Iterative Embedding with Robust Correction using Feedback of Error Observed

Praneeth Vepakomma
Dept. of Statistics, Rutgers University, NJ, USA
Dept. of Electrical & Computer Engineering
Florida International University, FL, USA
Motorola Solutions, USA

Ahmed Elgammal
Dept. of Computer Science
Rutgers University, NJ, USA

Abstract

Nonlinear dimensionality reduction techniques of today are highly sensitive to outliers. Almost all of them are spectral methods and differ from each other over their treatment of the notion of neighborhood similarities computed amongst the high-dimensional input data points. These techniques aim to preserve the notion of this similarity structure in the low-dimensional output. The presence of unwanted outliers in the data directly influences the preservation of these neighborhood similarities amongst the majority of the non-outlier data, as these points ocuring in majority need to simultaneously satisfy their neighborhood similarities they form with the outliers while also satisfying the similarity structure they form with the non-outlier data. This issue disrupts the intrinsic structure of the manifold on which the majority of the non-outlier data lies when preserved via a homeomorphism on a low-dimensional manifold. In this paper we come up with an iterative algorithm that analytically solves for a non-linear embedding with monotonic improvements after each iteration. As an application of this iterative manifold learning algorithm, we come up with a framework that decomposes the pair-wise error observed between all pairs of points and update the neighborhood similarity matrix dynamically to downplay the effect of the outliers, over the majority of the non-outlier data being embedded into a lower dimension.

1 Introduction

Nonlinear dimensionality reduction (NLDR) methods like Laplacian Eigenmaps (1), Locally Linear Embedding (2), Hessian Eigenmaps (3), and Local Tangent Space Alignment (4) try to preserve the local geometry of high dimensional data in a low dimensional space. NLDR methods aim to find a homeomorphic mapping and assume that a 'representation' of the local geometry of high-dimensional data can be preserved on a smooth manifold of much lower dimension, also referred to as its intrinsic dimension (6), (7), (8). The problem of finding such a mapping is also referred to as -Manifold Learning. In the presence of outliers the information required to find a homeomorphic mapping is corrupted and nonlinear dimensionality reduction methods of today fail to completely recover the manifold of interest.

In this paper, we propose an iterative method for manifold learning and use it to adaptively downweight the outliers based on the pair-wise error produced at any given iteration. The effect of outliers is reduced by simultaneously updating the priori local neighborhood information that needs to be preserved after the embedding. This is done using multiplicative updates derived from the pair-wise error produced between all pairs of points after any given iteration. Our iterative embedding algorithm guarantees monotonic improvement after each iteration as it is based on majorization-minimization, an optimization framework that guarantees monotonic convergence. This leads to every iteration of our algorithm doing better than the previous until it converges to smaller and smaller improvements after many iterations. The problem of manifold learning in the presence of noise or missing data was studied in (9), (10). The focus in this paper is instead, over the presence of outliers that do not have the same low-dimensional representation as of the data of interest, and often get falsely projected over the smoothened manifold with a lower degree of freedom.

Appearing in Proceedings of the 4^{th} Workshop on Machine Learning for Interactive Systems (MLIS) 2015, Lille, France. JMLR: W&CP volume 43 Copyright 2015 by the authors.

2 Interleaving Iterative Embedding & Robust Reweighting of Similarity Structure

The problem of nonlinear embedding in the presence of outliers around the high-dimensional data can be tackled at two levels. The first level is to deal with outlier detection in the high-dimensional space before the embedding. This is non-trivial and non-obvious, because of the 'curse of dimensionality'. The other choice, would be during or after the embedding. In this paper we focus on dealing with outliers during the embedding by first presenting an algorithm for embedding iteratively. Our iterative algorithm is inspired by Laplacian Eigenmaps, a non-iterative technique that uses a Gaussian kernel $exp(-\left\|Y_{i.} - Y_{j.}\right\|_2^2/\sigma)$ to generate weights W_{ij} from a high dimensional data matrix $Y_{n\times k}$, where $Y_{i.}, Y_{j.}$ are data points in $\Re^k$ and denote the rows i, j of Y respectively. σ is a tuning parameter that establishes the notion of the extent of neighborhood. In Laplacian Eigenmaps a low-dimensional embedding $X_{n\times p}$, in $\Re^p$ with $p < k$ is obtained by minimizing the following loss function:

