MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Nonrigid Embeddings for Dimensionality Reduction

Matthew Brand

TR2005-117 December 2005

Abstract

Spectral methods for embedding graphs and immersing data manifolds in low-dimensional speaces are notoriously unstable due to insufficient and/or numberically ill-conditioned constraint sets. Why show shy this is endemic to spectral methods, and develop low-complexity solutions for stiffening ill-conditioned problems and regulatizing ill-posed problems, with proofs of correctness. The regularization exploits sparse but complementary constraints on affine rigidity and edge lengths to obtain isometric embeddings. Am implemented algorithm is fast, accurate and industrial-strength: Experiments with problem sizes spanning four orders of magnitude show O (N) scaling. We demonstrate with speech data.

European Conference on Machine Learning (ECML)

Nonrigid embeddings for dimensionality reduction

Matthew Brand

Mitsubishi Electric Research Labs, Cambridge MA USA

Abstract. Spectral methods for embedding graphs and immersing data manifolds in low-dimensional spaces are notoriously unstable due to insufficient and/or numerically ill-conditioned constraint sets. Why show why this is endemic to spectral methods, and develop low-complexity solutions for stiffening ill-conditioned problems and regularizing ill-posed problems, with proofs of correctness. The regularization exploits sparse but complementary constraints on affine rigidity and edge lengths to obtain isometric embeddings. An implemented algorithm is fast, accurate, and industrial-strength: Experiments with problem sizes spanning four orders of magnitude show $O(N)$ scaling. We demonstrate with speech data.

1 Introduction

Embedding a graph under metric constraints is a central operation in nonlinear dimensionality reduction (NLDR), ad-hoc wireless network mapping, and visualization of relational data. Despite a recent wave of advances in spectral embeddings, it has not yet become a practical, reliable tool. At root is the difficulty of automatically generating embedding constraints that make the problem well-posed, well-conditioned, and solvable on practical time-scales. Well-posed constraints guarantee a unique solution. Well-conditioned constraints make the solution numerically separable from poor solutions. Spectral embeddings from local constraints are frequently ill-posed and almost always ill-conditioned. Both problems manifest as a tiny or zero eigengap in the spectrum of the embedding constraints, indicating that the graph is effectively *nonrigid* and there is an eigen-space of solutions whose optimality is numerically indistinguishable.

Section 2 shows why small eigengaps are endemic to spectral methods for combining local constraints, making it numerically infeasible to separate a solution from its modes of deformation. To remedy this, section 3 presents a linear-time method for stiffening an ill-conditioned problem at all scales, and prove that it inflates the eigengap between the space of optimal solutions and the space of suboptimal deformations.

If a problem is ill-posed, the graph is qualitatively nonrigid and the space of optimal solutions spans all of its degrees of freedom. Section 4 shows how to choose the most dispersed embedding from this space in a semidefinite programming problem (SDP) with a small number of variables and constraints, and proves feasability. Although SDP for graphs has $O(N^6)$ complexity, our methods give a problem reduction that yields embeddings of very large graphs in a matter of seconds or minutes, making million-point problems practical on a ordinary consumer PC.

2 Setting

This paper considers the family of Laplacian-like *local-to-global* graph embeddings, where the embedding of each graph vertex is constrained by the embeddings of its im-

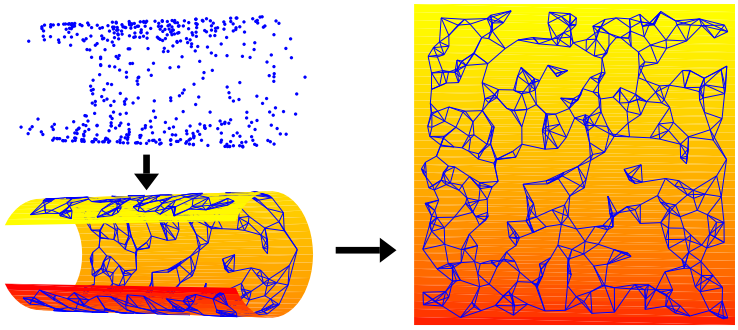

Fig. 1. $N = 500$ points are randomly sampled from a square patch of a cylindrical surface in $\mathbb{R}^{D=3}$, and connected in a $k = 4$ nearest neighbors graph which is then isometrically embedded in $\mathbb{R}^{d=2}$. Spectral embedding methods preserve affine structure of local star-shaped neighborhoods; convex optimization methods preserve edge lengths. Neither is sufficient for sparse graphs, while more densely connected graphs present exploding compute costs and/or may not embed without distortion and folds. Sparse graphs also yield numerically ill-conditioned problems. This paper shows how to obtain well-conditioned problems from very sparse neighborhood graphs and combine them with distance constraints to obtain high quality solutions in linear time.

mediate neighbors (in graph terminology, its 1-ring). For dimensionality reduction, the vertices are datapoints that are viewed as samples from a manifold that is somehow curled up in the ambient sample space, and the graph embedding constraints are designed to reproduce local affine structure of that manifold while unfurling it in a lower dimensional target space. Examples include Tutte's method [Tut63], Laplacian eigenmaps [BN02], locally linear embeddings (LLE) [RS00], Hessian LLE [DG03], charting [Bra03], linear tangent-space alignment (LTSA) [ZZ03], and geodesic nullspace analysis (GNA) [Bra04]. The last three methods construct local affine constraints of maximal possible rank, leading to the stablest solutions. Due to their simplicity, our analysis will be couched in terms of LTSA and GNA. All other methods employ an subset of their affine constraints, so our results will be applicable to the entire family of embeddings.

