On a Connection between Maximum Variance Unfolding, Shortest Path Problems and IsoMap

Alexander Paprotny
Inst. für Mathematik, TU Berlin
Straße des 17. Juni 136, D-10623 Berlin
paprotny@math.tu-berlin.de

Jochen Garcke
Inst. f. Numerische Simulation, Univ. Bonn
Wegelerstr. 6, D-53115 Bonn
Fraunhofer SCAI
Schloss Birlinghoven, D-53754 St. Augustin
garcke@ins.uni-bonn.de

Abstract

We present an equivalent formulation of the Maximum Variance Unfolding (MVU) approach to nonlinear dimensionality reduction in terms of distance matrices. This yields a novel interpretation of the MVU problem as a regularized version of the shortest path problem on a graph. This interpretation enables us to establish an asymptotic convergence result for the case that the underlying data are drawn from a Riemannian manifold which is isometric to a convex subset of Euclidean space.

1 Introduction

Problems related to high-dimensional data arise in many areas of science and engineering. Since a major part of practically relevant algorithms for classification, regression, and clustering suffer from the so-called *curse of dimensionality*, the necessity arises to exploit the fact that many of these high-dimensional data sets are intrinsically low-dimensional, i.e., may be described by few parameters. The goal of dimensionality reduction consists in recovering such low-dimensional parametrizations empirically. A classical approach to dimensionality reduction is Principal Component Analysis (PCA) [Hot33], which, however, assumes, that, up to noise, the data lie in a linear submanifold, so that it fails to recover the correct dimensionality of data residing in nonlinear structures. Therefore, the problem of nonlinear dimensionality reduction (NLD) has been drawing the attention of researchers, especially in recent years several new approaches were introduced.

Technically, in an NLD task, we are given a finite number of data points in high-dimensional Euclidean space that are assumed to reside in a low-dimensional submanifold. The goal consists in obtaining a low-dimensional representation of the data which respects the intrinsic geometry.

A major part of approaches to NLD is referred to as *spectral methods*. These rely on establishing a kernel matrix and eventually obtain a geometrically faithful low-dimensional representation of the data through a spectral decomposition of the former. What is particularly appealing about spectral methods is that the computation may be carried out by means of linear algebra rather than non-convex optimization, which is involved in regression-based approaches such as principal manifolds [SMSW01]. Examples of spectral methods for NLD are IsoMap [TdSL00], Laplacian Eigenmaps [BN03], Locally Linear Embedding [RS00], Local Tangent Space Alignment [ZZ02], and Maximum Variance Unfolding [WS06a, WS06b, WS04, WSS04, WSZS06]. Each of these approaches, however, brings along its own notion of a geometrically faithful low-dimensional representation of manifold data, and a unifying analysis is a matter of ongoing research.

We will focus on Maximum Variance Unfolding (MVU), a heuristic approach to nonlinear dimensionality reduction. Informally, the procedure may be outlined as follows: As is the case in many NLD heuristics, the first step consists in establishing a neighborhood relation on the sampled configuration, i.e., in determining which sample points are close to each other in terms of the intrinsic geometry. The desired low-dimensional representation is then taken to be a configuration of maximum variance amongst those that preserve the distances corresponding to edges of the neighborhood graph.

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

Besides intuitive and empirical plausibility, little is known about consistency of the method. Moreover, an empirical study of the asymptotic behavior of the procedure as the number of data points increases is computationally intractable. Note that this will prevent us from numerical studies in regard to the convergence results we will derive.

In this paper, we shall discuss an equivalent formulation of the MVU problem in terms of the distance matrix of the sought-after configuration. This amounts to a convex optimization problem over the cone of Euclidean distance matrices. We shall show that by omitting the constraint that the sought-after distance matrix be Euclidean, one arrives at an optimization problem whose unique optimizer is the distance matrix of the neighborhood graph. This suggests the interpretation of MVU as a regularized version of the shortest path problem on the neighborhood graph, which, additionally, reveals an unexpectedly intimate link to the IsoMap approach. Furthermore, we refine this insight by revealing that each optimizer of the MVU problem is also an optimal l^1-approximation to the distance matrix of the neighborhood graph. Equipped with these new insights and the theory of graph approximations to geodesic distances [TdSLB00] originally dedicated to IsoMap, we eventually derive an asymptotic convergence result restricted to the case that the underlying manifold is isometric to a convex domain in Euclidean space. Eventually, we shall argue that Laplacian Eigenmaps (LE) may be equivalently stated as a slightly modified version of MVU and briefly present some insights on a non-Euclidean version of Colored MVU [SSBG08], a weighted extension of MVU.

1.1 Related Work

To our knowledge, the connection between MVU and the shortest path problem on a graph has not yet been pointed out. We consider it, however, worth mentioning that Boyd et al. [SBXD06, XSB06] derive a connection between MVU and the problem of finding a fastest mixing Markov process (FMMP) on a graph by means of duality theory for semidefinite programming. This insight establishes a connection between MVU and LE which is fundamentally different from that pointed out in this paper.

The technique employed here to derive an SDP formulation of the shortest path problem parallels that in [BT96, Section 2.2.4, p. 36 ff], where it is used in the context of Markov decision processes.

Convergence of MVU is not discussed rigorously in any literature we are aware of. The asymptotic result presented in this paper is informally foreshadowed in [XSB06], which has actually inspired our analysis.

The metric equivalence assumption underlying our convergence result for MVU is justified by the convergence result for graph approximations to geodesic distances presented in [TdSLB00]. Interestingly, this result is intended as a justification for IsoMap.

2 $\mathbb{EDM}$-Formulations of Maximum Variance Unfolding and IsoMap

2.1 Preliminaries

What follows are basic notations we shall use throughout the paper when dealing with linear algebra or differential geometry, respectively.

