(\mathbb{R}^d)\} \subset \mathcal{S}_1.$$

Take $f \in \mathcal{F}_1$ which is not an isometry. Then there exists two points x and x' in M such that

$$\|f(x) - f(x')\| < \delta_M(x, x').$$

Take $f \in \mathcal{F}_1$ which is not an isometry. Then there exists two points x and x' in M such that

$$\|f(x) - f(x')\| < \delta_M(x, x').$$

By continuity of f, there exists a nonempty open subset U of $M \times M$ containing (x, x') such that

$$\|f(z) - f(z')\| < \delta_M(z, z'), \quad \forall (z, z') \in U.$$

In addition, $\mu(U) > 0$ by the assumption on μ.

Consequently

$$\begin{aligned}
\mathcal{E}(f) &= \int_{M\times M\setminus U} \|f(x)-f(x')\|^2\mu(\mathrm{d}x)\mu(\mathrm{d}x') \\
&\qquad + \int_U \|f(x)-f(x')\|^2\mu(\mathrm{d}x)\mu(\mathrm{d}x') \\
&< \int_{M\times M} \delta_M(x,x')^2\mu(\mathrm{d}x)\mu(\mathrm{d}x') \\
&= \sup_{f\in\mathcal{F}_1} \mathcal{E}(f).
\end{aligned}$$

Consequently

$$\begin{aligned}\mathcal{E}(f) &= \int_{M\times M\setminus U} \|f(x)-f(x')\|^2\mu(\mathrm{d}x)\mu(\mathrm{d}x') \\ &\qquad + \int_U \|f(x)-f(x')\|^2\mu(\mathrm{d}x)\mu(\mathrm{d}x') \\ &< \int_{M\times M} \delta_M(x,x')^2\mu(\mathrm{d}x)\mu(\mathrm{d}x') \\ &= \sup_{f\in\mathcal{F}_1} \mathcal{E}(f).\end{aligned}$$

So any function f in $\mathcal{F}_1$ which is not an isometry onto its image does not belong to $\mathcal{S}_1$.

Since for any isometry f in $\mathcal{S}_1$, $f \circ \psi^{-1} : D \to \mathbb{R}^d$ is an isometry, it can be extended to an isometry ζ of $\mathbb{R}^d$, so that $f = \zeta \circ \psi$, and we conclude that

$$\{\zeta \circ \psi \,:\, \zeta \in \mathrm{Isom}(\mathbb{R}^p)\} = \mathcal{S}_1.$$

Robustness to noise

When the setting is noisy, with noise level $\sigma \geq 0$, $x_1, \ldots, x_n$ are sampled from μ_σ, a (Borel) probability distribution on $\mathbb{R}^p$ with support

$$M_\sigma := \bar{B}(M, \sigma) = \{x \in \mathbb{R}^p : \mathrm{dist}(x, M) \leq \sigma\}.$$

Robustness to noise

When the setting is noisy, with noise level $\sigma \geq 0$, $x_1, \ldots, x_n$ are sampled from μ_σ, a (Borel) probability distribution on $\mathbb{R}^p$ with support

$$M_\sigma := \bar{B}(M, \sigma) = \{x \in \mathbb{R}^p : \mathrm{dist}(x, M) \leq \sigma\}.$$

The reach of a set $M \subset \mathbb{R}^p$ is defined as the largest σ such that, for any $x \in B(M, \sigma)$ there is a unique point $m \in M$ closest to x.

Robustness to noise

When the setting is noisy, with noise level $\sigma \geq 0$, $x_1, \ldots, x_n$ are sampled from μ_σ, a (Borel) probability distribution on $\mathbb{R}^p$ with support

$$M_\sigma := \bar{B}(M, \sigma) = \{x \in \mathbb{R}^p : \mathrm{dist}(x, M) \leq \sigma\}.$$

The reach of a set $M \subset \mathbb{R}^p$ is defined as the largest σ such that, for any $x \in B(M, \sigma)$ there is a unique point $m \in M$ closest to x.

Our perturbation analysis is based on the fact that $\mathcal{E}$ is continuous with respect to the noise level. This immediately implies that MVU is tolerant to noise.

Let $\mathcal{F}_{1,\sigma}$ denote the class of 1-Lipschitz functions on M_σ, and $\mathcal{S}_{1,\sigma}$ the set of solutions of Continuum MVU for M_σ.

Let $\mathcal{F}_{1,\sigma}$ denote the class of 1-Lipschitz functions on M_σ, and $\mathcal{S}_{1,\sigma}$ the set of solutions of Continuum MVU for M_σ.

Lemma

Let $M \subset \mathbb{R}^p$ be of positive reach and assume that $\mu_\sigma \to \mu_0$ weakly when $\sigma \to 0$. Then as $\sigma \to 0$, we have

$$\sup_{f \in \mathcal{F}_{1,\sigma}} \mathcal{E}_\sigma(f) \to \sup_{f \in \mathcal{F}_1} \mathcal{E}(f),$$

and

$$\sup_{f \in \mathcal{S}_{1,\sigma}} \inf_{g \in \mathcal{S}_1} \sup_{x \in M_\sigma} \inf_{z \in M} \|f(x) - g(z)\| \to 0.$$

Inconsistencies

We provide two emblematic situations where MVU fails to recover D. They are both consequences of MVU's robustness to noise.