LTSA and GNA take an N-vertex graph already embedded in an ambient space $\mathbb{R}^D$ with vertex positions $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N] \in \mathbb{R}^{D \times N}$, and re-embed it in a lower-dimensional space $\mathbb{R}^d$ with new vertex positions $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_N] \in \mathbb{R}^{d \times N}$, preserving local affine structure. Typically the graph is constructed from point data by some heuristic such as k-nearest neighbors. The embedding works as follows: Take one such neighborhood of k points and construct a local d-dimensional coordinate system $\mathbf{X}_m \doteq [\mathbf{x}_i, \mathbf{x}_j, \cdots] \in \mathbb{R}^{d \times k}$, perhaps by local principal components analysis. Now consider the nullspace matrix $\mathbf{Q}_m \in \mathbb{R}^{k \times (k-d-1)}$, whose orthonormal columns are orthogonal to the rows of $\mathbf{X}_m$ and to the constant vector $\mathbf{1}$. This nullspace is also orthogonal to any affine transform $A(\mathbf{X}_m)$ of the local coordinate system, such that any translation, rotation, or stretch that preserves parallel lines in the local coordinate system will satisfy $A(\mathbf{X}_m)\mathbf{Q}_m = \mathbf{0}$. Any other transform $T(\mathbf{X}_m)$ can then be separated into an affine component $A(\mathbf{X}_m)$ plus a nonlinear distortion, $N(\mathbf{X}_m) = T(\mathbf{X}_m)\mathbf{Q}_m\mathbf{Q}_m^\top$. The algorithm LTSA (resp. GNA) assembles these *nullspace projectors* $\mathbf{Q}_m\mathbf{Q}_m^\top$, $m = 1, 2, \cdots$ into a sparse matrix $\mathbf{K} \in \mathbb{R}^{N \times N}$ that sums

(resp. averages with weights) nonlinear distortions over all neighborhoods in the graph. Now let $\mathbf{V} \in \mathbb{R}^{d \times N}$ have row vectors that are orthonormal and that span the the column nullspace of $[\mathbf{K}, \mathbf{1}]$; i.e., $\mathbf{V}\mathbf{V}^\top = \mathbf{I}$ and $\mathbf{V}[\mathbf{K}, \mathbf{1}] = \mathbf{0}$. It follows immediately that if $\mathbf{V}$ exists and we use it as a basis for embedding the graph in $\mathbb{R}^d$, each neighborhood in that embedding will have *zero nonlinear distortion* with respect to its original local coordinate systems [ZZ03]. Furthermore, if the neighborhoods are sufficiently overlapped to make the graph affinely rigid in $\mathbb{R}^d$, the transform from the original data $\mathbf{X}$ to the embedding basis $\mathbf{V}$ must stretch every neighborhood *the same way* [Bra04]. Then we can estimate a linear transform $\mathbf{T} \in \mathbb{R}^{d \times d}$ that removes this stretch giving $\mathbf{Y} = \mathbf{T}\mathbf{V}$, such that the transform from $\mathbf{X}$ to $\mathbf{Y}$ involves only rigid transforms of local neighborhoods [Bra04]. I.e., the embedding $\mathbf{Y}$ is isometric.

When there is any kind of noise or measurement error in this process, a least-squares optimal approximate basis $\mathbf{V}$ can be obtained via thin SVD of $\mathbf{K} \in \mathbb{R}^{N \times N}$ or thin EVD of $\mathbf{K}\mathbf{K}^\top$. Because $\mathbf{K}$ is very sparse with $O(N)$ nonzero values, iterative subspace estimators typically exhibit $O(N)$ time scaling. When $\mathbf{K}$ is built with GNA, the corresponding singular values $\sigma_{N-1}, \sigma_{N-2}, \cdots$ measure the pointwise average distortion per dimension.

One of the central problems of this paper is that the eigenvalues of $\mathbf{K}\mathbf{K}^\top$—and indeed of *any* constraint matrix in local NLDR—grow quadratically near $\lambda_0 = 0$, which is the end of the spectrum that furnishes the embedding basis $\mathbf{V}$. (A proof is given in the first two propositions in the appendix.) Quadratic growth means that the eigenvalue curve is almost flat at the low end of the spectrum ($\lambda_{i+1} - \lambda_i \approx 0$) such that the eigengap that separates the embedding basis from other eigenvectors is negligible. A similar phenomenon is observed in the spectra of simple *graph* Laplacians[1] which are also sigmoidal with quadratic growth near zero.

3 Stiffening ill-conditioned problems with multiscale constraints

In graph embeddings the constraint matrix plays a role akin to the stiffness matrix in finite-element methods, and in both cases the eigenvectors associated with the near-zero eigenvalues specify an optimal parameterization and its modes of vibration. The problem facing the eigensolver (or any other estimator of the nullspace) is that convergence rate is a linear function of the relative eigengap $\frac{|\lambda_c - \lambda_{c+1}|}{\lambda_{\max} - \lambda_{\min}}$ or eigenratio $\frac{\lambda_{c+1}}{\lambda_c}$ between the desired and remaining principle eigenvalues [Kny01]. The numerical stability of the eigenvectors similarly depends on the eigengap [SS90]. As just noted, in local-to-global NLDR the eigengap and eigenratio are both very small, making it hard to separate the solution from its distorting modes of vibration. Intuitively, low-frequency vibrations make very smooth bends in the graph, which incur very small deformation penalties at the local constraint level. Since the eigenvalues sum these penalties, the eigenvalues associated with low-frequency modes of deformation have very small values, leading to poor numerical conditioning and slow convergence of eigensolvers. The problem gets much worse for large problems where fine neighborhood structure makes for closely spaced eigenvalues, making it impossible for iterative eigensolvers to accurately compute the smallest eigenvalues and vectors.