The symbol $\mathcal{M}$ refers to a (Riemannian) manifold, $\mathcal{E}^a$ to a-dimensional Euclidean space as a specific Riemannian manifold, dist to a metric, $\mathrm{dist}_{\mathcal{M}}$ to the geodesic metric on a Riemannian manifold. Furthermore, we use the abbreviation $\mathrm{dist}_{\mathcal{E}}$ for $\mathrm{dist}_{\mathcal{E}^a}$ when the dimension a follows from the context.

We shall deal extensively with the subsequently defined sets of matrices: $\mathbb{S}^N$ denotes the space of symmetric matrices of order N and $\mathbb{S}^N_{\succeq O}$ the set of symmetric positive semidefinite (SPSD) matrices of order N. Similarly, $\mathbb{S}^N_{\geq O}$ refers to the set of nonnegative (monotone) symmetric matrices of order N.

We denote by I the identity matrix, e_i the i-th unit vector, $\mathbf{1}$ the vector of all ones, and by $P_{\perp\mathbf{1}}$ the orthogonal projector along the linear hull of $\mathbf{1}$. Finally, for $N \in \mathbb{N}$, $\underline{N}$ denotes the set of the first N natural numbers (without 0).

Definition 1 *A (finite, weighted)* undirected graph *is a triplet $G := (V, E, d^w)$, where the* vertex set *V is a finite set, the* edge set *$E \subseteq \{e \mid e \subseteq V \wedge |e| = 2\}$, and the* edge weighting *$d^w : E \to \mathbb{R}_{>0},\ \{i, j\} \mapsto d^w_{ij}$.*

A path *γ of cardinality $|\gamma| = k \in \underline{N}$ in G is a tuple $\gamma = (\gamma_1, \ldots, \gamma_k) \in V^k$ such that for each $i = 1, \ldots, k-1$, $\{\gamma_i, \gamma_{i+1}\} \in E$. The set of all paths in a graph G is denoted by Π^G. We say that $\gamma \in \Pi^G$* connects *the vertices $v, w \in V$ if $\gamma_1 = v$ and $\gamma_{|\gamma|} = w$. The set of all paths connecting $v, w \in V$ is denoted by Π^G_{vw}. The* length *$\gamma \in \Pi^G$ is defined as*

$$l(\gamma) := \sum_{i=2}^{|\gamma|} \sqrt{d^w_{\gamma_{i-1}\gamma_i}}.$$

The graph distance *between $v, w \in V$ is defined as*

$$\mathrm{dist}_G(v, w) := \min_{\gamma \in \Pi^G_{vw}} l(\gamma).$$

The corresponding distance matrix *is defined as $D^G := (\mathrm{dist}^2_G(v, w))_{v,w \in V}$. We say that $\gamma \in \Pi^G_{vw}$ is a* shortest

path *if* $l(\gamma) = \mathrm{dist}_G(v, w)$. *We shall refer to the problem of computing* D^G *as the* shortest path problem.

Definition 2 *A finite sequence* $\{x_i\}_{i\in\underline{N}} \subseteq \mathcal{E}^D$ *is called* D-dimensional N-configuration. *We define the corresponding* configuration matrix *by* $X := [x_1, \ldots, x_N]$, Gram matrix *by* $G := G(X) := (\langle x_i, x_j\rangle)_{i,j\in\underline{N}} = X^T X$, *and* distance matrix *by* $D := D(X) := (\mathrm{dist}^2_{\mathcal{E}}(x_i, x_j))_{i,j\in\underline{N}}$.

The mean *of a configuration is defined as* $\frac{1}{N} X\mathbf{1}$. *A configuration is said to be* mean centered *if* $X\mathbf{1} = 0$.

2.2 Maximum Variance Unfolding

Maximum Variance Unfolding (MVU), proposed by Weinberger et al. [WS06a, WS06b, WS04, WSS04, WSZS06], is a heuristic for nonlinear dimensionality reduction. It relies on the intuition that a faithful low-dimensional representation of high-dimensional data residing on a low-dimensional, possibly nonlinear manifold may be obtained by an appropriate "unfolding" thereof. Specifically, we are given a configuration $\mathcal{Y} = \{y_i\}_{i\in\underline{N}} \subseteq \mathcal{E}^a$ which is assumed to lie on a d-dimensional Riemannian submanifold of $\mathcal{E}^a$ with $d \ll a$. The goal consists in obtaining a configuration $\mathcal{X} = \{x_i\}_{i\in\underline{N}} \subseteq \mathcal{E}^d$ which, in some sense, reflects the intrinsic geometry of the given configuration. More precisely, the configuration $\mathcal{X}$ should recover the geodesic distances between the points in $\mathcal{Y}$ to some extent. Since the geodesic distances are typically not available, we must rely on estimations thereof. As is common in spectral methods for NLD, we establish a neighborhood relation on $\mathcal{Y}$ and use Euclidean distances for approximations of geodesic distances between neighboring points. The physical intuition leading to the MVU approach now is as follows: In terms of the Euclidean metric, we pull the points in the data set maximally apart while preserving distances between neighboring points. Moreover, we require that the configuration be mean centered. Since, on one part, the intrinsic dimensionality d is obscure in many practical situations, on the other part, the heuristic assumes that pulling the data apart under the given constraints recover d automatically, any constraints stipulating the dimensionality of the sought-after configuration are omitted. Parametrized in the Gram matrix $K := X^T X$ of the latter, the approach amounts to the following semidefinite program:

$$\begin{aligned} \max_{K\in\mathbb{S}^N_{\succeq O}} \ & \langle I, K\rangle \\ \text{s.t.} \ & \langle E_{ij}, K\rangle \leq d^{\mathcal{E}}_{ij}, \ \{i,j\} \in E, \\ & \left\langle \mathbf{1}\mathbf{1}^T, K\right\rangle = 0. \end{aligned} \tag{2.1}$$

where

$$E_{ij} := (e_i - e_j)(e_i - e_j)^T, \tag{2.2}$$

E is the edge set of the neighborhood graph, and

$$\langle A, B\rangle = \sum_{i,j\in\underline{N}} a_{ij} b_{ij}, \ A, B \in \mathbb{S}^N.$$

A (low-dimensional) configuration corresponding to an optimizer is eventually retrieved from a spectral decomposition thereof.

