MERL – A MITSUBISHI ELECTRIC RESEARCH LABORATORY
http://www.merl.com

Continuous nonlinear dimensionality reduction by kernel eigenmaps

Matthew Brand

TR-2003-21 April 2003

Abstract

We equate nonlinear dimensionality reduction (NLDR) to graph embedding with side information about the vertices, and derive a solution to either problem in the form of a kernel-based mixture of affine maps from the ambient space to the target space. Unlike most spectral NLDR methods, the central eigenproblem can be made relatively small, and the result is a continuous mapping defined over the entire space, not just the datapoints. A demonstration is made to visualizing the distribution of word usages (as a proxy to word meanings) in a sample of the machine learning literature.

First circulated fall 2002.

Proceedings, International Joint Conference on Artificial Intelligence, IJCAI-2003.

Continuous nonlinear dimensionality reduction by kernel eigenmaps

Matthew Brand
Mitsubishi Electric Research Laboratories
Cambridge, MA 02460 USA

Abstract

We equate nonlinear dimensionality reduction (NLDR) to graph embedding with side information about the vertices, and derive a solution to either problem in the form of a kernel-based mixture of affine maps from the ambient space to the target space. Unlike most spectral NLDR methods, the central eigenproblem can be made relatively small, and the result is a continuous mapping defined over the entire space, not just the datapoints. A demonstration is made to visualizing the distribution of word usages (as a proxy to word meanings) in a sample of the machine learning literature.

1 Background: Graph embeddings

Consider a connected graph with weighted undirected edges specified by edge matrix $\mathbf{W}$. Let $W_{ij} = W_{ji}$ be the positive edge weight between connected vertices i and j, zero otherwise. Let $\mathbf{D} \doteq \text{diag}(\mathbf{W1})$ be a diagonal matrix where $D_{ii} = \sum_j W_{ij}$, the cumulative edge weights into vertex i. The following points are well known or easily derived in spectral graph theory [Fiedler, 1975; Chung, 1997]:

1. The generalized eigenvalue decomposition (EVD)
$$\mathbf{WV} = \mathbf{DV}\Lambda \tag{1}$$
has real eigenvectors $\mathbf{V} \doteq [\mathbf{v}_1, \cdots, \mathbf{v}_N]$ and eigenvalues $\Lambda \doteq \text{diag}([\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N])$.

2. Premultiplying equation (1) by $\mathbf{D}^{-1}$ makes the generalized eigenproblem into a *stochastic* eigenproblem
$$(\mathbf{D}^{-1}\mathbf{W})\mathbf{V} = \mathbf{V}\Lambda, \tag{2}$$
where $\mathbf{D}^{-1}\mathbf{W}$ is a stochastic transition matrix having nonnegative rows that sum to one. The largest eigenvalue of equation (1) is therefore *stochastic* ($\lambda_1 = 1$) and its paired eigenvector is *uniform* ($\mathbf{v}_1 = \mathbf{1}/\sqrt{N}$).

3. Expanding and collecting terms in W_{ij} reveals the geometric meaning of the eigenvalues:
$$\lambda_k = 1 - \textstyle\sum_{ij}(v_{ik} - v_{jk})^2 W_{ij}/2. \tag{3}$$
The d eigenvectors paired to eigenvalues λ_2 through λ_{d+1} therefore give an embedding of the vertices in $\mathbb{R}^d$ with minimal distortion vis-a-vis the weights, in the sense that a larger W_{ij} stipulates a shorter embedding distance. Formally, the embedding
$$\mathbf{Y}_{1:d} \doteq [\mathbf{v}_2, \cdots, \mathbf{v}_{d+1}]^\top = \arg\max_{\mathbf{YDY}^\top = \mathbf{I}} \text{trace}(\mathbf{YWY}^\top) \tag{4}$$
minimizes the distortion $d - \text{trace}(\mathbf{YWY}^\top)$
$$= \sum_{k=2}^{d+1}(1-\lambda_k) = \sum_{k=2}^{d+1}\sum_{ij}(v_{ik} - v_{jk})^2 W_{ij}/2 \tag{5}$$
for any integer $d \in [1, N]$. The norm constraint $\mathbf{YDY}^\top = \mathbf{I}$ sets the scale of the embedding and causes vertices of high cumulative weight to be embedded nearer to the origin.

4. $\mathbf{Y}$ can be rigidly rotated in $\mathbb{R}^d$ without changing its distortion. The distortion measure is also invariant to rigid translations, but the eigenproblem is not, thus there is an unwanted degree of freedom (DOF) in the solution. Due to stochasticity, this DOF is isolated in the uniform eigenvector $\mathbf{v}_1$, which is suppressed from the embedding without error (because $1 - \lambda_1 = 0$). Adding $\mathbf{tv}_1^\top$ to $\mathbf{Y}$ rigidly translates the embedding by $\mathbf{t} \in \mathbb{R}^d$.