We propose to solve this problem by stiffening the mesh with longer-range constraints that damp out lower-frequency vibrations. This can be done without looking

[1] E.g., see `http://www.cs.berkeley.edu/~demmel/ cs267/lecture20/lecture20.html`

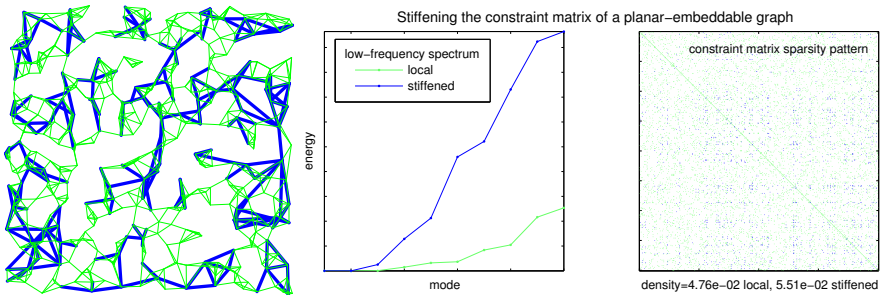

Fig. 2. Stiffening the embedding constraint matrix drives up the eigenvalues associated with low-frequency bending modes. In this example, the constraint matrix is derived from $N = 500$ points forming a 2D manifold embedded in $\mathbb{R}^{256}$. The original graph (green) is shown in green superimposed on a random multiscale stiffening (blue). The low-frequency tail of the eigenspectrum is plotted at center, before (green) and after (blue) stiffening. (The eigenvalue associated with the constant eigenvector $\mathbf{v}_0 = N^{-1/2} \cdot \mathbf{1}$ is suppressed.) The eigengap between the true 2D nullspace and the remaining approximate nullspace is improved by almost 2 orders of magnitude, whereas the original spectrum appears to have a 3D nullspace. The price is a modest 15% increase in constraint matrix density, shown at right as dark blue dots superimposed on the original sparsity pattern. However, the subspace computation is better conditioned and converges four times faster.

---

at the point data. Indeed, it must, because long-range distances in the ambient space are presumed to be untrustworthy. Instead we combine short-range constraints from overlapping rings in the graph, as follows:

ALGORITHM: Neighborhood expansion
1. Select a subgraph consisting of a small set of overlapped neighborhoods and compute an basis $\mathbf{V}_{subgraph}$ for embedding its points in $\mathbb{R}^d$.
2. Form a new neighborhood with at least d+1 points taken from the embedding basis and add (LTSA) or average (GNA) its nullspace projector into $\mathbf{K}$.

Because the $\mathbf{K}$ matrix penalizes distortions in proportion to the distances between the points, these larger-scale constraints can significantly drive up the eigenvalues outside the nullspace, enlarging the eigengap. It can be shown that

Proposition 1. *The nullspace of* $\mathbf{K}$ *is invariant to neighborhood expansions.*

See the appendix for all proofs. Neighborhood expansion is physically analogous to adding short ribs to a 2D plate to stiffen it against small-radius bends in 3D. However, in order to usefully improve the eigengap, one must brace against large-radius bends. Fortunately, stiffening lends itself very naturally to a multiscale scheme: We construct a set of neighborhood expansions that approximately covers the graph but adds constraints on just a small subset of all vertices. Note that this subset of vertices plus their parameterizations in the new neighborhoods constitutes a new embedding problem. Thus we may recursively stiffen this problem in the same manner, and so on until the original problem is stiffened at all scales:

ALGORITHM: Multiscale stiffening
1. Choose a constant fraction of vertices to be anchors.
2. Cover or partially cover the data with neighborhood expansions, adding constraints on any anchors that fall in an expansion.
3. Recurse only on the anchors, using their parameterizations in the neighborhood expansions.

Proposition 2. *If the number of neighborhoods and points is halved at each recursion, multiscale stiffening can be performed in $O(N)$ time with no more than a doubling of the number of nonzeros in the* **K** *matrix.*

For modern iterative nullspace estimators (e.g., LOBPCG [Kny01]), compute time of each iteration is typically linear in the number of nonzeros in **K** while convergence rate is supra-linear in the eigengap. Consequently, stiffening is a winning proposition. Figure 2 shows a simple example where stiffening the graph in figure 1 makes the spectrum rank-revealing and cuts the EVD time by 3/4. However, due to the difficulty of implementing the appropriate data structures efficiently in Matlab, there was no reduction in overall "wall time".

4 Regularizing ill-posed problems with edge length constraints

Even if the eigenvector problem is numerically well-conditioned, it may be the case that the graph is intrinsically nonrigid. This commonly happens when the graph is generated by a heuristic such as k-nearest neighbors. In such cases the embedding basis $\mathbf{V} \in \mathbb{R}^{c \times N}$ has greater dimension c than desired ($c > d$). For example, the initial constraints might allow for a variety of folds in $\mathbb{R}^d$, then **V** must span all possible folded configurations. The embedding is thus ill-posed, and some regularization is needed to choose from the space of possible embeddings. We will presume that in the most unfolded configuration, some subset of vertices are maximally dispersed. For example, we might maximize the distance between each vertex and all of its 4-hop neighbors. In order to prevent the trivial solution of an infinitely large embedding, we must fix the scale in each dimension by fixing some distances, i.e., edge lengths. Thus we seek an embedding that satisfies the affine constraints encoded in the **K** matrix, maximizes distances between a mutually repelling subset of vertices, and satisfies exact distance constraints on some subset of edges. For this we adapt the semidefinite graph embedding of [LLR95].

