Co-manifold learning with missing data

Gal Mishne [1] **Eric C. Chi** [2] **Ronald R. Coifman** [1]

1. Algorithm 1: CO-CLUSTER-MISSING

We provide more technical details on Algorithm 1, which is an instance of a DC algorithm. DC algorithms in turn belong to the broader class of majorization-minimization (MM) algorithms (Lange et al., 2000). The basic strategy behind an MM algorithm is to convert a hard optimization problem into a sequence of simpler ones. The MM principle requires majorizing the objective function $f(\mathbf{U})$ by a surrogate function $g(\mathbf{U} \mid \tilde{\mathbf{U}})$ anchored at $\tilde{\mathbf{U}}$. Majorization is a combination of the tangency condition $g(\mathbf{U} \mid \tilde{\mathbf{U}}) = f(\tilde{\mathbf{U}})$ and the domination condition $g(\mathbf{U} \mid \tilde{\mathbf{U}}) \geq f(\mathbf{U})$ for all $\mathbf{U} \in \mathbb{R}^{m \times n}$. The associated MM algorithm is defined by the iterates $\mathbf{U}_{t+1} = \arg\min_{\mathbf{U}} g(\mathbf{U} \mid \mathbf{U}_t)$. It is straightforward to verify that the MM iterates generate a descent algorithm driving the objective function downhill, i.e. that $f(\mathbf{U}_{t+1}) \leq f(\mathbf{U}_t)$ for all t.

Recall that in the co-clustering step, we seek a minimizer $\mathbf{U}(\gamma_r, \gamma_c)$ of the function:

$$f(\mathbf{U}) = \frac{1}{2}\|\mathcal{P}_\Theta(\mathbf{X} - \mathbf{U})\|_{\mathrm{F}}^2 + \gamma_r J_r(\mathbf{U}) + \gamma_c J_c(\mathbf{U}), \quad (1)$$

where

$$\begin{aligned} J_r(\mathbf{U}) &= \sum_{(i,j)\in\mathcal{E}_r} \Omega(\|\mathbf{U}_{i\cdot} - \mathbf{U}_{j\cdot}\|_2) \\ J_c(\mathbf{U}) &= \sum_{(i,j)\in\mathcal{E}_c} \Omega(\|\mathbf{U}_{\cdot i} - \mathbf{U}_{\cdot j}\|_2). \end{aligned} \quad (2)$$

We make the following assumptions on Ω.

Assumption 1.1 *The row and column graphs $\mathcal{E}_r$ and $\mathcal{E}_c$ are connected, i.e. the row graph is connected if for any pair of rows, indexed by i and j with $i \neq j$, there exists a sequence of indices $i \to k \to \cdots \to l \to j$ such that $(i,k), \ldots, (l,j) \in \mathcal{E}_r$. A column graph is connected under analogous conditions.*

[1]Department of Mathematics, Yale University, New Haven, CT, USA [2]Department of Statistics, North Carolina State University, Raleigh, NC, USA. Correspondence to: Gal Mishne <gal.mishne@yale.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

Assumption 1.2 *The function $\Omega : [0,\infty) \mapsto [0,\infty)$ is (i) concave and continuously differentiable on $(0,\infty)$, (ii) vanishes at the origin, i.e. $\Omega(0) = 0$, (iii) is increasing on $[0,\infty)$, and (iv) has finite directional derivative at the origin.*

In the main paper, we use the following function Ω

$$\Omega(z) \quad = \quad \frac{1}{2}\int_0^z \frac{1}{\sqrt{\zeta} + \epsilon} d\zeta, \quad (3)$$

where ϵ is a small positive number, e.g. 10^{-12}.

The following function

$$\begin{aligned} g(\mathbf{U} \mid \tilde{\mathbf{U}}) = \frac{1}{2}\|\tilde{\mathbf{X}} - \mathbf{U}\|_{\mathrm{F}}^2 + \gamma_r \sum_{(i,j)\in\mathcal{E}_r} \tilde{w}_{r,ij}\|\mathbf{U}_{i\cdot} - \mathbf{U}_{j\cdot}\|_2 \\ + \gamma_c \sum_{(i,j)\in\mathcal{E}_c} \tilde{w}_{c,ij}\|\mathbf{U}_{\cdot i} - \mathbf{U}_{\cdot j}\|_2 + \kappa \end{aligned}$$

majorizes our objective function (1) at $\tilde{\mathbf{U}}$, where κ is a constant that does not depend on $\mathbf{U}$ and $\tilde{w}_{r,ij}$ and $\tilde{w}_{c,ij}$ are weights that depend on $\tilde{\mathbf{U}}$, i.e.

$$\begin{aligned} \tilde{w}_{r,ij} &= \Omega'(\|\tilde{\mathbf{U}}_{i\cdot} - \tilde{\mathbf{U}}_{j\cdot}\|_2) \\ \tilde{w}_{c,ij} &= \Omega'(\|\tilde{\mathbf{U}}_{\cdot i} - \tilde{\mathbf{U}}_{\cdot j}\|_2), \end{aligned} \quad (4)$$

where Ω' denotes the first derivative of Ω.