5. Premultiplying by $\mathbf{V}^\top$ and rearranging equates equation 1 to the EVD of the graph Laplacian $\mathbf{D} - \mathbf{W}$:
$$\mathbf{V}^\top(\mathbf{D} - \mathbf{W})\mathbf{V} = \mathbf{I} - \Lambda. \tag{6}$$

6. Premultiplying by $\mathbf{D}^{-1/2}$ connects equation 1 to the (symmetric) EVD of the normalized Laplacian:
$$(\mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2})\mathbf{V}' = \mathbf{V}'\Lambda \tag{7}$$
with $\mathbf{V}' \doteq \mathbf{D}^{1/2}\mathbf{V}$.

In summary: Equation 1 gives an optimal embedding of a graph in $\mathbb{R}^d$ via eigenvectors $\mathbf{v}_2 \cdots \mathbf{v}_{d+1}$; eigenvalue $\lambda_1$is stochastic and the corresponding eigenvector $\mathbf{v}_1$ is uniform; this is an important property of the EVD solution because it isolates the problem's unwanted translational degree of freedom in a single eigenvector, leaving the remaining eigenvectors unpolluted.

Many embedding algorithms can be derived from this analysis, including the Fiedler vector [Fiedler, 1975], locally linear embeddings (LLE) [Roweis and Saul, 2000], and Laplacian eigenmaps [Belkin and Niyogi, 2002]. For example, direct solution of equation 1 gives the Laplacian eigenmap; as

a historical note, the symmetrized formulation was proposed by Fiedler in the 1970s and has been used for heuristic graph layout since the 1980s [Mohar, 1991].

2 Transformational embeddings

Now consider a more general problem: We are given some information about the vertices in a matrix $\mathbf{Z} \doteq [\mathbf{z}_1, \cdots, \mathbf{z}_N] \in \mathbb{R}^{d\times N}$, whose columns are generated by applying a vector-valued function $z(\cdot) \to \mathbf{z} \in \mathbb{R}^d$ to each vertex of the graph. We seek a linear operator which transforms $\mathbf{Z}$ to the optimal graph embedding: $G(\mathbf{Z}) \to \mathbf{Y}$. We will call this the "transformational embedding," to distinguish it from the "direct embedding" discussed above.

A natural candidate for the algebraic statement of the transformational embedding problem is the generalized EVD

$$(\mathbf{ZWZ}^\top)\mathbf{V} = (\mathbf{ZDZ}^\top)\mathbf{V}\Lambda, \tag{8}$$

because setting $\mathbf{Y} = \mathbf{V}^\top\mathbf{Z}$ makes this equivalent to the original direct embedding problem. Again, there is an equivalent symmetric eigenproblem: Make Cholesky[1] decomposition $\mathbf{R}^\top\mathbf{R} \leftarrow \mathbf{ZDZ}^\top$ into upper-triangular $\mathbf{R} \in \mathbb{R}^{d\times d}$ and let

$$\mathbf{B} \doteq (\mathbf{R}^{-\top}\mathbf{ZWZ}^\top\mathbf{R}^{-1}) \in \mathbb{R}^{d\times d}. \tag{9}$$

Then

$$\mathbf{BV}' = \mathbf{V}'\Lambda \tag{10}$$

with

$$\mathbf{V}' \doteq \mathbf{RV}, \quad \mathbf{V} = \mathbf{R}^{-1}\mathbf{V}'. \tag{11}$$

This gives an embedding $[\mathbf{v}_2, \mathbf{v}_3, \cdots]^\top\mathbf{Z}$, and a computational advantage: If $\mathbf{Z} \in \mathbb{R}^{d\times N}$ is a short matrix ($d \ll N$), the original $N \times N$ eigenproblem can be reduced to a very small $d \times d$ problem, and the matrix multiplications also scale as $O(d^2N)$ rather than $O(N^3)$, due to the sparsity of $\mathbf{W}$ and $\mathbf{D}$.

2.1 Correcting problematic eigenstructure

It is generally the case that $\mathbf{Y}^\top \notin \text{range}(\mathbf{Z}^\top)$—there is no linear combination of the rows of $\mathbf{Z}$ giving $\mathbf{Y}$, so the desired linear mapping $G(\mathbf{Z}) \to \mathbf{Y}$ does not exist. Equations 8–11 give the *optimal least-squares approximation* $G(\mathbf{Z}) = \mathbf{V}^\top\mathbf{Z} \approx \mathbf{Y}$. This approximation can have a serious flaw: If $\mathbf{1} \notin \text{range}(\mathbf{Z}^\top)$ then the first eigenvector $\mathbf{v}_1$ is *not* uniform; it cannot be discarded as the unwanted translational DOF. Worse, all the other eigenvectors will be variously contaminated by the unwanted DOF, resulting in an embedding polluted with artifacts. For this reason, we call direct solution of equation 8 a *raw* approximation.