Formally, let mixing matrix $\mathbf{U} \in \mathbb{R}^{c \times d}$ have orthogonal columns of arbitrary nonzero norm. Let error vector $\sigma = [\sigma_1, \cdots, \sigma_c]^\top$ contain the singular values of distortion matrix **K** associated with its left singular vectors, the rows of **V**. The matrix **U** will select a metrically correct embedding from the space of possible solutions spanned by the rows of **V**. The target embedding, $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_N] \doteq \mathbf{U}^\top \mathbf{V} \in \mathbb{R}^{d \times N}$, will have overall distortion $\|\mathbf{U}^\top \sigma\|$ and distance $\|\mathbf{y}_i - \mathbf{y}_j\| = \|\mathbf{U}^\top (\mathbf{v}_i - \mathbf{v}_j)\|$ between any two points ($\mathbf{v}_i$ being the ith column of **V**). The optimization problem is to minimize the distortion while maximizing the dispersion

$$\mathbf{U}^* = \max_{\mathbf{U}} -\|\mathbf{U}^\top \sigma\|^2 + \sum_{pq} r_{pq}^2 \|\mathbf{y}_p - \mathbf{y}_q\|^2 \tag{1}$$

for some choice of weights $r_{pq} \geq 0$, preserving distances

$$\forall_{ij \in \text{EdgeSubset}} \|\mathbf{y}_i - \mathbf{y}_j\| \leq D_{ij} \tag{2}$$

on at least d edges forming a simplex of nonzero volume in $\mathbb{R}^d$ (otherwise the embedding can collapse in some dimensions). We use inequality instead of equality because the D_{ij}, measured as straight-line distances, are chordal in the ambient space $\mathbb{R}^D$ rather than geodesic in the manifold, and thus may be inconsistent with a low dimensional embedding (or infeasible). The inequality allows some edges to be slightly shortened in favor of more dispersed and thus flatter, lower-dimensional embeddings. In general, we will enforce distance constraints corresponding to all or a random sample of the edges in the graph. Unlike [LLR95] (and [WSS04], discussed below), *the distance constraints do not have to form a connected graph*.

Using the identity $\|\mathbf{Y}\|_F^2 = \|\mathbf{U}^\top\mathbf{V}\|_F^2 = \text{trace}(\mathbf{U}^\top\mathbf{V}\mathbf{V}^\top\mathbf{U}) = \text{trace}(\mathbf{V}\mathbf{V}^\top\mathbf{U}\mathbf{U}^\top)$, we massage eqns. 1-2 into a small semidefinite program (SDP) on objective $\mathbf{G} \doteq \mathbf{U}\mathbf{U}^\top \succ \mathbf{0}$:

$$\max_{\mathbf{G}} \text{trace}((\mathbf{C} - \text{diag}(\sigma)^2)\mathbf{G}) \tag{3}$$

$$\text{with } \mathbf{C} \doteq \sum_{pq} r_{pq}^2 (\mathbf{v}_p - \mathbf{v}_q)(\mathbf{v}_p - \mathbf{v}_q)^\top \tag{4}$$

$$\text{subject to } \forall_{i,j \in \text{EdgeSubset}} \text{trace}((\mathbf{v}_i - \mathbf{v}_j)(\mathbf{v}_i - \mathbf{v}_j)^\top \mathbf{G}) \leq D_{ij}^2 . \tag{5}$$

In particular, when all points repel equally ($\forall_{pq} r_{pq} = 1$), then $\mathbf{C} = \mathbf{V}\mathbf{V}^\top = \mathbf{I}$, and $\text{trace}(\mathbf{C}\mathbf{G}) = \sum_{pq} \|\mathbf{y}_p - \mathbf{y}_q\|^2 = \|\mathbf{Y}\|_F^2$. Because $\mathbf{V} \perp \mathbf{1}$, the embedding is centered.

At the extreme of $c = d$, we recover pure LTSA/GNA, where $\mathbf{U} = \mathbf{T}$ is the upgrade to isometry (the SDP is unnecessary). At $c = D - 1$ we have an alternate formulation of the semidefinite graph embedding [LLR95], where $\text{range}(\mathbf{V}) = \text{span}(\mathbb{R}^N \perp \mathbf{1})$ replaces the centering constraints (the LTSA/GNA is unnecessary). In between we have a blend that we will call Nonrigid Alignment (NA). With iterative eigensolving, LTSA/GNA takes $O(N)$ time, but requires a globally rigid set of constraints. The semidefinite graph embedding does not require rigid constraints, but has $O(N^6)$ time scaling. Nonrigid Alignment combines the best of these methods by using LTSA/GNA to construct a basis that drastically reduces the semidefinite program. In addition, we have the option of combining an incomplete set of neighborhoods with an incomplete set of edge length constraints, further reducing both problems. (A forthcoming paper will detail which subsets of constraints guarantee affine rigidity.)

Although this method does require an estimate of the local dimension for the initial LTSA/GNA, it inherits from semidefinite graph embeddings the property that the spectrum of $\mathbf{X}$ gives a sharp estimate of the global embedding dimension, because the embedding is spanned by $\mathbf{V}$. In fact, one can safely over-estimate the local dimension—this reduces the local nullspace dimension and thus the global rigidity, but the additional degrees of freedom are then fixed in the SDP problem.