$$\sum_{i,j} W_{ij} d_{ij}^2(X) \tag{1}$$

over X, where $d_{ij}^2(X)$ is the squared Euclidean distance between row $X_{i.}, X_{j.}$ The solution is subject to an orthornormal constraint over $XG^{1/2}$ that depends on the diagonal matrix G, where $G_{ii} = \sum_j w_{ij}$ and is given by $X^T G X = I$ and thereby prevents a degenerate solution over X. In this paper we consider the case where W is a matrix of weights computed over a Y that is plagued by outliers. Our approach relies on the pair-wise error matrix E, obtained at any given solution $\hat{X}$ with entries $E_{ij} = W_{ij} d_{ij}^2(\hat{X})$. The pair-wise residuals E_{ij} can be decomposed into decoupled pointwise indices of the form c_i, c_j such that, $\sum_{i,j} \gamma(E_{ij}, (c_i, c_j))$ is minimum, based on the model $\gamma(.)$ that we would like to build over our error. We refer to c_i, c_j, as point-wise indices, in the rest of the paper. These indices can be used at this stage to inturn update the weight matrix W.

It would be of practical use to have an iterative update for non-linear embedding, where the pairwise error terms can be collected during the process of embedding, and decomposed into well regularized point-wise indices, which would in turn be used to dynamically update the weight matrix during the embedding. We aim for $c_i^{\prime}s$ being restricted to $\Re^+$ $\forall i$. Finally, we perform a regularized M-Estimation to estimate the point-wise indices in this framework.This gives us the following updates at iteration, t in its basic form:

$$\underset{X\neq 0}{\arg\min} \sum_{i,j} W_{ij}^t d_{ij}(X^t)^2 \tag{2}$$

$$\underset{\overrightarrow{c}}{\arg\min} \sum_{i,j} \gamma(E_{ij}^t, c_i^t, c_j^t) \tag{3}$$

where $E_{ij}^t = \sum_{i,j} W_{ij}^t d_{ij}(\hat{X}^t)^2$ with $\hat{X}^t$ being the minimizer of eqn 2 at any given iteration under a constraint that X is not a matrix of all zeros. In our scheme, W_{ij}^{t+1} is updated using a functional of the point wise indices as $W_{ij}^{t+1} = W_{ij}^t \gamma(c_i^t, c_j^t)$. We minimize (2) iteratively, and learn a new weight matrix based on the error at every iteration using 3.The next two sections build over these updates shown above, taking the issues of regularization, convergence and robustness into consideration.

3 Unified Iterative Framework for Nonlinear Dimensionality Reduction

In this section, we propose majorization-minimization based iterative updates and a linear constraint for nonlinear dimensionality reduction over the loss function given as: $\underset{X}{\arg\min} \sum_{i,j} W_{ij} d_{ij}^2(X)$. This can be represented as a trace optimization problem as follows:

$$\begin{aligned} &\underset{X}{\arg\min}\ \Theta(X) = Tr[X^T L X] \\ &L = D - W; D_{ii} = -\sum_j w_{ij} \end{aligned} \tag{4}$$