Our options for remedy are limited to those that modify the row-space of $\mathbf{Z}$ to reintroduce the uniform eigenvector. For reasons that will become obvious below, we will restrict ourselves to operations that can be applied to any column of $\mathbf{Z}$ without knowing any other column.

The simplest such operation is to append a uniform row to $\mathbf{Z}$, so that $\mathbf{z}_i \to [\mathbf{z}_i^\top, 1]^\top$. This makes the relation between $\mathbf{Z}$ and $\mathbf{Y}$ *affine* and guarantees that $\mathbf{v}_1^\top\mathbf{Z}$ is uniform, but it can also force the eigenvectors to model additional variance that is not part of the problem.

Working backward from the desiderata that the leading column of $\mathbf{V}^\top\mathbf{Z}$ should be uniform, let $\mathbf{K} \doteq \text{diag}(\mathbf{v}_1^\top\mathbf{Z})^{-1}$ such that $\mathbf{ZK}$ is a modified representation of the vertices with values of $z(\cdot)$ *reweighted* on a per-vertex basis: $\mathbf{z}_i \to \mathbf{z}_i/(\mathbf{v}_1^\top\mathbf{z}_i)$. Clearly $(\mathbf{V}^\top\mathbf{ZK})^\top$ has a uniform first column, since each row is divided by its first element.

It follows immediately that the related eigenproblem

$$(\mathbf{ZKWKZ}^\top)\mathbf{V}'' = (\mathbf{ZKDKZ}^\top)\mathbf{V}''\Lambda'' \tag{12}$$

is *stochastic*, and $\mathbf{Y}'' \doteq [\mathbf{v}_2'', \mathbf{v}_3'', \cdots]^\top\mathbf{ZK}$ is an embedding with the unwanted translational degree of freedom totally removed. Note that the raw and stochastic approximations are orthogonal (under metric $\mathbf{D}$): $\mathbf{Y}''\mathbf{DY}''^\top$ is a diagonal matrix; the other methods are not.

It should be noted that—when scaled to have equal norm trace$(\mathbf{YDY}^\top)$— none of these approximations has uniformly superior distortion scores; but in Monte Carlo trials with random graphs, we find a clear ordering from lowest to highest distortion: *reweighted*, *affine*, *stochastic*, *raw* (see figure 1).

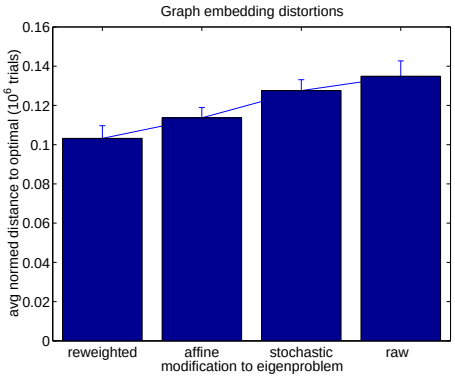

Figure 1: Comparison of methods for modifying the row-space of $\mathbf{Z}$. The graph shows distortion from the optimal embedding, averaged over 10^6 trials with 50-node matrices having random edge weights and random $\mathbf{Z} \in \mathbb{R}^{4\times 50}$.

The raw approximation is suboptimal because information about the d-dimensional embedding is spread over $d+1$ eigenvectors, no subset of which is optimal. The stochastic approximation is also suboptimal—it optimizes a different measure implied by equation 12. In practice, when computing embeddings of graphs whose embedding structure is known *a priori*, we find that the reweighted and stochastic approximations give results that are clearly very similar, and superior to the other approximations.

The need for *any* such correction stems from the fact that—the literatures of spectral graph theory and NLDR notwithstanding—equation 1 is *not* a completely correct statement of the embedding problem. We will show in a forthcoming paper that, as a statement of the embedding problem, equation 1 is both algebraically underconstrained and numerically ill-conditioned. In particular, point #2 is

[1] Any gram-like factorization will work. For example, given EVD $\mathbf{A\Omega A}^\top \leftarrow \mathbf{ZDZ}^\top$, $\mathbf{R} = \Omega^{1/2}\mathbf{A}^\top$. The Cholesky is especially attractive for its numerical stability, sparsity, and easy invertibility.

not strictly true: *The stochastic eigenvalue is not always paired to a uniform eigenvector*. This leads to pathologies that can ruin the embedding, whether obtained from the basic or derived formulations. NLDR algorithms that can be derived from equation 1 (e.g., [Roweis and Saul, 2000; Belkin and Niyogi, 2002; Teh and Roweis, 2003]) do not remediate the problem.