4.1 Reducing the SDP constraints

The SDP equality constraints can be rewritten in matrix-vector form as $\mathbf{A}^\top \text{svec}(\mathbf{G}) = \mathbf{b}$, where $\text{svec}(\mathbf{G})$ forms a column vector from the upper triangle of $\mathbf{X}$ with the off-diagonal elements multiplied by $\sqrt{2}$. Here each column of $\mathbf{A}$ contains a vectorized edge length constraint (e.g., $\text{svec}((\mathbf{v}_i - \mathbf{v}_j)(\mathbf{v}_i - \mathbf{v}_j)^\top)$ for an equality constraint) for some edge $i \leftrightarrow j$; the corresponding element of vector $\mathbf{b}$ contains the value D_{ij}^2. A major cost of the SDP solver lies in operations on the matrix $\mathbf{A} \in \mathbb{R}^{c^2 \times e}$, which may have a large number of

linearly redundant columns. Note that c^2 is relatively small due to the choice of basis, but e, the number of edges whose distance constraints are used in the SDP, might be very large. When the problem has an exact solution (equation 5 is feasible as an equality), this cost can be reduced by projection: Let $\mathbf{F} \in \mathbb{R}^{e \times f}$, $f \ll e$ be a column-orthogonal basis for the principal row-subspace of $\mathbf{A}$, which can be estimated in $O(ef^2c^2)$ time via thin SVD. From the Mirsky-Eckart theorem it trivially follows that the f equality constraints,

$$\mathbf{F}^\top \mathbf{A}^\top \mathrm{vec}(\mathbf{G}) = \mathbf{F}^\top \mathbf{b} \tag{6}$$

are either equivalent to or a least-squares optimal approximation of the original equality constraints. In our experience, for large, exactly solvable problems, it is not unusual to reduce the cardinality of constraint set by 97% without loss of information.

Proposition 3. *The resulting* SDP *problem is feasible.*

When the problem does not have an exact solution (equation 5 is only feasible as an inequality), one can solve the SDP problem with a small subset of randomly chosen edge length inequality constraints. In conjunction with the affine constraints imposed by the subspace $\mathbf{V}$, this suffices to satisfy most of the remaining unenforced length constraints. Those that are violated can be added to the active set and the SDP re-solved, possibly repeating until all are satisfied.

These reductions yield a practical algorithm for very large problems:

ALGORITHM: Nonrigid LTSA/GNA
1. Obtain basis: Compute extended approximate nullspace $\mathbf{V}$ and residuals σ_i of (stiffened) $\mathbf{K}$ matrix.
2. SDP: Find $\mathbf{G}$ maximizing eq. 3 subject to eq. 6 or eq. 5 with a constraint subset.
2a. Repeat 2 with violated constraints, if any.
3. Upgrade to isometry: Factor $\mathbf{G} \rightarrow \mathbf{U}\mathrm{diag}(\lambda)^2\mathbf{U}^\top$ and set embedding $\mathbf{Y} = \mathrm{diag}(\lambda)\mathbf{U}^\top\mathbf{V}$.

4.2 Related work

Recently [WSS04] introduced an algorithm that applies the LLR embedding to densely triangulated graphs, and [WPS05] introduced a related scheme called ℓSDE which uses a landmark basis derived from LLE to reduce the semidefinite program. We can highlight some substantial differences between our approach and ℓSDE: 1) Because LLE is quasi-conformal and has no isometry properties, one would expect that a much higher-dimensional LLE basis will be necessary to span the correct isometric embedding (this we have verified numerically), either substantially increasing the SDP time or decreasing solution quality if a lower-dimensional basis is used. 2) If the manifold has nonzero genus or concave boundary, the number of randomly selected landmarks—and thus basis dimensions—needed to span the isometric embedding can grow exponentially; not so for the LTSA/GNA basis, which depends only on local properties of the manifold. 3) graph triangulation increases the number of graph edges by a factor of $k^2$and the complexity of the SDP problem by k^6—a major issue because k itself should grow quadratically with the intrinsic dimension of the manifold. Thus we can solve problems 2 orders of magnitude larger in considerably less time, and report *exact* solutions.

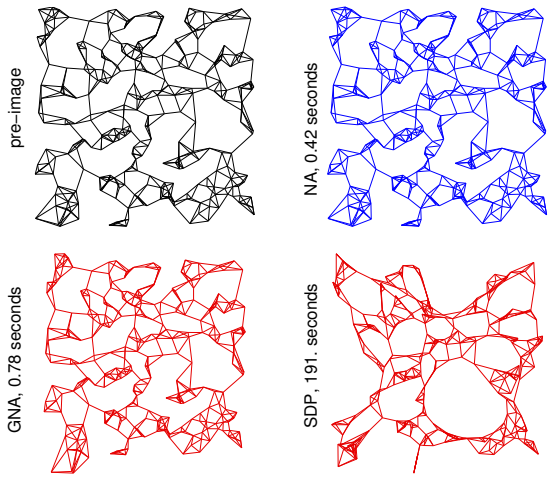

Fig. 3. A 2D NA embedding of a 4-neighbors graph on 300 points in $\mathbb{R}^{256}$ perfectly recovers the pre-image. The LTSA/GNA solution has five affine degrees of freedom associated with the distorted subgraphs on the bottom boundary. The SDP solution "foams" around large cycles where the graph is nonrigid.