L is also known as the graph laplacian. We build a majorization function (13), (14) over the above model, based on the fact that $[2Diag[L] - L]$ is diagonally dominant. This leads, to the following inequality for any matrix $M_{n\times p}$ given by: $(X - M)^T [2Diag[L] - L](X - M) \succeq 0$ and this inequality was used by Trosset, in [4], in a different context; to have a faster algorithm, as a substitution to the Guttman majorization based MDS. We get the following majorization inequality over our objective function in (4), by separating it from this inequality using

$$g(X, M) = Tr[X^T 2Diag(L)X] - 2Tr[X^T(2Diag(L) - L)M]$$

as

$$Tr(X^T L X) + g(X, M) \tag{5}$$

which is quadratic in X. Hence, we achieve the following bound over our objective function:

$$\begin{aligned} Tr(X^T L X) + f(M) &\leq g(X, M), \forall\, X \neq M \\ &= g(X, X),\ X = M \end{aligned}$$

that satisfies the supporting point requirement, and hence g(.) touches the objective function at the current iterate and the following majorization-minimization iteration holds true:

$$X^{t+1} = \underset{X}{\arg\min}\, g(X, M^t) \text{ and } M^{t+1} = X^t$$

Also, L_{ij} can be replaced by$L_{ij}.\gamma(c_i, c_j)$ without loss of any generality. It is important to note that these inequalities occur amongst the presence of additive terms that are independent of X unlike a typical majorization-minimization framework and hence, it is a relaxation. We now propose a linear constraint for nonlinear dimensionality reduction over the quadratic loss function proposed in (5). Our constraint prevents degenerate solutions, where the rows(or columns) of X coincide thereby preventing $d_{ij}(X)$ from going to zero. Its linearity, makes it easier to practically enforce it due to the quadratic nature of the loss function.

Row Unique Matrix: A matrix M is row-unique, if all the rows in the matrix are distinct.

Proposition: For any row-unique matrix $M_{n\times p}$, and for any given Laplacian matrix $L_{n\times n}$, if $Tr(X^TLM) \neq 0$, then there exist at least two rows in $X_{n\times p}$, that are distinct. $Tr(X^TLM) = \sum_{i<j} w_{ij}\phi_{ij}(X, M)$ where, $\phi_{ij}(X, M) = \sum_{a=1}^{p}(x_{ia} - x_{ja})(m_{ia} - m_{ja})$. Hence, for a row unique M, there exists at least two rows in X, such that $x_{i.} \neq x_{j.}$ in order to satisfy the inequality on $Tr(X^TLM)$. Note that, $\phi_{ij}(X, X) = d_{ij}^2(X)$.

We define our constraint in its basic form for nonlinear dimensionality reduction as follows: $Tr(X^TSM) = \nu$ where $\nu > 0$ is a user-defined constant and $S = n^{-1}I - ee^T$ is the graph laplacian, with all the weights being one. As a result of $g(.)$ being a quadratic majorizer we have

$$\lim_{t\to\infty} \|X_{t+1} - M_t\| \to 0$$

as a result of which, we have the following over $\phi(.)$ in our linear constraint

$$\lim_{t\to\infty} \phi_{ij}(X_{t+1}, M_t) \to d_{ij}^2(X) \in \Re^+$$

and hence we require that ν be non-negative inorder to simultaneously achieve convergence and enforce regularization.The following is the total loss function, $T(.)$ obtained when the constraint is combined with our majorizing function $g(.)$, defined in (5) with λ being a positive multiplier over the constraint: $T(X, \lambda) = g(X, M) + \lambda\left[Tr(X^TSM) - \nu\right]$. We get the following update, by setting the gradient equal to zero: $X_{t+1} = M_t - (0.5)\left[Diag(L)\right]^{-1}LM_t - 0.25\lambda[Diag(L)]^{-1}SM_t$ and solving for the constraint, we get the follwing update, for the multiplier:

$$\lambda = \frac{4(Tr[M_t^TSM_t] - \nu) - 2Tr(M_t^TLDiag(L)^{-1}SM_t)}{Tr(M_t^TSDiag(L)^{-1}SM_t)} \tag{6}$$

$$M_t = X_{t+1} \tag{7}$$

Hence, these are updates that satisfy the following set of inequalities, $\Theta(X_t) \leq g(X_t, M_{t-1}) \leq g(X_{t-1}, X_{t-1}) \leq \Theta(X_{t-1})$ and with every iterate, doing better than the previous, it proves the convergence of our updates.