A forthcoming paper makes a full analysis of these issues, identifies the correct problem statements for both equations 1 & 8, and gives closed-form optimal solutions to both problems. The approximation methods discussed in this section are still useful in that they are faster and give reasonably high-quality embeddings. For the NLDR method and datasets considered below, the result of the reweighted approximation is almost numerically indistinguishable from the optimal embedding, and requires substantially less calculation. The reweighting method can also be justified as a Padé approximation of the optimal solution.

3 Nonlinear dimensionality reduction

Let $\mathbf{X} \doteq [\mathbf{x}_1, \cdots, \mathbf{x}_N] \in \mathbb{R}^D$ be a set of points sampled from a low-dimensional manifold embedded in a high-dimensional ambient space. A reduced-dimension *embedding* $\mathbf{Y} \doteq [\mathbf{y}_1, \cdots \mathbf{y}_N] \in \mathbb{R}^d$ with $d < D \leq N$ is a set of low-dimensional points with the same local neighborhood structure. We desire instead a *mapping* $G : \mathbb{R}^D \to \mathbb{R}^d$, which will generalize the correspondence to the whole continuum, with reasonable interpolation and extrapolation to be expected in the neighborhood of the data. Spectral methods for NLDR typically require solution of many and/or very large eigenvalue or generalized eigenvalue problems [Kruskal and Wish, 1978; Kambhatla and Leen, 1997; Tenenbaum *et al.*, 2000; Roweis and Saul, 2000; Belkin and Niyogi, 2002], and with the exception of [Teh and Roweis, 2003; Brand, 2003], offer embeddings of points rather than mappings between spaces.

Here we show how a leverage the transformational embedding of section 2 into a continuous NLDR algorithm, specifically a kernel-based mixture of affine maps from the ambient space to the target space. To do so, we must show how the edge weight matrix $\mathbf{W}$ and vertex matrix $\mathbf{Z}$ are specified. Let $W_{ij} \doteq f(\mathbf{x}_i, \mathbf{x}_j)$ iff $\mathbf{x}_i$ and $\mathbf{x}_j$ satisfy some locality criterion, e.g., $\|\mathbf{x}_i - \mathbf{x}_j\| < \varepsilon$, otherwise $W_{ij} = 0$. As stated above, an embedding $\mathbf{Y}$ of $\mathbf{X}$ should satisfy

$$\mathbf{Y} = \arg\min_{\mathbf{Y}} \textstyle\sum_{i \neq j} \|\mathbf{y}_i - \mathbf{y}_j\|^2 W_{ij} \tag{13}$$

where larger W_{ij} penalize large distances between $\mathbf{y}_i$ and $\mathbf{y}_j$.

How should W_{ij} be computed? f is a measure of similarity: The graph-theoretic literature usually takes $f(\cdot,\cdot) = 1$, while NLDR methods typically take $f(\mathbf{x}_i, \mathbf{x}_j) \propto \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2)$ to be a Gaussian kernel, on analogy to heat diffusion models [Belkin and Niyogi, 2002]. The uninformative setting $W_{ij} = 1$ is only usable when there is a very large number of points (and edges), so that connectivity information alone suffices to determine metric properties of the embedding. The Gaussian setting has a complementary weakness: It can be very sensitive to small variations in distance to neighbors (that may be introduced by the curvature of the data manifold or measurement noise in the ambient space). f should be monotonically decreasing, relatively insensitive to noise ($\mathrm{d}f$ should be small), and it should lead to exact reconstructions of data sampled from manifolds that are already flat. Straightforward calculus shows that equation 13 has the desired minimum when $f(\mathbf{x}_i, \mathbf{x}_j) \propto \|\mathbf{x}_i - \mathbf{x}_j\|^{-1}$, or more generally, the multiplicative inverse of whatever distance measure is appropriate in the ambient space[2]. (By contrast, the LLE weightings are not correlated with distances.) To make the problem scale invariant, we scale $\mathbf{W}$ such that its largest nonzero off-diagonal value is 1 (consequently $\mathrm{d}f \leq 1$ everywhere f is computed).

Let us now situate some Gaussian kernels $p_k(\mathbf{x}) \doteq \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$ on the manifold. In this paper, we will take a random subset of data points as kernel centers, and set all $\Sigma_k = \sigma^2\mathbf{I}$; these kernels are radial basis functions. Let vector

$$\mathbf{z}_{ik} \doteq \begin{bmatrix} \mathbf{K}_i(\mathbf{x}_i - \mu_k) \\ 1 \end{bmatrix} \frac{p_k(\mathbf{x}_i)}{\sum_k p_k(\mathbf{x}_i)} \tag{14}$$

be the kth local homogeneous coordinate of $\mathbf{x}_i$ scaled by the posterior of the kth kernel. $\mathbf{K}_i$ is an optional local dimensionality-reducing linear projection. Let representation vector

$$z(\mathbf{x}_i) \doteq [\Downarrow_k \mathbf{z}_{ik}] = [\mathbf{z}_{i1}^\top, \cdots, \mathbf{z}_{iK}^\top]^\top \tag{15}$$

be the vertical concatenation of all such local coordinate vectors. Collect all such column vectors into a *basis* matrix $\mathbf{Z} \doteq [z(\mathbf{x}_1), \cdots, z(\mathbf{x}_N)]$.