4.3 Example

In this example, the source manifold is a square planar patch, which is embedding isometrically in $\mathbb{R}^4$ through the toric map that takes each ordinate $(x) \rightarrow (\sin x, \cos x)$. $\mathbb{R}^4$ is in turn embedded in $\mathbb{R}^8$ by the same map, and so on until the ambient space has $D = 256$ dimensions. The patch is randomly sampled in $\mathbb{R}^D$ and each point connected to its four nearest neighbors. The graph is too sparsely connected to determine a rigid embedding for either LTSA/GNA or the LLR SDP (see figure 3). Nonrigid GNA yields near-perfect embeddings. For example, figure 3 depicts the pre-image and three embeddings of a small $N = 300$ point, $K = 4$ neighbors graph. Ordinary LTSA/GNA has a 7-dimensional nullspace, indicating that some subgraphs have unwanted affine degrees of freedom. This can be resolved by increasing K, but that risks bringing untrusted edge lengths into the constraint set. SDE can fix most (but not necessarily all) of these DOFs by fully triangulating each neighborhood, but that increases the number of edges by a factor of K^2 and the SDP time complexity by a factor of K^6. Even for this small problem NA is almost three orders of magnitude faster than untriangulated SDE; that gap widens rapidly as problem size grows.

Empirically, NA exhibits the predicted linear scaling over a wide range of problem sizes. Working in MatLab on a 3GHz P4 with 1Gbyte memory, 10^2 points took roughly 0.3 seconds;10^3 points took roughly 2 seconds; 10^4 points took 21 seconds; 10^5 points took roughly 232 seconds; we see linear scaling in between. The dominant computation is the EVD, not the SDP.

5 Application to speech data

The TIMIT speech database is a widely available collection of audio waveforms and phonetic transcriptions for 2000+ sentences uttered by 600+ speakers. We sought to

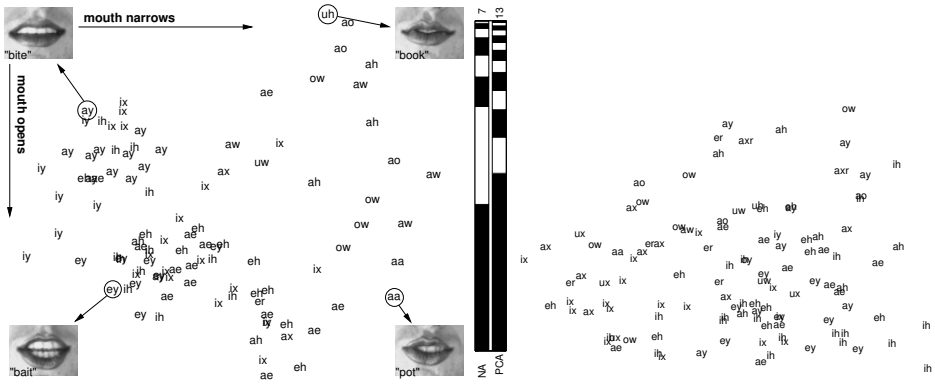

Fig. 4. LEFT: A thin slice along the two principal axes of an NA embedding of 2.5×10^5 vowel feature vectors. TIMIT phoneme labels are scatter-plotted according to their embedding coordinates. The distribution of phonemes is well correlated with mouth shape (see discussion in section 5). MIDDLE: Normalized spectra of the NA and PCA representations, showing the fraction of total variance captured in each dimension. RIGHT: An equivalent slice through the PCA representation slice scatter-plot is far less interpretable. Some sounds (e.g., ix in debit) depend little on lip shape and are thus distributed freely through both plots.

model the space of acoustic variations in vowel sounds. Starting with a standard representation, we computed a vector of $D = 13$ mel-cepstral features for each 10 millisecond frame that was labelled as a vowel in the transcriptions. To reduce the impact of transcription errors and co-articulatory phenomena, we narrowed the data to the middle half of each vowel segment, yielding roughly $N = 240,000$ samples in $\mathbb{R}^{13}$. Multiple applications of PCA to random data neighborhoods suggested that the data is locally 5-dimensional. An NA embedding of the 7 approximately-nearest neighbors graph with 5-dimensional neighborhoods and a 25-dimensional basis took slightly less than 11 minutes to compute. The spectrum is sharp, with >99% of the variance in 7 dimensions, >95% in 5 dimensions, and >75% in 2 dimensions. A PCA rotation of the raw data matches these percentages at 13, 9, and 4 dimensions respectively. Noting the discrepancy between the estimated local dimensionality and global embedding dimension, we introduced slack variables with low penalties to explore the possibility that the graph was not completely unfolding. Since this left the spectrum substantially unchanged, we conjecture that there may be topological loops or unnoticed 7-dimensional clusters, and indeed some projections of the embedding showed holes.

Figure 4 shows how the phonemes are organized in the two principal dimensions of the NA and PCA representations. The NA axes are clearly correlated with the physical degrees of freedom of the speech apparatus: Roughly speaking, as one moves to the right the mouth narrows horizontally, from iy (beet) and ey (bait) to ao (bought) and aw (bout); as one moves up the mouth narrows vertically with the lower lip moving forward and upward, from ah (but) and eh (bet) to ow (boat) and uh (book). The third dimension (not shown) appears to be correlated with the size of the resonant chamber at

the back of the mouth, i.e. tongue position. After considerable study, it is still not clear how to interpret the raw PCA axes.