The formulation (2.1), which we shall refer to as *classical MVU*, is presented and discussed in [SBXD06, XSB06]. It deviates from the original version proposed in [WS06a, WS06b, WS04, WSS04, WSZS06] in that the local distance constraints have been relaxed to inequalities. We stress that all results we will present in Section 3 also hold for equality constraints (cf. Remark 4).

Since we will study the equivalent $\mathbb{EDM}$-cone formulation of (2.1), we give the following definitions:

Definition 3 *1. The set of* Euclidean distance matrices of order N *is defined as*

$$\mathbb{EDM}^N := \left\{(\mathrm{dist}^2_{\mathcal{E}}(x_i, x_j))_{i,j\in\underline{N}} \mid \exists\, n \in \mathbb{N} : x_i \in \mathbb{R}^n, \ i \in \underline{N}\right\}. \tag{2.3}$$

2. The set of distance matrices of order N *is defined as (with* $D := (d_{ij})_{i,j\in\underline{N}}$*)*

$$\mathbb{DM}^N := \left\{D \in \mathbb{S}^N_{\geq O} \mid d_{ii} = 0, \ i \in \underline{N} \wedge \sqrt{d_{ij}} \leq \sqrt{d_{ik}} + \sqrt{d_{kj}}, \ i, j \in \underline{N}\right\}. \tag{2.4}$$

Obviously, $\mathbb{EDM}^N \subseteq \mathbb{DM}^N$. Furthermore, we obtain

Proposition 1 *The sets* $\mathbb{DM}^N$ *and* $\mathbb{EDM}^N$ *defined in Definition 3 are proper closed convex cones*[1].

Since the proof is not essential for the observations of this article it is only provided in the supplementary material.

Remark 1 *We stress that the notion of a distance matrix stipulated here essentially differs from that in [Ver04]. Specifically, the notion of a distance matrix considered therein may be obtained by omitting the square roots in the definition of ours. Therefore, showing that* $\mathbb{DM}^N$ *is a cone is non-trivial. For* $\mathbb{EDM}^N$ *this is well known (see, e.g., [Dat05]), but the* $\mathbb{DM}^N$ *case is not handled in [Dat05] or [Ver04], nor in other related literature we have found.*

[1] A subset C of a Banach space is called cone if $\lambda C \subseteq C \ \forall\ \lambda \geq 0$. A cone is called *proper* if $v \in C \wedge -v \in C$ implies $v = 0$.

Distance matrices may be considered as representations of metrics on graphs. In particular, we shall be interested in graph approximations to geodesic distances in the manifold learning setting.

For the sake of simplicity, we shall henceforth identify configurations with their corresponding configuration matrices.

It is well-known that each matrix $S \in \mathbb{S}^N_{\succeq O}$ may be decomposed according to $S = X^T X$ for some $X \in \mathbb{R}^{r\times N}$, where r denotes the rank of S. Conversely, every Gram matrix is symmetric and positive semidefinite. Therefore, $\mathbb{S}^N_{\succeq O}$ may be considered as the set of all Gram matrices corresponding to N-configurations.

One may easily verify that the distance matrix of a configuration is invariant under rigid transformation. The Gram matrix, however, is only invariant under unitary transformation, but depends on the absolute position of the configuration in the chosen coordinate system. To achieve a $1-1$ correspondence, we restrict ourselves to mean centered configurations. The corresponding set of Gram matrices is given by $P_{\perp\mathbf{1}}\mathbb{S}^N_{\succeq O}P_{\perp\mathbf{1}}$. Given $D \in \mathbb{EDM}^N$, the Gram matrix of a corresponding mean centered configuration is given by $-\frac{1}{2}P_{\perp\mathbf{1}}DP_{\perp\mathbf{1}}$. Conversely, as may easily be verified by the well-know parallelogram formula, given $S \in \mathbb{S}^N_{\succeq O}$, the distance matrix of any corresponding configuration is $(s_{ii}+s_{jj}-2s_{ij})_{i,j\in\underline{N}}$. For more details on Euclidean distance matrices, please refer to [Dat05, Sec. 5.10].

The above discussed correspondences between configurations, distance and Gram matrices enable us to cast optimization problems related to configurations in terms of Gram matrices so as to obtain semidefinite optimization problems. Besides this computational advantage, the correspondences provide new interpretations of existing methods, which we shall exploit in the following.

In this paper, we shall obtain novel insights by studying the subsequent equivalent $\mathbb{EDM}$-cone formulation of (2.1):

$$\begin{aligned} \max_{D\in\mathbb{EDM}^N} \ & \langle \mathbf{1}\mathbf{1}^T, D\rangle \\ \text{s.t.} \ & \langle e_{ij}, D\rangle \le d^{\mathcal{E}}_{ij}, \ \{i,j\}\in E, \end{aligned} \tag{2.5}$$

where $e_{ij} := e_i e_j^T$. Here and in what follows, $D^{\mathcal{E}} = (d^{\mathcal{E}}_{ij})_{i,j\in\underline{N}}$ denotes the Euclidean distance matrix corresponding to the configuration $\mathcal{Y}$, i.e., $d^{\mathcal{E}}_{ij} := \mathrm{dist}^2_{\mathcal{E}}(y_i,y_j),\ i,j\in\underline{N}$.