Minimizing $g(\mathbf{U} \mid \tilde{\mathbf{U}})$ is equivalent to minimizing the objective function of the convex biclustering problem for which efficient algorithms have been introduced (Chi et al., 2017). Thus, in the $t+1$th iteration, our MM algorithm solves a convex biclustering problem where the missing values in $\mathbf{X}$ have been replaced with the values of $\tilde{\mathbf{U}} = \mathbf{U}_t$ and the weights $\tilde{w}_{r,ij}$ and $\tilde{w}_{c,ij}$ have been computed based on $\tilde{\mathbf{U}} = \mathbf{U}_t$ according to (4). Note that the weights are continuously updated throughout the optimization as opposed to the fixed weights in Chi et al. (2017). This introduces a notion of the scale of the solution into the weights.

We first construct a majorization of the data-fidelity term. It is easy to verify that the following function of $\mathbf{U}$

$$g_1(\mathbf{U} \mid \tilde{\mathbf{U}}) \quad = \quad \frac{1}{2}\|\tilde{\mathbf{X}} - \mathbf{U}\|_{\mathrm{F}}^2, \quad (5)$$

where $\tilde{\mathbf{X}} = \mathcal{P}_\Theta(\mathbf{X}) + \mathcal{P}_{\Theta^c}(\tilde{\mathbf{U}})$, majorizes the data-fidelity term $\frac{1}{2}\|\mathcal{P}_\Theta(\mathbf{X}) - \mathcal{P}_\Theta(\mathbf{U})\|_{\mathrm{F}}^2$ at $\tilde{\mathbf{U}}$.

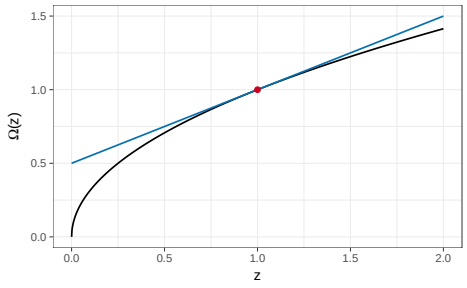

Figure 1: Majorization of the Ω function (black) given in (3) by its first-order Taylor approximation at 1 (blue).

We next construct a majorization of the penalty term. Recall that the first-order Taylor approximation of a differentiable concave function provides a global upper bound on the function. Therefore, under Assumption 1.2, we have the following inequality

$$\Omega(z) \quad \leq \quad \Omega(\tilde{z}) + \Omega'(\tilde{z})(z - \tilde{z}), \quad \text{for all } z, \tilde{z} \in [0, \infty).$$

Figure 1 shows the relationship between Ω given in (3) with $\epsilon = 10^{-12}$ and its first-order Taylor approximation at $\tilde{z} = 1$.

Thus, we can majorize the penalty term $\gamma_r J_r(\mathbf{U}) + \gamma_c J_c(\mathbf{U})$ with the function

$$\begin{aligned} g_2(\mathbf{U} \mid \tilde{\mathbf{U}}) \quad &= \quad \gamma_r \sum_{(i,j)\in\mathcal{E}_r} \tilde{w}_{r,ij} \|\mathbf{U}_{i\cdot} - \mathbf{U}_{j\cdot}\|_2 \qquad (6) \\ &+ \quad \gamma_c \sum_{(i,j)\in\mathcal{E}_c} \tilde{w}_{c,ij} \|\mathbf{U}_{\cdot i} - \mathbf{U}_{\cdot j}\|_2 + \kappa, \end{aligned}$$

where κ is a constant that does not depend on $\mathbf{U}$ and $\tilde{w}_{r,ij}$ and $\tilde{w}_{c,ij}$ (4) are weights that depend on $\tilde{\mathbf{U}}$. The sum of functions (5) and (6)

$$\begin{aligned} g(\mathbf{U} \mid \tilde{\mathbf{U}}) \quad &= \quad g_1(\mathbf{U} \mid \tilde{\mathbf{U}}) + g_2(\mathbf{U} \mid \tilde{\mathbf{U}}) \qquad (7) \\ &= \quad \frac{1}{2}\|\tilde{\mathbf{X}} - \mathbf{U}\|_{\mathrm{F}}^2 \\ &+ \quad \gamma_r \sum_{(i,j)\in\mathcal{E}_r} \tilde{w}_{r,ij} \|\mathbf{U}_{i\cdot} - \mathbf{U}_{j\cdot}\|_2 \\ &+ \quad \gamma_c \sum_{(i,j)\in\mathcal{E}_c} \tilde{w}_{c,ij} \|\mathbf{U}_{\cdot i} - \mathbf{U}_{\cdot j}\|_2 + \kappa \end{aligned}$$

majorizes our objective function (1) at $\tilde{\mathbf{U}}$.

2. Convex Biclustering Algorithm

We give a high level overview of the algorithm used to compute CVX-BCLUST in Algorithm 1 CO-CLUSTERING-MISSING. Note that we use the variable splitting approach given in (**?**). Given data matrix $\tilde{\mathbf{X}}$, cost parameters γ_r and γ_c, and weights $\tilde{w}_{r,ij}$ and $\tilde{w}_{c,ij}$, we seek the unique minimizer $\mathbf{U}$ of the convex optimization problem

$$\begin{aligned} \text{minimize } \frac{1}{2}\|\tilde{\mathbf{X}} - \mathbf{U}\|_{\mathrm{F}}^2 &+ \gamma_r \sum_{(i,j)\in\mathcal{E}_r} \tilde{w}_{r,ij} \|\mathbf{U}_{i\cdot} - \mathbf{U}_{j\cdot}\|_2 \\ &+ \gamma_c \sum_{(i,j)\in\mathcal{E}_c} \tilde{w}_{c,ij} \|\mathbf{U}_{\cdot i} - \mathbf{U}_{\cdot j}\|_2. \end{aligned}$$