A low-dimensional representation is advantageous for speech recognition because it makes it practical to model phoneme classes with full covariance Gaussians. A long-standing rule-of-thumb in speech recognition is that a full-covariance Gaussian is competitive with a mixture of 3 or 4 diagonal-covariance Gaussians [LRS83]. The important empirical question is whether the NA representation offers a better separation of the classes than the PCA. This can be quantified (independently of any downstream speech processing) by fitting a Gaussian to each phoneme class and calculating the symmetrized KL-divergence between classes. Higher divergence means that one will need fewer bits to describe classification errors made by a (Gaussian) quadratic classifier. We found that the *divergence between classes in the $d = 5$ NA representation was on average approximately 2.2 times the divergence between classes in the $d = 5$ PCA representation, with no instances where the NA representation was inferior*. Similar advantages were observed for other values of d, even, surprisingly, $d = 1$ and $d = D$.

Even though both representations are unsupervised, we may conclude that preserving short-range metric structure (NA) is more conducive to class separation than preserving long-range distances (PCA). We are now working on a larger embedding of all phonemes which, when combined with the GNA out-of-sample extension, will be incorporated into a speech recognition engine.

6 Discussion

We have demonstrated that rigidity is a key obstacle for viable nonlinear dimensionality reduction, but by stiffening the constraint set and recasting the upgrade to isometry as a small SDP problem, problems that are severely ill-posed and ill-conditioned can be solved—in linear time. At time of submission, we have successfully embedded problems of up to 10^6 points, and it appears that the principal challenge in using these methods will be the most advantageous choice of basis dimension. The is a matter of finding the eigengap of ill-posed problems, and we hope to make connections with an existing literature on large-scale physical eigenproblems. Another issue is the initial problem of graph building—at 10^5 points, the approximate nearest-neighbor algorithms that make graph-building tractable begin to make substantial errors. For NLDR to be practical above 10^7 points—the size of bioinformatic and econometric problems—the problem of reliable graph-building will have to be solved.

References

BN02. Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. volume 14 of *Advances in Neural Information Processing Systems*, 2002.

Bra03. Matthew Brand. Charting a manifold. In *Advances in Neural Information Processing Systems*, volume 15, 2003.

Bra04. Matthew Brand. From subspaces to submanifolds. In *Proceedings, British Machine Vision Conference*, 2004.

DG03. David L. Donoho and Carrie Grimes. Hessian eigenmaps. *Proceedings, National Academy of Sciences*, 2003.

Kny01. A. V. Knyazev. Toward the optimal preconditioned eigensolver. *SIAM Journal on Scientific Computing*, 23(2):517–541, 2001.

LLR95. N. Linial, E. London, and Y. Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245, 1995.

LRS83. S. Levinson, L. Rabiner, and M. Sondhi. An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition. *Bell System Technical Journal*, 62(4):1035–1074, 1983.

RS00. Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, December 22 2000.

SS90. G.W. Stewart and Ji-Guang Sun. *Matrix perturbation theory*. Academic Press, 1990.

Tut63. W.T. Tutte. How to draw a graph. *Proc. London Mathematical Society*, 13:743–768, 1963.

WPS05. K.Q. Weinberger, B.D. Packer, and L.K. Saul. Nonlinear dimensionality reduction by semidefinite programming and kernel matrix factorization. In *Proc. AI & Statistics*, 2005.

WSS04. K. Q. Weinberger, F. Sha, and L. K. Saul. Learning a kernel matrix for nonlinear dimensionality reduction. In *Proc. 21st ICML*, 2004.

ZZ03. Z. Zhang and H. Zha. Nonlinear dimension reduction via local tangent space alignment. In *Proc., Conf. on Intelligent Data Engineering and Automated Learning*, number 2690 in Lecture Notes on Computer Science, pages 477–481. Springer-Verlag, 2003.

A Analysis of local-to-global spectral models and misc. proofs

We can view the constraint matrix $\mathbf{K}$ as a discrete approximation to a convolution of a candidate embedding $\mathbf{Z}$ with a filter If we plot columns of $\mathbf{K}$, this filter resembles an inverted Laplacian. Analysis shows that this is indeed the case:

Proposition 4. *Let $\mathbf{Z} \doteq [\mathbf{z}_1, \cdots, \mathbf{z}_N] \in \mathbb{R}^{d \times N}$ with $\mathbf{z}_i = z(\mathbf{y}_i)$ be a data parameterization given by some C^2 multivalued map $z : \mathcal{M} \to \mathbb{R}^d$ on the intrinsic coordinates $\mathbf{y}_i$. Let*

$$\mathbf{K} \doteq \left(\sum_m \mathbf{S}_m \mathbf{Q}_m \mathbf{Q}_m^\top diag(\mathbf{w}_m) \mathbf{S}_m^\top \right) diag(\sum_m \mathbf{S}_m \mathbf{w}_m)^{-1} \tag{7}$$

where binary indexing matrix $\mathbf{S}_m \in \{0,1\}^{N \times k}$ select k points forming the mth neighborhood and neighborhood weight vector $\mathbf{w}_m \in \mathbb{R}^k$ assigns points weights according to their distance from the neighborhood center: ($\{\mathbf{w}_m\}_i \propto \exp(-\|\{\mathbf{X}_m\}_i - \overline{\mathbf{X}_m}\|^2/2\sigma^2)/\sigma$). Then each column of $\mathbf{K}$ is a discrete difference of Gaussians operator with the parameterization error $\|\mathbf{ZK}\|_F^2$ approximating $\|z - G * z - \bigtriangledown^2 G * z\|^2$, the difference between z and a smoothed version of itself, minus its convolution with a Laplacian-of-Gaussian operator.