Remark 2 *Equivalence of the programs* (2.1) *and* (2.5) *may be established by replacing* K *in* (2.1) *by the invertible parametrization*

$$\mathbb{EDM}^N \to -P_{\perp\mathbf{1}}\mathbb{S}^N_{\succeq O}P_{\perp\mathbf{1}},\ D\mapsto -\frac{1}{2}P_{\perp\mathbf{1}}DP_{\perp\mathbf{1}}.$$

2.3 IsoMap

IsoMap [TdSL00] is one of the earliest approaches to manifold learning. The procedure may be outlined as follows:

1. Establish a neighborhood graph and use Euclidean distances for edge weights.
2. Obtain the graph distance matrix D^G.
3. Obtain a Gramian $\tilde{K}$ by setting all negative eigenvalues of $-0.5P_{\perp\mathbf{1}}D^GP_{\perp\mathbf{1}}$ to zero.
4. Obtain a low-dimensional representation of the data by means of PCA of the Gramian $\tilde{K}$.

Remark 3 *As a by-product of step 3, we obtain a spectral decomposition of the optimizer, which immediately provides us with the desired PCA in the last step. Therefore, in a practical implementation, one would merge the final two steps. We list them separately merely for theoretical reasons that should become clear in the course of our discussion.*

The key idea underlying the algorithm consists in the approximation of geodesic distances by graph distances in a neighborhood graph. As for the latter, the following convergence result has been established.

Theorem 1 (Main Theorem A in [TdSLB00]) *Let* $\epsilon_{max},\ \epsilon_{min}, \delta > 0$, $0 \le \lambda_1, \lambda_2 < 1$ *and suppose that*

1. $\mathcal{M}$ *is compact and geodesically convex.*
2. $\forall\ i,j\in\underline{N}:\ \mathrm{dist}_{\mathcal{E}}(y_i,y_j)\le\epsilon_{\min}\Rightarrow\{i,j\}\in E,$
3. $\forall\ \{i,j\}\in E:\ \mathrm{dist}_{\mathcal{E}}(y_i,y_j)\le\epsilon_{\max},$
4. $\forall\ m\in\mathcal{M}:\ \exists\ i\in\underline{N}:\ \mathrm{dist}_{\mathcal{M}}(y_i,m)\le\delta,$
5. $\epsilon_{max}\le\min\left\{s_0, \frac{2}{\pi}r_0\sqrt{24\lambda_1}\right\}$, *where* s_0 *denotes the minimum branch separation of* $\mathcal{M}$, *and* r_0 *the minimum radius of curvature,*
6. $\delta\le 0.25\lambda_2\epsilon_{min}$.

Then we have

$$\begin{aligned}(1-\lambda_1)\,\mathrm{dist}_{\mathcal{M}}(y_i,y_j) &\le \mathrm{dist}_G(y_i,y_j)\\ &\le (1+\lambda_2)\,\mathrm{dist}_{\mathcal{M}}(y_i,y_j).\end{aligned} \tag{2.6}$$

Moreover, the authors of [TdSLB00] provide extensions of this result to a probabilistic setting and the k-NN rule. Hence, not only do the results presented in [TdSLB00] put graph approximations to geodesic

distance on a sound footing, but they also enable convergence guarantees in practically realistic, probabilistic settings.

The rationale for step 3 is that, even if the considered manifold is isometric to a convex subset of Euclidean space, and, hence, for any configuration from $\mathcal{M}$, the matrix $-\frac{1}{2}P_{\perp\mathbf{1}}D^{\mathcal{M}}P_{\perp\mathbf{1}}$ is positive semidefinite, where $D^{\mathcal{M}}$ denotes the geodesic distance matrix corresponding to the configuration, while the matrix $-\frac{1}{2}P_{\perp\mathbf{1}}D^{G}P_{\perp\mathbf{1}}$ need not be, even if it approximates $D^{\mathcal{M}}$ arbitrarily well. This is simply due to the fact that $\mathbb{EDM}^N$ is closed.

We shall consider an equivalent $\mathbb{EDM}$-formulation of step 3, which, since removing components corresponding to negative eigenvalues is equivalent to finding the best semidefinite approximation in terms of the Frobenius norm

$$||A||_F := \sqrt{\langle A, A\rangle},$$

is given by

$$\min_{D\in\mathbb{EDM}^N} ||P_{\perp\mathbf{1}}(D - D^G)P_{\perp\mathbf{1}}||_F. \tag{2.7}$$

In the main part of this paper, we shall establish an interesting connection between IsoMap and MVU. One of the main messages to be conveyed by this paper consists in the insight that asymptotic convergence of MVU may be obtained from the IsoMap-related convergence Theorem 1.

3 MVU as a Regularized Shortest Path Problem

We shall be interested in a generalization of MVU which deals with an abstract graph endowed with arbitrary positive edge weights and constrains the sought-after distance matrix to an arbitrary subset of $\mathbb{DM}^N$. This abstraction is inconsequential with respect to the proof of the subsequent theorem. In brevity, it states that the generalized MVU problem is equivalent to computing the best approximation to the distance matrix of the neighborhood graph with respect to $||\cdot||_1$.