To summarize thus far, our goal now is to find a linear transform $\mathbf{y}_i \doteq G(z(\mathbf{x}_i))$ of the basis (kernel-weighted coordinates) that is maximally consistent with our local distance constraints, specifically

$$\mathbf{Y} = [\mathbf{y}_1, \cdots \mathbf{y}_N] = \arg\min_{\mathbf{Y}} \sum_{ij} \frac{\|\mathbf{y}_i - \mathbf{y}_j\|^2}{\|\mathbf{x}_i - \mathbf{x}_j\|} \tag{16}$$

This is isomorphic to the graph embeddings of section 2; the methods developed there apply directly to $\mathbf{W}$ and $\mathbf{Z}$. The continuous mapping from ambient to embedding space immediatly follows from the continuity and smoothness of $z(\cdot)$:

$$G(\mathbf{x}) = \mathbf{G}z(\mathbf{x}),$$

where the EVD determines the transformation $\mathbf{G} \doteq [\mathbf{v}_2, \cdots, \mathbf{v}_{d+1}]^\top$ of the continuous kernel representation defined over the entire ambient space:

$$z(\mathbf{x}) \doteq \left[\Downarrow_k \begin{bmatrix} \mathbf{K}_i(\mathbf{x} - \mu_k) \\ 1 \end{bmatrix} \frac{p_k(\mathbf{x})}{\sum_k p_k(\mathbf{x})}\right]. \tag{17}$$

[2] Proof: Consider three points $\{x_1 = 0, x_2 = \lambda, x_3 = 1\}$ on a 1D manifold. What similarity measure $W_{ab} = f(\|x_a - x_b\|)$ causes the distortion $(y_2 - y_1)^2 W_{12} + (y_2 - y_3)^2 W_{23}$ to have a global minimum at $y_2 = \lambda$? Without loss of generality, we fix the global location and scale of the embedding by fixing the endpoints: $\{y_1 = 0, y_3 = 1\}$. Solving for the unique zero of the distortion's first derivative, we obtain the optimum at $y_2 = W_{23}/(W_{12} + W_{23})$. Since this is a harmonic relation, the unique continuous satisficing measure is $W_{ab} = (\|x_a - x_b\|)^{-1}$. This sets $W_{12} = 1/\lambda$ and $W_{23} = 1/(1-\lambda)$; some simple algebra confirms that indeed $y_2 = \lambda$ at the optimum. The induction to multiple points in multiple dimensions is direct.

As a matter of numerical prudence, we recommend using the reweighted approximation:

$$G(\mathbf{x}) = \frac{\mathbf{G}z(\mathbf{x})}{\mathbf{v}_1^\top z(\mathbf{x})}. \tag{18}$$

At first blush, it would seem that reweighting should not be necessary: By construction, $\mathbf{1} \in \text{range}(\mathbf{Z}^\top)$, thus $\mathbf{v}_1^\top z(\mathbf{x})$—and the denominator—should be uniform at the datapoints. However, as mentioned above, even when the algebra predicts this structure, numerical eigensolvers may not find it.

To obtain an approximate inverse mapping, we map the means and covariances of each kernel $p_k(\cdot)$ into the target space to obtain kernels $p'_k(\mathbf{y}) \doteq \mathcal{N}(\mathbf{y}|\mu', \Sigma')$ there. Then, breaking $\mathbf{G} = [\mathbf{G}_1, \cdots, \mathbf{G}_K]$ into blocks corresponding to each kernel, take the Moore-Penrose pseudo-inverse of each, and set $\mathbf{G}^+ \doteq [\mathbf{G}_1^+, \cdots, \mathbf{G}_k^+]$. If using the reweighted map, the approximate inverse map is

$$G^+(\mathbf{y}) \approx \mathbf{G}^+ z^+(\mathbf{y}) \cdot (\mathbf{v}_1^\top \mathbf{G}^+ z^+(\mathbf{y})), \tag{19}$$

where $z^+(\mathbf{y}) \doteq \left[\Downarrow_k \begin{bmatrix} \mathbf{y} - \mu'_k \\ 1 \end{bmatrix} \frac{p'_k(\mathbf{y})}{\Sigma_k p'_k(\mathbf{y})} \right]$.