The compositions of the nonsmooth 2-norm with an affine mapping of $\mathbf{U}$, i.e. $\|\mathbf{U}_{i\cdot} - \mathbf{U}_{j\cdot}\|_2$ and $\|\mathbf{U}_{\cdot i} - \mathbf{U}_{\cdot j}\|_2$, make solving the above optimization problem challenging. Consequently, we use variable splitting to reformulate the above unconstrained optimization problem as the following equivalent equality constrained problem

$$\begin{aligned} \text{minimize } \frac{1}{2}\|\tilde{\mathbf{X}} - \mathbf{U}\|_{\mathrm{F}}^2 &+ \gamma_r \sum_{(i,j)\in\mathcal{E}_r} \tilde{w}_{r,ij} \|\mathbf{v}_{r,ij}\|_2 \\ &+ \gamma_c \sum_{(i,j)\in\mathcal{E}_c} \tilde{w}_{c,ij} \|\mathbf{v}_{c,ij}\|_2. \end{aligned}$$

subject to

$$\begin{aligned} \mathbf{v}_{r,ij} \quad &= \quad \mathbf{U}_{i\cdot} - \mathbf{U}_{j\cdot} \quad \text{for all } (i,j) \in \mathcal{E}_r \\ \mathbf{v}_{c,ij} \quad &= \quad \mathbf{U}_{\cdot i} - \mathbf{U}_{\cdot j} \quad \text{for all } (i,j) \in \mathcal{E}_c, \end{aligned}$$

where the new dummy variables $\mathbf{v}_{r,ij}(\mathbf{v}_{c,ij})$ are the pairwise differences between the ith and jth rows (columns) of $\mathbf{U}$.

The Lagrangian dual problem to the above equivalent equality constrained problem is a constrained least squares problem which can be solved using projected gradient descent (**?**). Both the computational complexity and memory requirements of using the projected gradient descent algorithm to compute CVX-BCLUST are linear in the size of the data $\tilde{\mathbf{X}}$ (**?**). The number of times that we have to compute CVX-BCLUST depends on the inherent structure in the data. If the rows and columns have more clustered structure, convergence is rapid and CVX-BCLUST does not need to be applied many times. On the other hand, if there is less clustered and more manifold structure, more applications of CVX-BCLUST will be needed to attain convergence.

3. Proof of Proposition 1

The MM algorithm generates a sequence of iterates that has at least one limit point, and the limit points are stationary points of the objective function (1) To reduce notational clutter, we suppress the dependency of f on γ_r and γ_c since they are fixed during Algorithm 1. We prove Proposition 1 in three stages. First, we show that all limit points of the MM algorithm are fixed points of the MM algorithm map. Second, we show that fixed points of the MM algorithm are stationary points of f in (1). Finally, we show that the MM algorithm has at least one limit point.

3.1. Limit points are fixed points

The convergence theory of MM algorithms relies on the properties of the algorithm map $\psi(\mathbf{U})$ that returns the next iterate given the last iterate. For easy reference, we state a simple version of Meyer's monotone convergence theorem (Meyer, 1976), which is instrumental in proving convergence in our setting.

Theorem 1 *Let $f(\mathbf{U})$ be a continuous function on a domain S and $\psi(\mathbf{U})$ be a continuous algorithm map from S into S satisfying $f(\psi(\mathbf{U})) < f(\mathbf{U})$ for all $\mathbf{U} \in S$ with $\psi(\mathbf{U}) \neq \mathbf{U}$. Then all limit points of the iterate sequence $\mathbf{U}_k = \psi(\mathbf{U}_{k-1})$ are fixed points of $\psi(\mathbf{U})$.*

In order to apply Theorem 1, we need to identify elements in the assumption with specific functions and sets corresponding to the problem of minimizing (1). Throughout the following proof, it will sometimes be convenient to work with the column major vectorization of a matrix. The vector $\mathbf{b} = \mathrm{vec}(\mathbf{B})$ is obtained by stacking the columns of $\mathbf{B}$ on top of each other.

The function f: Take $\mathcal{S} = \mathbb{R}^{m\times n}$ and $f : \mathcal{S} \mapsto \mathbb{R}$ to be the objective function in (1) and majorize f with $g(\mathbf{U} \mid \tilde{\mathbf{U}})$ given in (7). The function f is continuous. Let $\psi(\tilde{\mathbf{U}}) = \arg\min_{\mathbf{U}} g(\mathbf{U} \mid \tilde{\mathbf{U}})$ denote the algorithm map for the MM algorithm. Since $g(\mathbf{U} \mid \tilde{\mathbf{U}})$ is strongly convex in $\mathbf{U}$, it has a unique global minimizer. Consequently, $f(\psi(\mathbf{U})) < f(\mathbf{U})$ for all $\psi(\mathbf{U}) \neq \mathbf{U}$.