Theorem 2 *Let $C \subseteq \mathbb{DM}^N$ and $G := (\underline{N}, E, d^w)$ be a weighted graph. If G is connected, then the following programs are equivalent:*

$$\begin{aligned} &\max_{D\in C} \left\langle \mathbf{1}\mathbf{1}^T, D\right\rangle \\ &s.t.\ \ \langle e_{ij}, D\rangle \leq d^w_{ij},\ \{i,j\} \in E, \end{aligned} \tag{3.1}$$

and

$$\begin{aligned} &\min_{D\in C} ||D - D^G||_1 \\ &s.t.\ \ \langle e_{ij}, D\rangle \leq d^w_{ij},\ \{i,j\} \in E. \end{aligned} \tag{3.2}$$

Proof. Let D be feasible for (3.1). Then, for all $i, j \in \underline{N}$ and all $\gamma \in \Pi^G_{ij}$, the triangle inequality implies

$$\sqrt{d_{ij}} \leq \sum_{k=2}^{|\gamma|} \sqrt{d_{\gamma_{k-1}\gamma_k}} \leq \sum_{k=2}^{|\gamma|} \sqrt{d^w_{\gamma_{k-1}\gamma_k}} = l(\gamma).$$

In particular, the inequality holds for any shortest path between i, j. Thus, we obtain $d_{ij} \leq d^G_{ij}$ for all $i, j \in \underline{N}$, which yields

$$\left\langle \mathbf{1}\mathbf{1}^T, D\right\rangle - \left\langle \mathbf{1}\mathbf{1}^T, D^G\right\rangle = \left\langle \mathbf{1}\mathbf{1}^T, D - D^G\right\rangle = \sum_{i,j\in\underline{N}} d_{ij} - d^G_{ij} = -\sum_{i,j\in\underline{N}} |d_{ij} - d^G_{ij}| = -||D - D^G||_1.$$

Since adding a constant to the objective does not affect contour lines and maximizing is equivalent to minimizing the additive inverse, the result follows. □

The subsequent results are immediate consequences of the above.

Corollary 1 *Let $G := (\underline{N}, E, d^w)$ be a weighted graph. If G is connected, then D^G is the unique solution of*

$$\begin{aligned} &\max_{D\in\mathbb{DM}^N} \left\langle \mathbf{1}\mathbf{1}^T, D\right\rangle \\ &s.t.\ \ \langle e_{ij}, D\rangle \leq d^w_{ij},\ \{i,j\} \in E. \end{aligned} \tag{3.3}$$

This result tells us, that the shortest-path-problem on a graph is equivalent to a variant of MVU, called *non-Euclidean MVU*, where now the distance matrix is no longer constrained to be Euclidean.

Corollary 2 *Let $G := (\underline{N}, E, d^w)$ be a weighted graph. If G is connected, then the following holds.*

1. *If $D^G \in \mathbb{EDM}^N$, then D^G is the unique solution of* (2.5).
2. *The program* (2.5) *is equivalent to*

$$\begin{aligned} &\min_{D\in\mathbb{EDM}^N} ||D - D^G||_1 \\ &s.t.\ \ \langle e_{ij}, D\rangle \leq d^w_{ij},\ \{i,j\} \in E. \end{aligned} \tag{3.4}$$

In general, D^G need not be Euclidean, even if $D^{\mathcal{M}}$ is for all possible configurations. Furthermore, since $\mathbb{EDM}^N$ is closed, it is even possible for a sequence of configurations of increasing size and "density" that the bi-Lipschitz bounds relating $D^{\mathcal{M}}$ and D^G according to Theorem 1 become arbitrarily tight while D^G remains strictly outside $\mathbb{EDM}^N$. Hence, considering the non-Euclidean version of MVU (3.3), its solution D^G may provide an arbitrarily good approximation to the intrinsic distance matrix $D^{\mathcal{M}}$. However, if $D^{\mathcal{M}}$ is in $\mathbb{EDM}^N$, one typically prefers an approximation therein. This

is achieved by restricting the search space to $\mathbb{EDM}^N$ and thus arriving at classical MVU. This justifies our speaking of a regularized shortest path problem. In the light of manifold learning, it may be even clearer to speak of a regularized geodesic distance approximation problem.

In the same fashion, with the $\mathbb{EDM}$-formulation (2.7) in mind, IsoMap may be considered as a regularized shortest path problem, except that it is based on a different cost function and does not incorporate any constraints other than the distance matrix to be Euclidean.

According to its authors, MVU was intended as an approach which, as opposed to IsoMap, avoids taking global geodesic distance estimates into account. We consider it remarkable that global distance estimates are incorporated implicitly, seemingly unnoticed by others so far.

Remark 4 *All results presented so far in this section also hold for the original version of MVU [WS06a, WS06b, WS04, WSS04, WSZS06], which is obtained by replacing the family of constraints*

$$\langle e_{ij}, D\rangle \leq d_{ij}^w,\ \{i,j\} \in E,$$

by

$$\langle e_{ij}, D\rangle = d_{ij}^w,\ \{i,j\} \in E,$$

in all of the above optimization problems. This is due to the fact that the proof of Theorem 2 carries over to the latter case verbatim.

4 Asymptotic Result

Let us now formulate the setting of our result on the asymptotic behavior of MVU.

Assumption 1
1. *$\mathcal{M}$ is a d-dimensional compact Riemannian manifold isometric to a convex subset of $\mathcal{E}^d$,*
2. *$\mathcal{Y} := \{y_1, \dots, y_N\} \subseteq \mathcal{M}$ with geodesic distance matrix $D^{\mathcal{M}}$,*
3. *$G := (\underline{N}, E, d^w)$,*
4. *the distance matrix D^G satisfies $(1-\epsilon)D^{\mathcal{M}} \leq D^G \leq (1+\epsilon)D^{\mathcal{M}}$ for some $\epsilon \geq 0$.*

Remark 5 *The last item of Assumption 1 may be justified by Theorem 1 and the therefrom derived probabilistic convergence results on graph approximations to geodesic distances presented in [TdSLB00]. Specifically, we may assume that $\epsilon \to 0$ as $N \to \infty$, i.e., for an increasing number of suitably drawn samples. This observation qualifies the subsequent theorem as an asymptotic result.*

Combined with the subsequent lemma, the derivations presented in Section 3 yield the convergence result. The notation is as follows: For a set S, $\mathbb{R}^S$ denotes the set of real-valued functions on S. Furthermore the restriction of a function $f \in \mathbb{R}^S$ to a subset $\tilde{S} \subseteq S$ is denoted by $f|_{\tilde{S}}$.