Proof. (prop. 4) For simplicity, we will first consider the case of a 1D manifold sampled at regular intervals. Recall that $\mathbf{K}$ is an average of neighborhood nullspace projectors, each of the form $\mathbf{N}_m = \mathbf{Q}_m \mathbf{Q}_m^\top = \mathbf{I} - \frac{1}{k}\mathbf{1}\mathbf{1}^\top - \mathbf{P}_m \mathbf{P}_m^\top$, where $\mathbf{P}_k \in \mathbb{R}^{k \times d}$ is an orthogonal basis of centered local coordinates $\mathbf{X}_m - \overline{\mathbf{X}_m}\mathbf{1}^\top$. Because orthogonalization is a linear operation, $\frac{1}{k} - \{\mathbf{N}_m\}_{i \neq j}$ is proportional to $\|\{\mathbf{X}_m\}_i - \overline{\mathbf{X}_m}\| \cdot \|\{\mathbf{X}_m\}_j - \overline{\mathbf{X}_m}\|$, the product of the distances of points i and j from the clique centroid. Viewing the elements of the matrix $\mathbf{P}_m \mathbf{P}_m^\top$ as surface heights, we have a quadratic saddle surface, maximally positive in the upper left and lower right corners, and maximally negative in the upper right and lower left corners. In our simplified case, $\mathbf{P}_m = k^{-1/2} \cdot [-j, 1-j, \cdots, j-1, j]^\top$ where

$k = 2j+1$ is the size of each neighborhood, and elements in each column of $\mathbf{K}$ are Gaussian-weighted sums along the diagonals of $\mathbf{N}_m$. Precisely, for the pth non-boundary neighborhood, the nth nonzero subdiagonal element in a column of $\mathbf{K}$ is

$$\begin{aligned} K_{p+n,p} &= -\frac{1}{k}\sum_{i=n}^{i=2j}(1+(i-j)(i-j-n)\frac{3}{j(j+1)})e^{-(i-j)^2} \\ &= -\frac{1}{k}\frac{3}{j(j+1)}\sum_{i=n}^{i=2j}\{(1-(i-j)^2)e^{-(i-j)^2} \\ &\qquad -(1-n(i-j))e^{-(i-j)^2}+\frac{j(j+1)}{3}e^{-(i-j)^2}\}. \end{aligned}$$

Note that $(1-(i-j)^2)e^{-(i-j)^2}$ is a Laplacian-of-Gaussian, and that if we hold $i=n$ and iterate over n (the elements of a column in $\mathbf{K}$), we obtain a difference of Gaussians and LoG's, each with finite support; summing over i gives a superposition of these curves, each with a different support. To generalize to non-regular sampling, simply increment i by the difference between neighboring points. To generalize to multidimensional manifolds, note that the above arguments apply to any subset of points forming a geodesic line on $\mathcal{M}$, and by the linearity of $\mathbf{K}$ and the Laplacian operator, to any linear combination of different subsets of points forming different geodesics.

Proposition 5. *The near-zero eigenvalues of $I-G-\bigtriangledown^2 G$ grow quadratically.*

Proof. (prop. 5) Consider the harmonic equation, which describes how the graph vibrates in the space normal to its embedding: $-(I-G-\bigtriangledown^2 G)Y(x,t) = d^2Y(x,t)/d^2t$, with $Y(x,t)$ being the displacement at time t and position x (in manifold-intrinsic coordinates). For periodic motion, set $Y(x,t) = \sin(\omega t)\cdot Y(x)$, with $Y(x)$ being a vibrational mode. After substitution and cancellation, the harmonic equation simplifies to $(I-G-\bigtriangledown^2 G)Y(x) = \omega^2\cdot Y(x)$, confirming that the mode $Y(x)$ is an eigenfunction of the operator $I-G-\bigtriangledown^2 G$. One can verify by substitution that $Y(x)=\sin(ax+b)$ for $a\in\{1,2,\cdots,N\}, b\in\mathbb{R}$ is an orthogonal basis for solutions (eigenvectors) with eigenvalues on the sigmoid curve $\omega^2 = 1-(1+a^2/\sqrt{2\pi})e^{-a^2}$. A series expansion around $a=0$ reveals that the leading term is quadratic.

Proof. (prop. 1) Expansion generates a new neighborhood whose parameterization is affine to those of its constituent neighborhoods, thus its nullspace is orthogonal to $\mathbf{K}$.

Proof. (prop. 2) Because of halving, at any scale the number of vertices in each neighborhood expansion is, on average, a constant $v \ll N$ that is determined only by the intrinsic dimensionality and the average size of the original local neighborhoods. Halving also guarantees that the total number of neighborhood expansions is $\sum_i(\frac{1}{2})^i N < N$. Together these establish $O(N)$ time. In each of the fewer than N neighborhood expansions, a point receives on average d constraints from new neighbors—the same or less than it receives in each of the N original neighborhoods.

Proof. (prop. 3) Since $\mathbf{F}$ is a variance-preserving rotation of the constraints, one can always rotate the f-dimensional row-space of $\mathbf{F}=[\mathbf{f}_1,\cdots,\mathbf{f}_f]$ so that $\forall_i \mathbf{f}_i^\top \mathbf{b} > 0$. Then any infeasible solution $\tilde{\mathbf{G}}$ can be scaled by $z>0$ such that $\forall_i \mathbf{f}_i^\top \mathbf{A}^\top \text{svec}(z\tilde{\mathbf{G}}) \le \mathbf{f}_i^\top \mathbf{b}$, with any differences made up by nonnegative slack variables.