Continuity of the algorithm map ψ: Continuity of ψ follows from the fact that the solution to the convex biclustering problem is jointly continuous in the weights and data matrix (Chi et al., 2017)[Proposition 2]. The weight $\tilde{w}_{r,ij}(\tilde{\mathbf{U}}) = \Omega'(\|\mathbf{U}_{i\cdot} - \mathbf{U}_{j\cdot}\|_2)$ is a continuous function of $\tilde{\mathbf{U}}$, since Ω' is continuous according to Assumption 1.2. The weight $\tilde{w}_{c,ij}(\tilde{\mathbf{U}})$ is likewise continuous in $\tilde{\mathbf{U}}$. The data matrix passed into the convex biclustering algorithm is $\tilde{\mathbf{X}} = \mathcal{P}_\Theta(\mathbf{X}) + \mathcal{P}_{\Theta^c}(\tilde{\mathbf{U}})$, which is a continuous function of $\tilde{\mathbf{U}}$ since the projection mapping $\mathcal{P}_{\Theta^c}$ is continuous.

3.2. Fixed points are stationary points

Let $\mathbf{L}_{ij} = (\mathbf{e}_i - \mathbf{e}_j)^\mathsf{T} \otimes \mathbf{I}$ and $\tilde{\mathbf{L}}_{ij} = \mathbf{I} \otimes (\mathbf{e}_i - \mathbf{e}_j)^\mathsf{T}$, where $\otimes$ denotes the Kronecker product. Let $\Delta_{ij} = \mathbf{L}_{ij}\mathbf{u}$ and $\widetilde{\Delta}_{ij} = \tilde{\mathbf{L}}_{ij}\mathbf{u}$. Then

$$\begin{aligned} \mathrm{vec}(\mathbf{U}_{i\cdot} - \mathbf{U}_{j\cdot}) &= \Delta_{ij} \\ \mathrm{vec}(\mathbf{U}_{\cdot i} - \mathbf{U}_{\cdot j}) &= \widetilde{\Delta}_{ij}. \end{aligned}$$

The directional derivative of f in the direction $\mathbf{v}$ at a point $\mathbf{u}$ is given by

$$\Omega'(\|\Delta_{ij}\|_2; \mathbf{v}) = \begin{cases} \Omega'(\|\Delta_{ij}\|_2)\langle \mathbf{L}_{ij}\mathbf{v}, \frac{\Delta_{ij}}{\|\Delta_{ij}\|_2}\rangle & \Delta_{ij} \neq \mathbf{0} \\ \Omega'(\|\Delta_{ij}\|_2)\|\mathbf{L}_{ij}\mathbf{v}\|_2 & \text{otherwise.} \end{cases}$$

A point $\mathbf{u}$ is a stationary point of f, if for all direction vectors $\mathbf{v}$

$$\begin{aligned} 0 \leq \langle \mathcal{P}_\Theta(\mathbf{u} - \mathbf{x}), \mathbf{v}\rangle \quad &+ \quad \gamma_r \sum_{(i,j)\in\mathcal{E}_r} \Omega'(\|\Delta_{ij}\|_2; \mathbf{v}) \\ &+ \quad \gamma_c \sum_{(i,j)\in\mathcal{E}_c} \Omega'(\|\widetilde{\Delta}_{ij}\|_2; \mathbf{v}), \end{aligned}$$

where $\mathcal{P}_\Theta(\mathbf{u} - \mathbf{x}) = \mathrm{vec}(\mathcal{P}_\Theta(\mathbf{U}) - \mathcal{P}_\Theta(\mathbf{X}))$.

A point $\mathbf{u}$ is a fixed point of ψ, if $\mathbf{0}$ is in the subdifferential of $g(\mathbf{u} \mid \mathbf{u})$, i.e.

$$\begin{aligned} \mathbf{0} \in \{\mathcal{P}_\Theta(\mathbf{u} - \mathbf{x})\} &+ \gamma_r \sum_{(i,j)\in\mathcal{E}_r} \Omega'(\|\Delta_{ij}\|_2)\partial\|\Delta_{ij}\|_2 \\ &+ \gamma_c \sum_{(i,j)\in\mathcal{E}_c} \Omega'(\|\widetilde{\Delta}_{ij}\|_2)\partial\|\widetilde{\Delta}_{ij}\|_2, \end{aligned} \tag{8}$$

where the set on the right is the subdifferential $\partial g(\mathbf{u} \mid \mathbf{u})$.

If $\Delta_{ij} \neq \mathbf{0}$, then $\partial\|\Delta_{ij}\|_2 = \left\{\mathbf{L}_{ij}^\mathsf{T}\frac{\Delta_{ij}}{\|\Delta_{ij}\|_2}\right\}$. On the other hand, if $\Delta_{ij} = \mathbf{0}$, then $\partial\|\Delta_{ij}\|_2 = \partial\|\mathbf{0}\|_2 = \{\mathbf{d} : \|\mathbf{d}\|_2 \leq 1\}$.

Fix an arbitrary direction vector $\mathbf{v}$. The inner product of $\mathbf{v}$ with an element in the set on right hand side of (8) is given by

$$\begin{aligned} \langle \mathcal{P}_\Theta(\mathbf{u} - \mathbf{x}), \mathbf{v}\rangle &+ \gamma_r \sum_{(i,j)\in\mathcal{E}_r} \Omega'(\|\Delta_{ij}\|_2)\langle \mathbf{d}_{ij}, \mathbf{v}\rangle \\ &+ \gamma_c \sum_{(i,j)\in\mathcal{E}_c} \Omega'(\|\widetilde{\Delta}_{ij}\|_2)\langle \mathbf{d}_{ij}, \mathbf{v}\rangle, \end{aligned} \tag{9}$$

where $\mathbf{d}_{ij} \in \partial\|\Delta_{ij}\|_2$ and $\tilde{\mathbf{d}}_{ij} \in \partial\|\widetilde{\Delta}_{ij}\|_2$.