Lemma 1 *Let S be a set, $\tilde{S} \subseteq S$, $C \subseteq \mathbb{R}^{\tilde{S}}$, $f \in \mathbb{R}^S$, and $\tilde{f} \in \mathbb{R}^{\tilde{S}}$. Moreover, let $||\cdot||$ be a norm on $\mathbb{R}^{\tilde{S}}$ and $c, \epsilon \geq 0$. If*

$$||\tilde{f} - f|_{\tilde{S}}|| \leq c\epsilon, \tag{4.1}$$
$$(1-\epsilon)f|_{\tilde{S}} \in C, \tag{4.2}$$

then

$$||\hat{f} - f|_{\tilde{S}}|| \leq (2c + ||f|_{\tilde{S}}||)\epsilon \ \ \forall\, \hat{f} \in \operatorname*{argmin}_{\overline{f} \in C} ||\overline{f} - \tilde{f}||. \tag{4.3}$$

Proof.

$$\begin{aligned} ||\hat{f} - f|_{\tilde{S}}|| &\leq ||\hat{f} - \tilde{f}|| + ||\tilde{f} - f|_{\tilde{S}}|| \\ &\overset{(2)}{\leq} ||(1-\epsilon)f|_{\tilde{S}} - \tilde{f}|| + ||\tilde{f} - f|_{\tilde{S}}|| \\ &\leq \epsilon||f|_{\tilde{S}}|| + 2||\tilde{f} - f|_{\tilde{S}}|| \\ &\overset{(1)}{\leq} (||f|_{\tilde{S}}|| + 2c)\epsilon. \end{aligned}$$

□

Theorem 3 (Asymptotic behavior of MVU) *Let Assumption 1 hold. Then any solution of (2.5) satisfies*

$$||D - D^{\mathcal{M}}||_1 \leq 3||D^{\mathcal{M}}||_1\epsilon \leq 3(N \operatorname{diam}(\mathcal{M}))^2\epsilon, \tag{4.4}$$

where $\operatorname{diam}(\mathcal{M}) := sup_{x,y\in\mathcal{M}} \operatorname{dist}_{\mathcal{M}}(x,y)$.

Proof. The proof is a mere application of Lemma 1 and Corollary 2. Assigning therein

- $S := \mathcal{M} \times \mathcal{M}$, $\tilde{S} := \mathcal{X} \times \mathcal{X}$,
- $C := \left\{\overline{f} \in \mathbb{R}^{\tilde{S}} | (\overline{f}(x_i, x_j))_{i,j\in\underline{N}} \text{ is feasible for (2.5)}\right\}$,
- $f := \operatorname{dist}^2_{\mathcal{M}}(\cdot,\cdot)$, $\tilde{f} : (x_i, x_j) \mapsto d^G_{ij}$,
- $||\cdot|| := ||(\cdot(x_i, x_j)_{i,j\in\underline{N}})||_1$,
- $c := ||D^{\mathcal{M}}||_1$, $D^{\mathcal{M}} := (\operatorname{dist}^2_{\mathcal{M}}(x_i, x_j))_{i,j\in\underline{N}}$,

it remains to verify the conditions (4.1) and (4.2). Statement 4 of Assumption 1 yields

$$\begin{aligned} ||\tilde{f} - f|_{\tilde{S}}|| = ||D^G - D^{\mathcal{M}}||_1 &= \sum_{i,j\in\underline{N}} |d^G_{ij} - d^{\mathcal{M}}_{ij}| \\ &\leq \sum_{i,j\in\underline{N}} \epsilon d^{\mathcal{M}}_{ij} = \epsilon||D^{\mathcal{M}}||_1, \end{aligned}$$

from which (4.1) follows. To obtain (4.2), please note that Statement 2 of Assumption 1 implies that $D^{\mathcal{M}} \in \mathbb{EDM}^N$. Invoking the left inequality of Statement 4 of Assumption 1 and the fact that $\mathbb{EDM}^N$ is a cone, it follows that $(1-\epsilon)D^{\mathcal{M}} \in \mathbb{EDM}^N$ and is therefore feasible for (2.5). Hence, the hypothesis of Lemma 1 holds and the first inequality of (4.4) follows from (4.3) and the fact that $||f|_{\tilde{S}}|| = ||D^{\mathcal{M}}||_1$. The second inequality is eventually obtained from the estimate

$$||D^{\mathcal{M}}||_1 = \sum_{i,j \in \underline{N}} d_{ij}^{\mathcal{M}} \leq \max_{i,j \in \underline{N}} d_{ij}^{\mathcal{M}} \leq (\operatorname{diam} \mathcal{M})^2. \quad \square$$

The above result is a step towards an answer to the question of to what extent the geometrically intrinsic structure of the underlying manifold is revealed by MVU. Specifically, it states that, under certain circumstances, any distance matrix obtained from MVU provides an approximation to the geodesic distance matrix of the given configuration and the average error becomes arbitrarily small as the number of points in the configuration increases. Interestingly, the crucial point (4) of Assumption 1 may be established in virtue of Theorem 1, which was intended as a convergence result for IsoMap by its authors.