4 Robust Multiplicative Updates

In this section, we deal with the estimation of robust pointwise indices from the error obtained after every iteration of eqn.s (6), (7) inorder to reweight the weights at each iteration. We aim to downweight the effect of outliers during a nonlinear embedding and help retain local information, that is required to achieve a homeomorphic mapping of the topology of interest. We provide majorization minimization based updates, to perform a regularized M-estimation of these indices with a differentiable ψ type robust function. We minimize the following function, that is defined over the residual, e using a robust function $\rho(.)$ given by: $\sum_{i,j}\rho\left(e_{ij}c_i^kc_j^k\right)$; $e_{ij} = W_{ij}d_{ij}^2(X^k)$. The Geman Mcclure $\rho(.)$ function and its first derivative, which is the influence function $\psi(.)$ is given by: $\rho\,(x) = \frac{x^2}{(\sigma+x^2)}$; $\psi\,(x) = \frac{2x\sigma}{(\sigma+x^2)^2}$. (15) suggested a beautiful result for calculating a majorizer if $h(.)$ is an even, differentiable function such that the ratio $h'(x)/x$ is decreasing on $(0,\infty)$ and the sharpest quadratic majorizer is given by $\frac{h'(y)}{2y}(x^2 - y^2) + h(y)$. Our $\rho(.)$ function in (18) does not require an alternative construction for a majorizer as $\lim_{x\to\infty}\frac{\psi(x)}{x} = 0,\ \rho\,(x) = \rho\,(-x)$ and hence we have the following sharpest quadratic majorizer up to a constant: $\xi_1(c, z) = \sum_{i<j}\frac{e_{ij}^2\sigma c_i^2c_j^2}{(\sigma+z_i^2z_j^2)}$. We majorize $c_i^2c_j^2$ inorder to achieve independence of variables in $\overrightarrow{c}$ over the gradient as required for constraint qualification and hence, it also give us closed form updates instead of relying on a block relaxation framework that involves cyclic updates. We employ the following majorizer that is obtained through the arithmetic-geometric mean inequality, $c_i^2c_j^2 \;\leq\; z_i^2z_j^2\left[\frac{1}{2}\left(\frac{c_i}{z_i}\right)^4 + \frac{1}{2}\left(\frac{c_j}{z_j}\right)^4\right] \;=\; \beta(c, z);\ \forall c \neq z$ where $\beta(c, c) = c_i^2c_j^2$ and $\beta_{c,z}$ also provides us with implicit positivity constraints over$\overrightarrow{c}$ in a majorization setting where, $z_t = c_{t-1}$. Employing (22) over (21) we get the following majorizer $\xi_m(.)$ over the chosen robust function: $\xi_m(c, z) = \sum_{i<j}\frac{e_{ij}^2\sigma}{(\sigma+z_i^2z_j^2)^2}\; z_i^2z_j^2\left[\frac{1}{2}\left(\frac{c_i}{z_i}\right)^4 + \frac{1}{2}\left(\frac{c_j}{z_j}\right)^4\right]$. We require that the entries in $\overrightarrow{c}$ corresponding to outliers in the data be sparse with the rest of the indices being large and spread out. We use a combination of L_1 and L_2 norms with coefficients that control the tradeoff between the sparsity induced by the L_1 and the reguarization of large values with the easy to optimize L_2 norm. This framework was previously introduced to improve the performance of the lasso, and to encourage the grouping effect among the predictors in the regression setting. This gives us the following loss function with λ_1,λ_2 being the coefficients over the norms and ξ_m being our majorizer: $l\,(c, z) = \xi_m(c, z) + \lambda_1\left\|c\right\|_1 + \lambda_2\left\|c\right\|_2$. The contribution of λ_1 and λ_2 can be easily reparametrized using a single variable $\alpha = \frac{\lambda_2}{\lambda_1+\lambda_2}$ giving us the following problem: $\widehat{c} = \arg\min_{c}\; l(c, z)$ with the constraint using, $r \in \Re^+$