Then the supremum of the right hand side of (9) over all $\mathbf{d}_{ij} \in \partial\|\Delta_{ij}\|_2$ and $\tilde{\mathbf{d}}_{ij} \in \partial\|\widetilde{\Delta}_{ij}\|_2$ is nonnegative, because $\mathbf{0} \in \partial g(\mathbf{u} \mid \mathbf{u})$. Consequently, all fixed points of ψ are also stationary points of f.

3.3. The MM iterate sequence has a limit point

To ensure the existence of a limit point, we show that the function f is coercive, i.e. $f(\mathbf{U}_t) \to \infty$ for any sequence $\|\mathbf{U}_t\|_\mathrm{F} \to \infty$. Recall that according to Assumption 1.1 we assume that the row and column edge sets $\mathcal{E}_r$ and $\mathcal{E}_c$ form connected graphs. Therefore, $J_r(\mathbf{U}) = J_c(\mathbf{U}) = 0$ if and only if $\mathbf{U} = a\mathbf{1}\mathbf{1}^\mathsf{T}$ (Chi et al., 2017, Proposition 3). The edge-incidence matrix of the column graph $\mathbf{\Phi}_c \in \mathbb{R}^{|\mathcal{E}_c|\times n}$ encodes its connectivity and is defined as

$$\phi_{c,li} = \begin{cases} 1 & \text{If node } i \text{ is the head of edge } l, \\ -1 & \text{If node } i \text{ is the tail of edge } l, \\ 0 & \text{otherwise.} \end{cases}$$

The row edge-incidence matrix $\mathbf{\Phi}_r \in \mathbb{R}^{|\mathcal{E}_r| \times m}$ is defined similarly. Assume that Θ non-empty, i.e. at least one entry of the matrix has been observed. Finally, assume that Ω is also coercive.

Note that any sequence $\mathbf{U}_t = a_t \mathbf{1}\mathbf{1}^\mathsf{T} + \mathbf{B}_t$ where $\langle \mathbf{B}_t, \mathbf{1}\mathbf{1}^\mathsf{T} \rangle = 0$. Note that $J_r(\mathbf{U}_t) = J_r(\mathbf{B}_t)$ and $J_c(\mathbf{U}_t) = J_c(\mathbf{B}_t)$. Let $\mathbf{U}_t$ be a diverging sequence, i.e. $\|\mathbf{U}_t\|_\mathrm{F} \to \infty$. There are two cases to consider.

Case I: Suppose that $\|\mathbf{B}_t\|_\mathrm{F} \to \infty$. Let

$$\mathbf{L} = \begin{pmatrix} \mathbf{I} \otimes \mathbf{\Phi}_r \\ \mathbf{\Phi}_c \otimes \mathbf{I} \end{pmatrix} \in \mathbb{R}^{|\mathcal{E}_r| m + |\mathcal{E}_c| n \times mn},$$

and let $\sigma_{\min}$ denote the smallest singular value of $\mathbf{L}$. Note that the null space of $\mathbf{L}$ is the span of $\mathbf{1}$. Therefore, since $\langle \mathbf{1}, \mathbf{b}_t \rangle = 0$

$$\|\mathbf{L}\mathbf{b}_t\|_2 \geq \sigma_{\min} \|\mathbf{B}_t\|_\mathrm{F}. \tag{10}$$

Also note that

$$\mathbf{L}\mathbf{b}_t = \begin{pmatrix} \mathrm{vec}(\mathbf{\Phi}_r \mathbf{B}_t) \\ \mathrm{vec}(\mathbf{B}_t \mathbf{\Phi}_c^\mathsf{T}) \end{pmatrix}.$$

Since the mapping $\mathbf{x} = \begin{pmatrix} \mathbf{x}_1^\mathsf{T} & \mathbf{x}_2^\mathsf{T} \end{pmatrix}^\mathsf{T} \mapsto \max\{\|\mathbf{x}_1\|_2, \|\mathbf{x}_2\|_2\}$ is a norm, and all finite dimensional norms are equivalent, there exists some $\eta > 0$ such that

$$\eta \|\mathbf{L}\mathbf{b}_t\|_2 \leq \max\left\{ \|\mathbf{\Phi}_r \mathbf{B}_t\|_\mathrm{F}, \|\mathbf{B}_t \mathbf{\Phi}_c^\mathsf{T}\|_\mathrm{F} \right\}. \tag{11}$$

By the triangle inequality

$$\begin{aligned} &\max\left\{ \|\mathbf{\Phi}_r \mathbf{B}_t\|_\mathrm{F}, \|\mathbf{B}_t \mathbf{\Phi}_c^\mathsf{T}\|_\mathrm{F} \right\} \\ &\leq \max\left\{ \sum_{(i,j) \in \mathcal{E}_r} \|\mathbf{L}_{ij} \mathbf{b}_t\|_2, \sum_{(i,j) \in \mathcal{E}_c} \|\tilde{\mathbf{L}}_{ij} \mathbf{b}_t\|_2 \right\}. \end{aligned} \tag{12}$$