The result guarantees convergence only in terms of average error. But what justifies our usage of the term "convergence" is Theorem 1, which gives us conditions so that the constant ϵ can become arbitrarily small, in particular the configuration needs to be sufficiently large and dense. Furthermore, the magnitude $\operatorname{diam} \mathcal{M}$ depends exclusively on the underlying manifold rather than the considered configuration. In particular, it is finite for a compact manifold. Granted, convergence "on average" is a somewhat weak result. On the other hand, to our awareness, ours is the first rigorous convergence analysis of MVU.

5 Additional Observations

5.1 Laplacian Eigenmaps as a Modified Maximum Variance Unfolding

The σ-Laplacian of a weighted graph $G = (\underline{N}, E, d^w)$ is defined as

$$L_\sigma^G := \sum_{\{i,j\} \in E} \omega_{ij} E_{ij},\ \omega_{ij} := \exp(-\sigma^{-1} d_{ij}^w),\ \{i,j\} \in E,$$

where E_{ij} is as stipulated in (2.2). *Laplacian Eigenmaps*, devised in [BN03], is a spectral NLD method which deploys the quadratic form induced by some σ-Laplacian of a suitably chosen neighborhood graph as an objective. As above, the edge set of neighborhood graph is typically established by means of the ϵ- or k-NN-rule with Euclidean distances in ambient space for edge weights. Specifically, the approach consists in using the column set of a solution of the subsequent optimization as a low-dimensional representation of the sampled configuration $\mathcal{Y}$:

$$\min_{X \in \mathbb{R}^{d \times N}} trXLX^T \text{ s.t. } XX^T = I,\ X\mathbf{1} = 0. \tag{5.1}$$

Clearly, X is an optimizer of (5.1) if and only if its columns form an orthogonal basis for an eigenspace of L corresponding to a collection of d smallest nonzero eigenvalues. Thus, computationally, the problem may be efficiently solved by means of linear algebra rather than optimization procedures. Nevertheless, it is advantageous from a theoretical point of view to consider the original formulation as an optimization problem. To fit (5.1) into the semidefinite programming framework, it is required that the quadratic constraint be eliminated. To do so, we invoke the following result adapted from [Dat05, p. 308], which enables us to cast an eigenvalue problem of the form (5.1) as a semidefinite program.

Proposition 2 *Let $A \in \mathbb{S}^N_{\succeq O}$. Then the spectral projector onto the eigenspace of A corresponding to its d-smallest eigenvalues is an optimizer of*

$$\min_{K \in \mathbb{S}^N_{\succeq O}} \langle A, K \rangle \ s.t.\ \langle I, K \rangle \geq d, I - K \in \mathbb{S}^N_{\succeq O}. \tag{5.2}$$

Furthermore, if the $d+1$- and d-smallest eigenvalues of A are distinct, then the solution is unique.

Please note that the last constraint of (5.2) is equivalent to $||K||_2^* \leq 1$, where $||\cdot||_2^*$ denotes the operator norm induced by the Euclidean inner product on $\mathbb{R}^N$. Equipped with this result, we may cast (5.1) as the equivalent semidefinite program

$$\min_{K \in \mathbb{S}^N_{\succeq O}} \langle L, K \rangle \text{ s.t. } \langle I, K \rangle \geq d,\ ||K||_2^* \leq 1. \tag{5.3}$$

Interchanging the constraint with the objective by means of a Lagrange multiplier, we arrive at the following insight.

Theorem 4 *Let X be an optimizer of* (5.1). *Then $X^T X$ is an optimizer of*

$$\begin{aligned} \max_{K \in \mathbb{S}^N_{\succeq O}} & \langle I, K \rangle \\ s.t.\ & \langle E_{ij}, K \rangle \leq \operatorname{dist}^2_{\mathcal{E}}(x_i, x_j),\ \{i,j\} \in E, \\ & \left\langle \mathbf{1}\mathbf{1}^T, K \right\rangle = 0, \\ & ||K||_2^* \leq 1. \end{aligned} \tag{5.4}$$

A more detailed derivation is provided in the supplementary material.

Please note that (5.4) is basically an instance of MVU, except for the additional imposing a bound on the norm

of K. Thus, the message of the foregoing reasoning is that Laplacian Eigenmaps may be cast in a slightly modified MVU framework.

5.2 Non-Euclidean Colored MVU

We conclude the main part of this paper by a brief outlook to a generalization of non-Euclidean MVU in which nonnegative objectives other than the matrix of all ones are allowed. Specifically, we consider

$$\min_{D\in\mathbb{DM}^N} \langle W, D\rangle \text{ s.t. } \langle e_{ij}, D\rangle \leq d^w_{ij},\ \{i,j\}\in E, \quad (5.5)$$

for a connected graph $G = (\underline{N}, E, d^w)$ and some weight matrix $W \geq O$. As an aside, it is worth mentioning that (5.5) is the $\mathbb{DM}$ formulation of the non-Euclidean version of Colored MVU [SSBG08], a generalization of MVU allowing for positive semidefinite objective matrices other than the identity in (2.1). The subsequent proposition provides a preliminary characterization of the set of optimizers of (5.5).

Proposition 3 *Assume that $G := (\underline{N}, E, d^\omega)$ be connected. Then $D \in \mathbb{DM}^N$ is an optimizer of (5.5) if and only if $D \leq D^G$ and $d_{ij} = d^G_{ij}$, $\{i,j\}\in\tilde{E}$, where $\tilde{E} := \{\{i,j\} \mid w_{ij} > 0\}$. In particular, D^G is an optimizer.*

The next result provides a characterization of the case where the optimizer D^G is unique.