such that: $(1-\alpha)\left\|c\right\|_1 + \alpha\left\|c\right\|_2 \le r$. We majorize $\sum_{1=1}^{n}\sqrt{c_i^2}$ using a linear approximation of its taylor expansion to deal with the L_1 norm as shown below: $\xi_m(c,z) + \lambda_1\sum_{i=1}^{n}\frac{c_i^2+z_i^2}{2|z_i|} + \lambda_2\sum_{i=1}^{n}c_i^2 = \xi(c,z)$. This gives us the following quadratic equation which needs to be solved at every iteration: $c_i^2 = \frac{1}{k_i}\left(\frac{\lambda_1}{|z_i|} + 2\lambda_2 c_i\right)$ and for the model in (26) we have the following update to obtain c_i: $c_i^2 = \frac{\gamma p_i}{k_i}$; $p_i = \left[\frac{\alpha-1}{|z_i|} - 2\alpha\right]$ such that: and $\gamma = \frac{t}{(1-\alpha)/\sqrt{k_i}\sum_{i=1}^{n}\left(\sqrt{p_i}+\alpha p_i\right)}$.

5 Experiments

In this section, we present the results of our iterative algorithm presented in section 3, along with the results of these updates, when combined with the robust framework in section 4. We refer to our technique as 'Robust Nonlinear Embedding' or 'RNE' in this section. We initially tested RNE on this standard dataset, in the presence of outliers that were uniformly generated around the topology. Fig 1. shows the results of our expriment on this dataset under the presence of outliers. The first image in Fig 1. is the Torroidal Helix before adding outliers. The second image has 5% outliers added around it. The ideal recovery upon this embedding from a Homeomorphic perspective has to be a circular loop. The third image shows the result obtained by Laplacian Eigenmap where the result is severely distorted because of the outliers. The fourth image shows the result of our RNE, which is close to the ideal of being a circular loop. The fifth image in this series shows the monotonic convergence of the error of our iterative embedding algorithm when applied on the non-corrupted Torroidal Helix prior to even interleaving it with the robust outlier correction mechanism. Figure 1; shows that our proposed algorithm recovers the topology reasonably well, in comparison to Laplacian Eigenmaps in the presence of outliers. This experiment was actually first run on Laplacian Eigenmaps, with different neighborhood parameters, and the parameters that gave the best possible embedding were chosen, and the corresponding weight matrix was constructed. We then ran our proposed (RNE) algorithm, using the weight matrix constructed above. For a real-life data experiment we used the famous USPS Handwritten Digits standard dataset to measure the precision and recall of RNE over the Digit 1 corrupted with increasing levels of outliers generated by uniform sampling from the rest of the digits in this dataset. The measurements were made using the indices generated by RNE. Indices that were close to zero, were counted as points detected as outliers and indices with larger values were counted as points considered as inliers, and then these measurements were compared with respect to the ground truth. Table 1 shows the Precision/Recall measured using this construction over a repeated series of experiments with increasing levels of outliers. As the % of outliers increased from 10% to 40% the precision-recall have reduced from a precision of 98.99% and a recall of 98.5% to a precision of 87.45% and a recall of 84.5% respectively in our detection rate upon the completion of the entire iterative embedding.

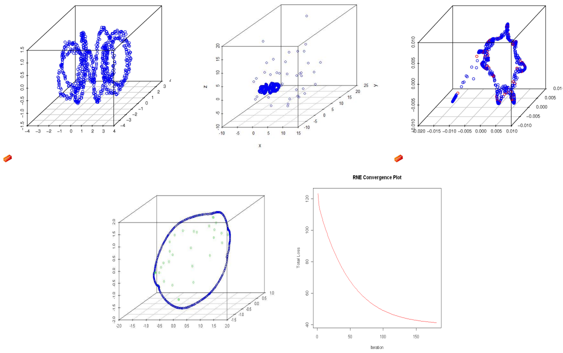

Figure 1: Noisy Toroidal Helix: 5% Outliers

Outlier %	Precision/Recall
10% Outliers	98.99 / 98.5
20% Outliers	98.98 / 98.0
30% Outliers	98.45 / 95.5
40% Outliers	87.45 / 84.5