Inconsistencies

We provide two emblematic situations where MVU fails to recover D. They are both consequences of MVU's robustness to noise.

In both cases, we consider the simplest situation where $M = D \subset \mathbb{R}^2$ (so that $\psi = \mathrm{id}$) and μ is the uniform distribution.

Nonconvex without holes

Suppose $M_0 \subset \mathbb{R}^2$ is a curve homeomorphic to a line segment, but not a line segment, and for $\sigma > 0$, let $M_\sigma = \bar{B}(M_0, \sigma)$.

Nonconvex without holes

Suppose $M_0 \subset \mathbb{R}^2$ is a curve homeomorphic to a line segment, but not a line segment, and for $\sigma > 0$, let $M_\sigma = \bar{B}(M_0, \sigma)$.

Then there is a numeric constant $\sigma_0 > 0$ such that, when $\sigma < \sigma_0$, ψ does not maximize the energy $\mathcal{E}_\sigma$.

Nonconvex without holes

Suppose $M_0 \subset \mathbb{R}^2$ is a curve homeomorphic to a line segment, but not a line segment, and for $\sigma > 0$, let $M_\sigma = \bar{B}(M_0, \sigma)$.

Then there is a numeric constant $\sigma_0 > 0$ such that, when $\sigma < \sigma_0$, ψ does not maximize the energy $\mathcal{E}_\sigma$.

To see this, use the lemma to get $\mathcal{S}_{1,\sigma} \to \mathcal{S}_{1,0}$.

Then notice that $\psi \notin \mathcal{S}_{1,0}$, because $\mathcal{S}_{1,0}$ is made of all the functions that map M to a line segment isometrically.

So there is $\sigma_0 > 0$ such that $\psi \notin \mathcal{S}_{1,\sigma}$ for all $\sigma < \sigma_0$. This also implies that no rigid transformation of $\mathbb{R}^2$ is part of $\mathcal{S}_{1,\sigma}$.

Nonconvex without holes

Suppose $M_0 \subset \mathbb{R}^2$ is a curve homeomorphic to a line segment, but not a line segment, and for $\sigma > 0$, let $M_\sigma = \bar{B}(M_0, \sigma)$.

Then there is a numeric constant $\sigma_0 > 0$ such that, when $\sigma < \sigma_0$, ψ does not maximize the energy $\mathcal{E}_\sigma$.

To see this, use the lemma to get $\mathcal{S}_{1,\sigma} \to \mathcal{S}_{1,0}$.

Then notice that $\psi \notin \mathcal{S}_{1,0}$, because $\mathcal{S}_{1,0}$ is made of all the functions that map M to a line segment isometrically.

So there is $\sigma_0 > 0$ such that $\psi \notin \mathcal{S}_{1,\sigma}$ for all $\sigma < \sigma_0$. This also implies that no rigid transformation of $\mathbb{R}^2$ is part of $\mathcal{S}_{1,\sigma}$.

If we now let $D = M = M_\sigma$ for some $0 < \sigma < \sigma_0$, we see that we do not recover D up to a rigid transformation.

Convex boundary and convex hole

Let K_a denote the symmetric, axis-aligned ellipse of $\mathbb{R}^2$ with semi-major axis length equal to a and perimeter equal to 2π.

Necessarily, $1 \leq a < \pi/2$, with the extreme cases being the unit circle ($a = 1$) and the interval $[-\pi/2, \pi/2]$ swept twice ($a = \pi/2$).

Convex boundary and convex hole

Let K_a denote the symmetric, axis-aligned ellipse of $\mathbb{R}^2$ with semi-major axis length equal to a and perimeter equal to 2π.

Necessarily, $1 \leq a < \pi/2$, with the extreme cases being the unit circle ($a = 1$) and the interval $[-\pi/2, \pi/2]$ swept twice ($a = \pi/2$).

Denote by $b = b(a)$ the semi-minor axis length of K_a, implicitly defined by

$$\int_0^{2\pi} \sqrt{a^2 \sin^2 t + b^2 \cos^2 t}\, \mathrm{d}t = 2\pi.$$

We have

$$\begin{aligned} F(a) &:= \int_{K_a} \|x\|^2 \mathrm{d}x \\ &= \int_0^{2\pi} \left(a^2 \cos^2 t + b^2 \sin^2 t\right) \sqrt{a^2 \sin^2 t + b^2 \cos^2 t}\, \mathrm{d}t. \end{aligned}$$

Note that

$$F(1) = 2\pi \text{ (circle)} \quad > \quad F(\pi/2) = \pi^2/12 \text{ (interval).}$$

Since F is continuous, there is $a_\star$ such that $F(a) < F(1)$ for $a > a_\star$. (We actually believe that $a_\star = 1$.)