Let $M = \max\{|\mathcal{E}_r|, |\mathcal{E}_c|\}$ then

$$\begin{aligned} &\max\left\{ \sum_{(i,j) \in \mathcal{E}_r} \|\mathbf{L}_{ij} \mathbf{b}_t\|_2, \sum_{(i,j) \in \mathcal{E}_c} \|\tilde{\mathbf{L}}_{ij} \mathbf{b}_t\|_2 \right\} \\ &\leq M \max\left\{ \max_{(i,j) \in \mathcal{E}_r} \|\mathbf{L}_{ij} \mathbf{b}_t\|_2, \max_{(i,j) \in \mathcal{E}_c} \|\tilde{\mathbf{L}}_{ij} \mathbf{b}_t\|_2 \right\}. \end{aligned} \tag{13}$$

Putting inequalities (10), (11), (12), and (13) together gives us

$$\frac{\eta \sigma_{\min}}{M} \|\mathbf{B}_t\|_\mathrm{F} \leq \max\left\{ \max_{(i,j) \in \mathcal{E}_r} \|\mathbf{L}_{ij} \mathbf{b}_t\|_2, \max_{(i,j) \in \mathcal{E}_c} \|\tilde{\mathbf{L}}_{ij} \mathbf{b}_t\|_2 \right\}. \tag{14}$$

Since Ω is increasing according to Assumption 1.2, it follows that

$$\begin{aligned} &\Omega\left(\frac{\eta \sigma_{\min}}{M} \|\mathbf{B}_t\|_\mathrm{F} \right) \\ &\leq \max\left\{ \Omega\left(\max_{(i,j) \in \mathcal{E}_r} \|\mathbf{L}_{ij} \mathbf{b}_t\|_2 \right), \Omega\left(\max_{(i,j) \in \mathcal{E}_c} \|\tilde{\mathbf{L}}_{ij} \mathbf{b}_t\|_2 \right) \right\}. \end{aligned} \tag{15}$$

Inequality (15) implies that

$$\begin{aligned} &\min\{\gamma_r, \gamma_c\} M \Omega\left(\frac{\eta \sigma_{\min}}{M} \|\mathbf{B}_t\|_\mathrm{F} \right) \\ &\leq \min\{\gamma_r, \gamma_c\} \max\{J_r(\mathbf{U}_t), J_c(\mathbf{U}_t)\} \\ &\leq \gamma_r J_r(\mathbf{U}_t) + \gamma_c J_c(\mathbf{U}_t). \end{aligned}$$

Consequently, since Ω is increasing and $\|\mathbf{B}_t\|_\mathrm{F} \to \infty$ implies that $f(\mathbf{U}_t) \to \infty$.

Case II: Suppose $\|\mathbf{B}_t\|_\mathrm{F} \leq B$ for some B. Then $|a_t| \to \infty$. Note that we have the following inequality

$$\begin{aligned} f(\mathbf{U}_t) &\geq \sum_{(i,j) \in \Theta} (x_{ij} - b_{k,ij} - a_t)^2 \\ &\geq \sum_{(i,j) \in \Theta} a_t^2 - 2a_t(x_{ij} - b_{k,ij}) \\ &= |\Theta| a_t^2 - 2a_t \sum_{(i,j) \in \Theta} (x_{ij} - b_{k,ij}) \\ &\geq |\Theta| a_t^2 - 2a_t \sup_{\|\mathbf{B}_t\|_\mathrm{F} \leq B} \sum_{(i,j) \in \Theta} (x_{ij} - b_{k,ij}) \\ &= |\Theta| \left[a_t^2 - 2a_t C \right] \\ &= |\Theta| \left[(a_t - C)^2 - C^2 \right], \end{aligned}$$

where $C = |\Theta|^{-1} \sup_{\|\mathbf{B}_t\|_\mathrm{F} \leq B} \sum_{(i,j) \in \Theta} (x_{ij} - b_{k,ij})$.

The function $(a_t - C)^2$ diverges since $|a_t| \to \infty$. Therefore, the function f is coercive.

4. Filling in missing data

We present the original underlying structure of 3D points used to generate the Euclidean distance matrix $\mathbf{X}$ for the datasets **linkage** and **linkage2** in Figure 2 and Figure 5. In Figure 3 and Figure 6, on the left we plot the original complete matrix where the rows and columns have been ordered according to the geometry of the 3D points. On the right we plot the matrix we analyze whose rows and columns have been permuted and 50% of the entries have been removed. In Figure 4 and Figure 7 we display the matrix $\tilde{\mathbf{X}}^{(l,k)}$ for three pairs of values l, k to demonstrate the smoothing that is occurring across the different scales of the rows and columns.

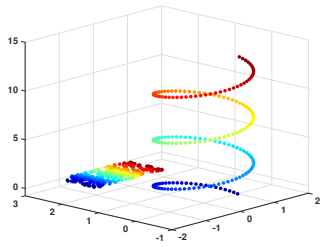

Figure 2: Points in 3D used to generate the Euclidean distance matrix $\mathbf{X}$ in the **linkage** dataset. Rows correspond to the helix, columns to the 2D surface. Points are colored corresponding to the embedding of rows and columns in Figure 2 of the main text.