4 Illustrative example

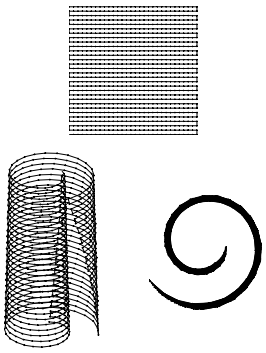

Figure 2: The swiss roll.

We will use a variant of the "swiss roll", a standard test manifold in the NLDR community, to illustrate the arguments and methods developed in this paper. We sampled a twisted version of the manifold regularly on a 30×30 grid and added a small amount of Gaussian isotropic noise. Figure 2 shows the ideal $\mathbb{R}^2$ parameterization and two views of the ambient $\mathbb{R}^3$ embedding. Points are shown joined into a line to aid visual interpretation of the embeddings. All experiments in this section use a $\mathbf{W}$ matrix that was generated using the 12 nearest neighbors to each point and the inverse distance function.

The Laplacian eigenmap embedding (figure 3) shows the embedding specified by the $\mathbf{W}$ matrix. Note that it exhibits some folding at the corners and top and bottom edges, due partly to problems with the uniform eigenvector and exacerbated by the fact that spectral embeddings tend to compress near the boundaries. The Laplacian eigenmap requires solution of a large 900×900 eigenproblem, and offers no mapping off the points. Kernel eigenmaps will be approximations to this embedding.

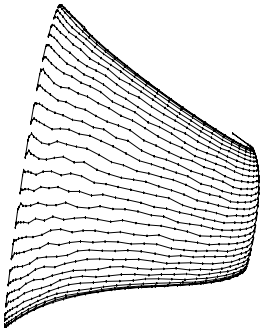

Figure 3: Laplacian eigenmap embedding.

We now show some kernel eigenmaps computed using the transformational embedding of section 2. All embedding methods are given the same inputs.

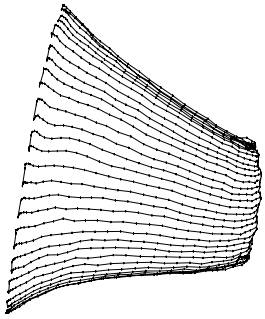

Figure 4: Kernel eigenmap embedding, raw result.

Figure 4 shows a raw kernel eigenmap embedding computed using a basis ($\mathbf{Z}$ matrix) created from 64 Gaussian unit-σ kernels placed on random points. This required solving a much more manageable 256×256 eigenproblem. 100 trials were performed with different sets of randomly placed kernels. In all trials, the reweighted and stochastic maps gave the best reconstructions, while the raw and affine maps exhibited substantial folding at the edges and corners of the embedding.

Figure 5 shows a *reweighted* kernel eigenmap computed from the same $\mathbf{W}$ and $\mathbf{Z}$ as figures 3 & 4. The result is smoother and actually exhibits *less* folding than the original Laplacian eigenmap.

The problem can be regularized by putting positive mass on the diagonal of $\mathbf{W}$ (e.g., $\mathbf{W} \to \mathbf{W} + \mathbf{I}$), thereby making the recovered kernel eigenmap more isometric (bottom figure 5). This regularization is appropriate when it is believed that all neighborhoods are roughly the same size.

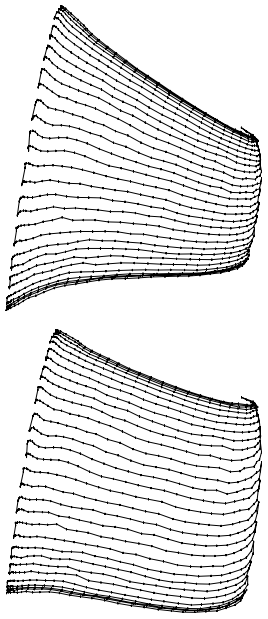

Figure 5: Kernel eigenmap embedding, reweighted and regularized results.

The recently proposed Locality Preserving Projection (LPP) [He and Niyogi, 2002], is essentially the raw approximation (direct solution of equation 8) with $W_{ij} = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/t}$ and $\mathbf{Z} = \mathbf{X}$, thereby giving a linear projection from the ambient space to the target space that best preserves local relationships.

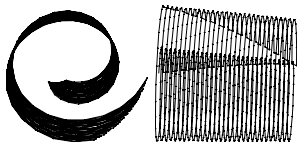

Figure 6: LPP embedding and our affine upgrade.

LPP is admirably simple, but it can be shown that the affine approximation from section 2 will always have less distortion. LPP can also suffer from loss of the uniform eigenvector. Figure 6 shows embeddings of the swiss roll produced by LPP and by an affine modification of it that is equivalent to our method with a trivial single uniform-

density kernel. Upgrading LPP to an affine projection captures more of the data's structure. Even so, there is no affine "view" of this manifold that avoids folding.