Table 1: Precision/Recall with Varying Percentage of Outliers

From a visual perspective, the Fig.2 shows the comparison of the embeddings recovered by Laplacian Eigenmaps and RNE respectively. The first image shows the result of Laplacian Eigenmaps where the outliers have been placed relatively closer to the embedding of the 1's. Similarly, in some cases the outliers and inliers have got mixed up as well, as in by being placed in close proximity to each other. In comparison, in the second image, the 1's have densely amassed themselves on an arc like geometry and a vast majority of the outliers have got separated from this structure formed by 1's. Empirical evidence was collected to see the effect of the parameter ν in our constraint. We used data depth, an affine invariant, robust measure of scatter to find that the scatter increases with increasing ν to an extent, following which the change in scatter flattens out.

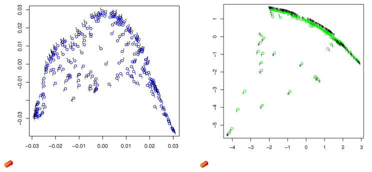

Figure 2: USPS Digit 1

References

[1] Mikhail Belkin, and Partha Niyogi. Laplacian Eigenmaps for dimensionality reduction and data representation. *Neural Computation*,15(6):1373–1396, 2003.

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

[3] David L. Donoho and Carrie Grimes Hessian Eigenmaps: Locally Linear Embedding Techniques for High-Dimensional Data. *Proceedings of the National Academy of Sciences of the United States of America*, 100(10):5591-5596, 2003.

[4] Zhenyue Zhang and Hongyuan Zha. Principal Manifolds and Nonlinear Dimension Reduction via Local Tangent Space Alignment. *SIAM Journal of Scientific Computing*, 26:313-338, 2002.

[5] Ingwer Borg, and Patrick J. F. Groenen. Modern multidimensional scaling: theory and applications. *Springer Series in Statistics*, 2005

[6] Y. Freund, S. Dasgupta, M. Kabra, and N. Verma. Learning the structure of manifolds using random projections. *Advances in Neural Information Processing Systems*, 2008

[7] Richard G. Baraniuk and , Michael B. Wakin. Random projections of smooth manifolds. *Foundations of Computational Mathematics*, 941-944, 2006.

[8] James Theiler. Statistical precision of dimension estimators. *Physical Review A*, 41(6):3038–3051, 1990.

[9] Samuel Gerber , Tolga Tasdizen, and Ross Whitaker Robust Non-linear Dimensionality Reduction using Successive 1-Dimensional Laplacian Eigenmaps. *In International Conference on Machine Learning*, 2007.

[10] Miguel A. Carreira-Perpi, and Zhengdong Lu. Manifold learning and missing data recovery through unsupervised regression. *International Conference on Data Mining*, 2011.

[11] Mukund Balasubramanian, and Eric L. Schwartz. The Isomap algorithm and topological stability. *SCIENCE*, 295:7, 2002.

[12] Matthew Brand. Charting a Manifold. *Advances in Neural Information Processing Systems*, 961–968, 2003.

[13] Bharath K. Sriperumbudur, and Gert R. G. Lanckriet. On the Convergence of the Concave-Convex Procedure. *Advances in Neural Information Processing systems*, 2009

[14] Kenneth Lange, David R. Hunter, and Ilsoon Yang. Optimization Transfer Using Surrogate Objective Functions.

[15] Jan de Leeuw and Kenneth Lange. Sharp Quadratic Majorization in One Dimension *Computational Statistics Data Analysis*,Volume 53, Issue 7, 2009

[16] E. Kokiopoulou, J. Chen, and Y. Saad. Trace optimization and eigenproblems in dimension reduction methods. *Numerical Linear Algebra with Applications*, 18(3):565–602, 2011

[17] John Aldo Lee, and Michel Verleysen. Quality assessment of nonlinear dimensionality reduction based on K-ary neighborhoods. *JMLR Proceedings, page*, Vol. 4, 21-35. 2008