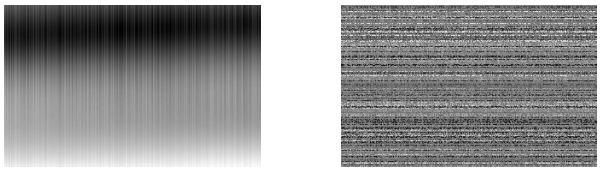

Figure 3: **linkage** dataset: (Left) Complete matrix $\mathbf{X}$. (Right) Matrix whose rows and columns and columns have been permuted and 50% of the values have been removed.

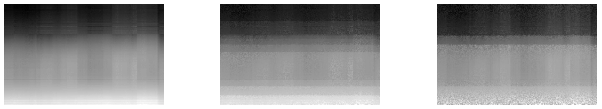

Figure 4: **linkage** dataset: Filled-in matrices $\tilde{\mathbf{X}}$ at multiple scales: $\tilde{\mathbf{X}}^{(-3,-2)}$,$\tilde{\mathbf{X}}^{(1,0)}$,$\tilde{\mathbf{X}}^{(5,2)}$. Rows and columns have been reordered based on the manifold embedding following (Ankenman, 2014).

5. Metric distortion

To further evaluate the different methods beyond just clustering accuracy, we calculate a metric distortion measure with respect to the underlying parametrization of the geometry. The distortion arises from two sources: the missing data and the embedding method itself.

Denote $\boldsymbol{\Theta}$ as the underlying parametrization, and let Ψ be the embedding of the observed data to a low-dimensional space. The metric distortion due to the embedding is

$$\text{distortion}(\Psi) = \text{expansion}(\Psi) \times \text{contraction}(\Psi),$$

where

$$\text{expansion}(\Psi) = \max_{\theta_i,\theta_j \in \Theta} \frac{\|\Psi(i) - \Psi(j)\|_2}{\|\theta_i - \theta_j\|_2}$$

and

$$\text{contraction}(\Psi) = \max_{\theta_i,\theta_j \in \Theta} \frac{\|\theta_i - \theta_j\|_2}{\|\Psi(i) - \Psi(j)\|_2}.$$

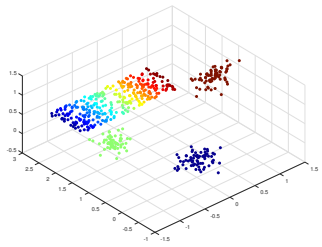

Figure 5: Points in 3D used to generate the Euclidean distance $\mathbf{X}$ in the **linkage2** dataset. Rows correspond to the three 3D Gaussians, columns to the 2D surface. Points are colored corresponding to the embedding of rows and columns in Figure 2 of the main text.

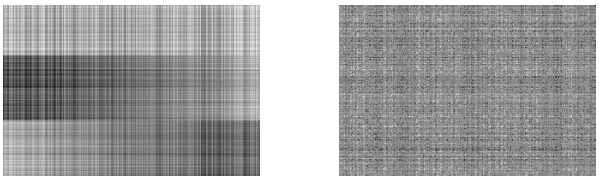

Figure 6: **linkage2** dataset: (Left) Complete matrix $\mathbf{X}$. (Right) Matrix whose rows and columns and columns have been permuted and 50% of the values have been removed.

We plot the distortion of the different methods for the **linkage** and **linkage2** datasets with respect to their underlying geometries (1D helix, 2D surface, 3D clusters) in Figures 8–9. The least distortion is achieved by diffusion maps on the full data (without missing entries) which we plot as a baseline (black plot). The co-manifold approach out-performs Diffusion maps with missing data and both linear embeddings of FRPCAG. For the **linkage** dataset NLPCA out-performs co-manifold but for the **linkage2** dataset co-manifold out-performs NLPCA up to high percentage of missing values.

As diffusion maps itself introduces a distortion, we also plot the metric distortion of diffusion maps with missing data and the co-manifold embeddings with respect to the diffusion distance on the full data (without missing entries) in Figures 10–11. For both datasets, the diffusion embedding yielded by the co-manifold approach introduces less distortion than the the diffusion embedding of the data with missing values.

References

Jerrod I. Ankenman. *Geometry and Analysis of Dual Networks on Questionnaires*. PhD thesis, Yale University, 2014.

Eric C. Chi, Genevera I. Allen, and Richard G. Baraniuk. Convex Biclustering. *Biometrics*, 73(1):10–19, 2017.

Kenneth Lange, David R. Hunter, and Ilsoon Yang. Op-

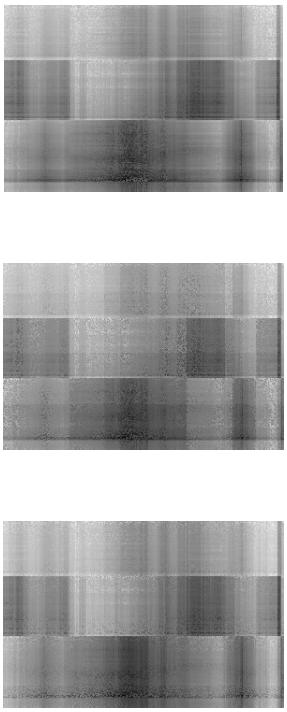

Figure 7: **linkage2** dataset: Filled-in matrices $\tilde{\mathbf{X}}$ at multiple scales: $\tilde{\mathbf{X}}^{(-4,-3)}$,$\tilde{\mathbf{X}}^{(-1,1)}$,$\tilde{\mathbf{X}}^{(5,-3)}$. Rows and columns have been reordered based on the manifold embedding following (Ankenman, 2014).