Definition 4 *Let $G = (V, E, d^w)$ be a weighted graph. We introduce the notion of a* geodesic covering *of G to refer to a set $C \subset \{\{i,j\} \mid x, y \in V\}$ such that for all $v, w \in V$, there is $\{x, y\} \in C$ such that some shortest path between x and y in G traverses v and w.*

Theorem 5 *Assume that G be connected. Then the graph distance matrix D^G is the unique solution of (5.5) if and only if $\tilde{E}$ is a geodesic covering.*

Finding applications for this result is as yet a matter of ongoing research. We hope, however, that, under additional requirements on W, this result is extendible to the Euclidean case. Apart from this, it might be possible to exploit this insight for efficiently solving the shortest path problem on a graph, provided that a suitable objective matrix W be available beforehand.

6 Conclusion

We have established a connection between MVU and the shortest path problem on the underlying neighborhood graph. This connection enables us to consider the shortest path problem as a non-Euclidean version of MVU, and, conversely, MVU as a regularized shortest path problem. Moreover, we have argued that the latter also applies to IsoMap, which establishes a surprising connection between the two approaches to manifold learning. Furthermore, by virtue of the latter insight, we obtain a convergence result under reasonable assumptions, partly justified by the convergence theory of graph approximations to geodesic distances established in [TdSLB00]. Apart from this, we have demonstrated that Laplacian Eigenmaps is essentially equivalent to a modified version of MVU and considered Colored MVU in the distance matrix view. Unfortunately, the observed link between MVU and LE is too weak to account for the fact that the two approaches may yield rather different results in practice.

Please note that our asymptotic result guarantees convergence in terms of a (relative) average error. Since the latter is the optimization objective of MVU, we can hardly expect pointwise convergence. Inspired by [ZZ07], who consider a continuum version of IsoMap, we are currently studying a continuum version of MVU. We hope that stronger convergence results can be obtained by considering MVU as a discretization thereof.

Furthermore, we have not addressed the case of a manifold which is isometric to a connected rather than a convex set, e.g., a manifold with "holes". As a step towards understanding this case, we are currently working on a generalization of Theorem 1 to manifolds with a possibly non-convex boundary. The proof of Theorem 3, however, crucially relies on the assumption that the geodesic distance matrix of the sampled configuration be Euclidean, which, in general, need not be the case if the manifold is isometric to a connected Euclidean domain. Here, as well, we believe that a better understanding of the continuum version may enable further insights.

Finally note that our observations are valid in a more general setting than the afore considered case of embedded manifolds where Euclidean distances in ambient space are used as local distance estimators. Specifically, the results presented in Section 3 linking MVU to the shortest path problem and IsoMap do not only hold for neighborhood graphs of Euclidean configurations, but for arbitrary connected graphs. As for the asymptotic analysis carried out in Section 4, the only assumption on the distance estimator is a certain metric equivalence between the induced metric and the restriction of the geodesic metric onto the neighborhood graph.

Acknowledgements

The work was supported by the German Federal Ministry of Education and Research (BMBF) in the framework of its "Mathematics for innovations in industry and service" program within the project 03MS652D.

References

[BN03] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–1396, 2003.

[BT96] Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, Massachusetts, 1996.

[Dat05] Jon Dattorro. *Convex optimization & Euclidean distance geometry*. Meboo, 2005.

[Hot33] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441, 1933.

[RS00] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[SBXD06] Jun Sun, Stephen Boyd, Lin Xiao, and Persi Diaconis. The fastest mixing markov process on a graph and a connection to a maximum variance unfolding problem. *SIAM Review*, 48:2006, 2006.

[SMSW01] Alex J. Smola, Sebastian Mika, Bernhard Schölkopf, and Robert C. Williamson. Regularized principal manifolds. *Journal of Machine Learning Research*, 1:179–209, 2001.

[SSBG08] Le Song, Alex Smola, Karsten Borgwardt, and Arthur Gretton. Colored maximum variance unfolding. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *NIPS 20*, pages 1385–1392, 2008.

[TdSL00] J. B. Tenenbaum, V. de Silva, and J.C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 22 December 2000.

[TdSLB00] J.B. Tenenbaum, V. de Silva, J.C. Langford, and M. Berstein. Graph approximations to geodesics on embedded manifolds. December 20 2000.

[Ver04] A.M. Vershik. Random metric spaces and universality. *Russian Mathematical Surveys*, 59(2):259, 2004.

[WS04] Kilian Q. Weinberger and Lawrence K. Saul. Unsupervised learning of image manifolds by semidefinite programming. In *CVPR (2)*, pages 988–995, 2004.

[WS06a] Kilian Q. Weinberger and Lawrence K. Saul. An introduction to nonlinear dimensionality reduction by maximum variance unfolding. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pages 1683–1686. AAAI, 2006.

[WS06b] Kilian Q. Weinberger and Lawrence K. Saul. Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*, 70(1):77–90, 2006.

[WSS04] Kilian Q. Weinberger, Fei Sha, and Lawrence K. Saul. Learning a kernel matrix for nonlinear dimensionality reduction. In Carla E. Brodley, editor, *ICML '04*, pages 839–846. ACM, 2004.

[WSZS06] Kilian Q. Weinberger, Fei Sha, Qihui Zhu, and Lawrence K.. Saul. Graph laplacian regularization for large-scale semidefinite programming. In Bernhard Schölkopf, John C. Platt, and Thomas Hoffman, editors, *NIPS 19*, pages 1489–1496. MIT Press, 2006.

[XSB06] Lin Xiao, Jun Sun, and Stephen Boyd. A duality view of spectral methods for dimensionality reduction. In *ICML '06*, pages 1041–1048. ACM Press, 2006.

[ZZ02] Zhenyue Zhang and Hongyuan Zha. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *SIAM Journal of Scientific Computing*, 26(1):313–338, December 2002.

[ZZ07] Hongyuan Zha and Zhenyue Zhang. Continuum isomap for manifold learnings. *Computational Statistics & Data Analysis*, 52:184–200, 2007.