5 Visualizing word usages

In statistical analyses of natural language, similar usage patterns for two words are taken to indicate that they have similar meanings or strongly related meanings. Latent semantic analysis (LSA) is a linear dimensionality reduction of a term-document co-occurence matrix. The principal components of this matrix give an embedding in which similarly used words are similarly located. Literally, co-location is a proxy for collocation (the propensity of words to be used together) and synonymy. We may expect that the kernel eigenmap offers a more powerful nonlinear analysis:

The NIPS12 corpus[3] features a matrix counting occurrences of 13000+ words in 1700+ documents. We modeled the first 1000 words and the last 200 documents in the matrix. This roughly corresponds to one year's papers, a reasonable "snapshot" of the ever-changing terminology of the field. We stemmed the words and combined counts for the same roots, then determined distance between two word roots as the cosines of the angles between their log-domain-transformed occurrence vectors ($x_{ij} \to \log_2(1+x_{ij})$). The $\mathbf{W}$ matrix was generated by adding an edge from each word to its 30 closest neighbors in cosine-space. The representation $\mathbf{Z}$ was made using 4 random words as kernel centers. Figure 7 discusses the resulting 2D embedding, in which technical terms are clearly grouped by field and many of the more common English words are tightly clustered by common semantics. The first two LSA dimensions (also shown in figure 7) of the same data are reveal significantly less semantic structure.

6 Discussion

The kernel eigenmap generates continuous nonlinear mapping functions for dimensionality reduction and manifold reconstruction. Suitable choices of kernels can reproduce the behavior of several other NLDR methods. One could put a kernel at every local group of points, perform local dimensionality reduction (e.g., a PCA) at each kernel, and thereby obtain from equations 8 and 17 an NLDR algorithm much like charting [Brand, 2003] or automatic alignment [Teh and Roweis, 2003]. Or, as in the demonstrations above, the kernel eigenmap can simultaneously determine the local dimensionality reductions and their global merger.

The kernel eigenmap typically substitutes a small dense EVD for the the large sparse EVD of graph embedding problems. In the sparse case, a specialized power method can compute the desired eigenvectors in significantly less than the $O(N^3)$ time required for a full EVD. In the kernel setting, similar efficiencies apply because both $\mathbf{W}$ and $\mathbf{Z}$ are typically sparse, allowing fast construction of the reduced EVD problem $\mathbf{ZWZ}^\top$; this too is amenable to fast power methods. Of course, the most important efficiency of the kernel method is its ability to embed new points quickly via the function $G(\mathbf{x})$—there is no need to compute a new global embedding or revise the EVD.

The reweighting scheme, although theoretically mooted by our subsequent discovery of a better problem formulation and closed-form solution, is still practically viable as a fast approximation for large problems, and as a post-conditioning step for unavoidable numerical error of any NLDR algorithm based on eigenvalue decompositions.

In this paper we have used random kernels. There are numerous avenues to discovering stronger methods by investigating placement and tuning of the kernels, stability of the embedding and its topological structure, and sample complexity. In short, all the issues that proved fertile ground for research in classification and regression can be studied anew in the context of estimating the geometry and topology of manifolds.

[3] Courtesy S. Roweis, available from the U. Toronto website.

References

[Belkin and Niyogi, 2002] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. Technical Report TR-2002-01, University of Chicago Computer Science, 2002.

[Brand, 2003] Matthew Brand. Charting a manifold. In *Proc. NIPS-15*, 2003.

[Chung, 1997] Fan R.K. Chung. *Spectral graph theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, 1997.

[Fiedler, 1975] Miroslav Fiedler. A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. *Czech. Math. Journal*, 25:619–633, 1975.

[He and Niyogi, 2002] Xiafei He and Partha Niyogi. Locality preserving projections. Technical Report TR-2002-09, University of Chicago Computer Science, October 2002.

[Kambhatla and Leen, 1997] N. Kambhatla and Todd Leen. Dimensionality reduction by local principal component analysis. *Neural Computation*, 9, 1997.

[Kruskal and Wish, 1978] J. B. Kruskal and M. Wish. *Multidimensional Scaling*. Sage Publications, Beverly Hills, CA, 1978.

[Mohar, 1991] B. Mohar. The laplacian spectrum of graphs. In Y. Alavi, editor, *Graph Theory, Combinatorics and Applications*, pages 871–898. J. Wiley, New York, 1991.

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

[Teh and Roweis, 2003] Yee Whye Teh and Sam T. Roweis. Automatic alignment of hidden representations. In *Proc. NIPS-15*, 2003.

[Tenenbaum *et al.*, 2000] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, December 22 2000.