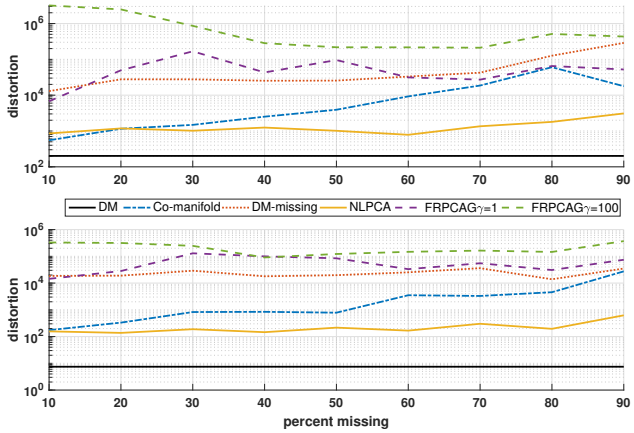

Figure 8: Metric distortion of embedding **linkage** dataset with respect to the 1D helix (top) and 2D surface (bottom).

timization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, 9(1): 1–20, 2000.

Robert R. Meyer. Sufficient conditions for the convergence of monotonic mathematical programming algorithms. *Journal of Computer and System Sciences*, 12(1): 108–121, 1976.

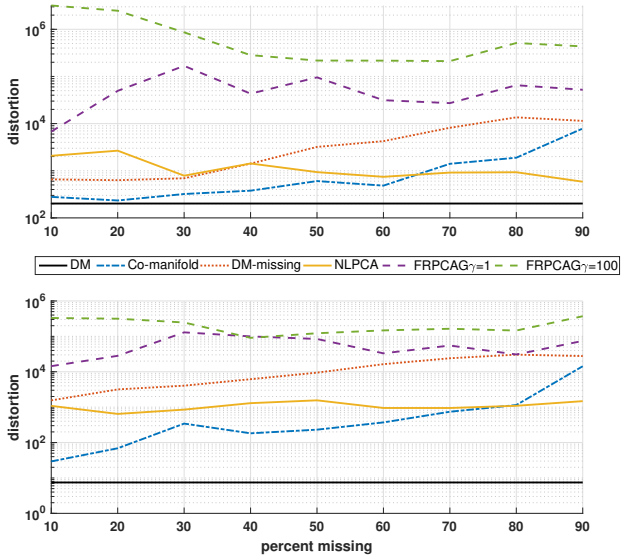

Figure 9: Metric distortion of embedding **linkage2** dataset with respect to the 3D Gaussian clusters (top) and 2D surface (bottom).

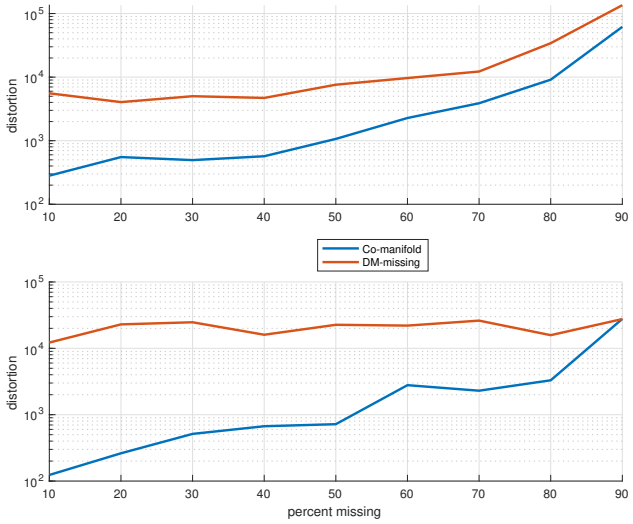

Figure 10: Metric distortion of embedding **linkage** dataset with increasing missing values with respect to diffusion maps without missing values. 1D helix (top) and 2D surface (bottom).

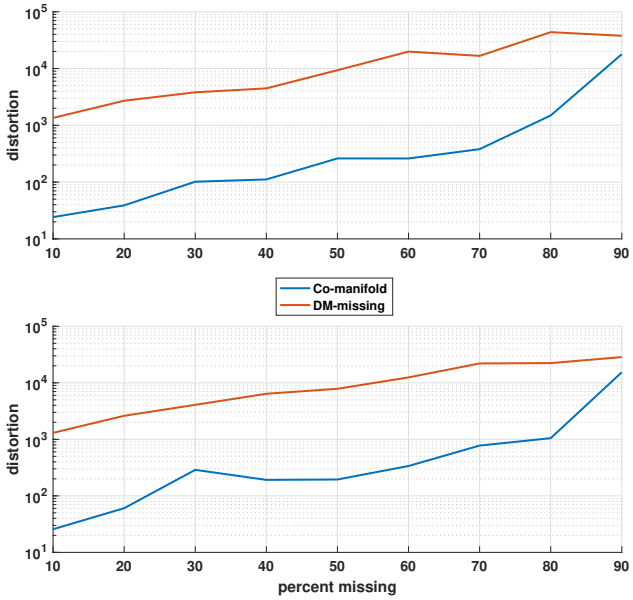

Figure 11: Metric distortion of embedding **linkage2** dataset with increasing missing values with respect to diffusion maps without missing values. 3D Gaussian clusters (top) and 2D surface (bottom)