Note that

$$F(1) = 2\pi \text{ (circle)} \quad > \quad F(\pi/2) = \pi^2/12 \text{ (interval).}$$

Since F is continuous, there is $a_\star$ such that $F(a) < F(1)$ for $a > a_\star$. (We actually believe that $a_\star = 1$.)

Fix $a \in (a_\star, \pi/2)$ and let $M_0 = K_a = \phi^{-1}(K_1)$, where

$$\phi : \mathbb{R}^2 \to \mathbb{R}^2, \quad \phi(x_1, x_2) = (x_1/a, x_2/b(a)).$$

Note that $\psi = \mathrm{id} \notin \mathcal{S}_{1,0}$, since

$$\begin{aligned}
\mathcal{E}_0(\psi) &= \frac{1}{(2\pi)^2}\int_{M_0\times M_0}\|x-x'\|^2 \mathrm{d}x\mathrm{d}x' \\
&= \frac{1}{\pi}\int_{M_0}\|x\|^2\mathrm{d}x \qquad \text{(by symmetry)} \\
&= \frac{1}{\pi}F(a) \\
&< \frac{1}{\pi}F(1) \\
&= \frac{1}{\pi}\int_{M_0}\|\phi(x)\|^2\mathrm{d}x \\
&= \mathcal{E}_0(\phi).
\end{aligned}$$

Note that $\psi = \mathrm{id} \notin \mathcal{S}_{1,0}$, since

$$\begin{aligned}
\mathcal{E}_0(\psi) &= \frac{1}{(2\pi)^2} \int_{M_0 \times M_0} \|x - x'\|^2 \mathrm{d}x \mathrm{d}x' \\
&= \frac{1}{\pi} \int_{M_0} \|x\|^2 \mathrm{d}x \qquad \text{(by symmetry)} \\
&= \frac{1}{\pi} F(a) \\
&< \frac{1}{\pi} F(1) \\
&= \frac{1}{\pi} \int_{M_0} \|\phi(x)\|^2 \mathrm{d}x \\
&= \mathcal{E}_0(\phi).
\end{aligned}$$

Let $M_\sigma = \bar{B}(M_0, \sigma)$. There is a numeric constant $\sigma_0 > 0$ such that, when $\sigma < \sigma_0$, ψ does not maximize the energy $\mathcal{E}_\sigma$, and we conclude again that if $D = M = M_\sigma$, MVU does not recover D up to a rigid transformation.

Quantitative convergence result

Theorem

Suppose that M is either thin or thick, of dimension d, and that μ is comparable to the uniform distribution on M (as before).

Quantitative convergence result

Theorem

Suppose that M is either thin or thick, of dimension d, and that μ is comparable to the uniform distribution on M (as before).

Assume that $r_n \to 0$ such that $r_n \gg r_n^\dagger := (\log(n)/n)^{1/d}$ and take any $a_n \to \infty$.

Quantitative convergence result

Theorem

Suppose that M is either thin or thick, of dimension d, and that μ is comparable to the uniform distribution on M (as before).

Assume that $r_n \to 0$ such that $r_n \gg r_n^\dagger := (\log(n)/n)^{1/d}$ and take any $a_n \to \infty$.

Then, with probability one,

$$\begin{aligned} & \Big| \sup\{\mathcal{E}(Y) : Y \in \mathcal{Y}_{n,r_n}\} - \sup\{\mathcal{E}(f) : f \in \mathcal{F}_1\} \Big| \\ \leq \quad & a_n \Big(r_n + \frac{r_n^\dagger}{r_n} + n^{-1/(d+2)} \Big), \end{aligned}$$

for n large enough.

Thin and thick sets

Thin sets. M is a d-dimensional compact, connected, C^2 submanifold with C^2 boundary (if nonempty). In addition, $M \subset M_\star$, where $M_\star$ is a d-dimensional, geodesically convex C^2 submanifold.

Thin and thick sets

Thin sets. M is a d-dimensional compact, connected, C^2 submanifold with C^2 boundary (if nonempty). In addition, $M \subset M_\star$, where $M_\star$ is a d-dimensional, geodesically convex C^2 submanifold.

Thick sets. M is a compact, connected subset that is the closure of its interior and has a C^2 boundary.

Some open problems

Convergence rate for the solution(s). We obtained a convergence rate for the energy $\sup_Y \mathcal{E}(Y)$. How about the convergence rate of $\hat{y}$ to $\mathcal{S}_1$?

Some open problems

Convergence rate for the solution(s). We obtained a convergence rate for the energy $\sup_Y \mathcal{E}(Y)$. How about the convergence rate of $\hat{y}$ to $\mathcal{S}_1$?

Flattening property of MVU. When M satisfies the Isometry Assumption, does Continuous MVU always flatten the manifold M to its intrinsic dimension?

EAC and Pelletier, B. (2013). On the convergence of maximum variance unfolding. *JMLR*, 14(1), 1747-1770.

Gracias