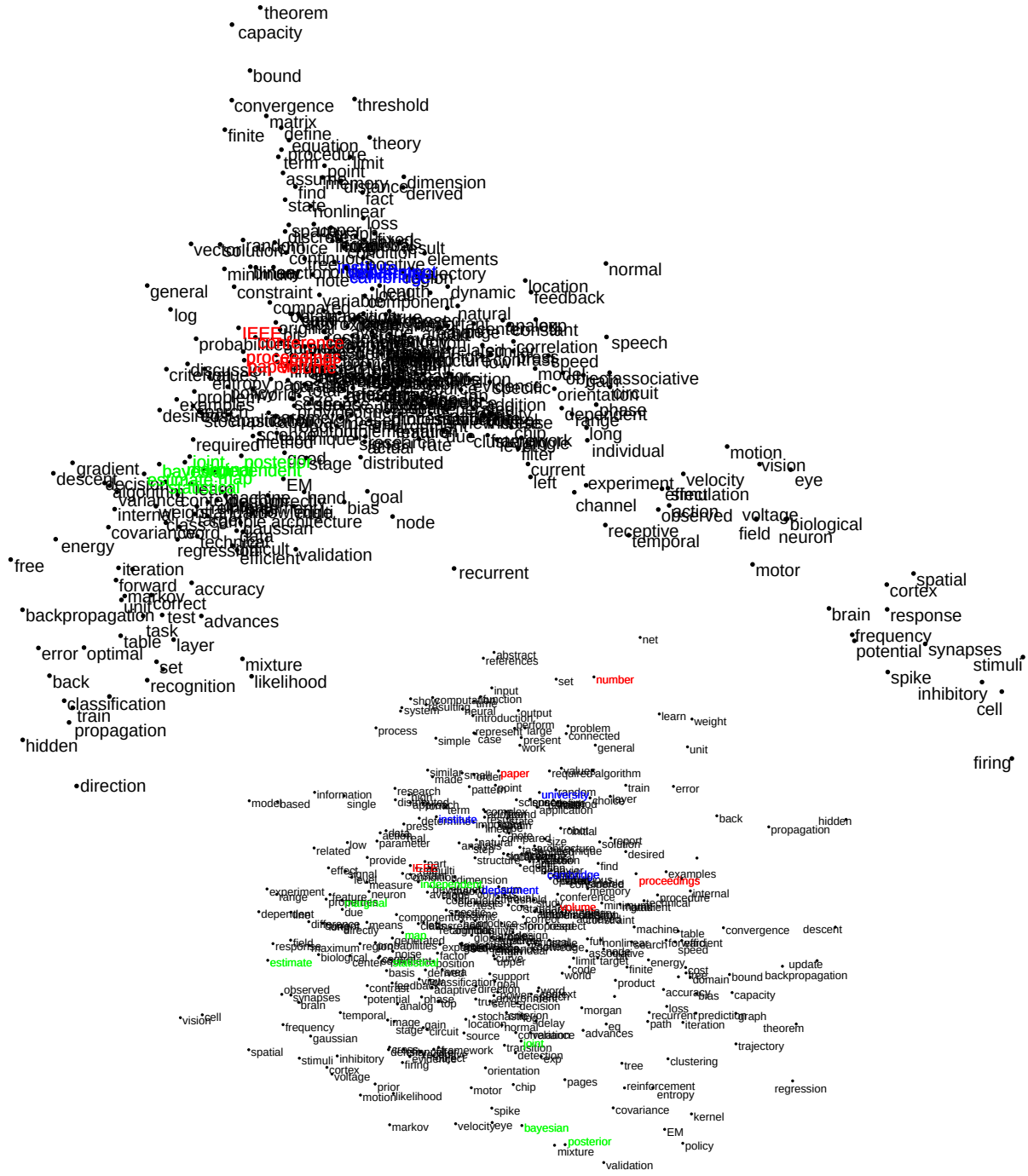

Figure 7: ABOVE: A 2D kernel eigenmap of word usages in recent NIPS papers. To improve legibility we show just a subset of the data; some labels have been shifted slightly to reduce overlap. Word roots are shown in their first occurring unstemmed variant. The three lobes of the distribution roughly correspond to favored terminology in the submission areas of Algorithms & Architectures (left), Neuroscience (right), and Theory (top). Words with broader usage are more tightly distributed in the center (presumably because they are more likely to co-occur in general discourse), with several clusters of words having strongly related meanings. Three of these clusters have been colored: red for publishing terms (IEEE, conference, number, paper, proceedings, volume), green for probability terms (bayesian, estimate, independent, map, marginal, posterior, joint, statistical), and blue for locations (cambridge, department, institute, university). BELOW, SMALLER: A linear embedding obtained from a latent semantic analysis of the same data. Though collocated words are often co-located, when compared with the kernel eigenmap result, semantic structures are far less obvious.