Local Multidimensional Scaling for Nonlinear Dimension Reduction, Graph Layout and Proximity Analysis

Lisha Chen and Andreas Buja
University of Pennsylvania

This draft: July 26, 2006

Abstract

In recent years there has been a resurgence of interest in nonlinear dimension reduction methods. Among new proposals are so-called "Local Linear Embedding" (LLE) and "Isomap". Both use local neighborhood information to construct a global low-dimensional embedding of a hypothetical manifold near which the data fall.

In this paper we introduce a family of new nonlinear dimension reduction methods called "Local Multidimensional Scaling" or LMDS. Like LLE and Isomap, LMDS only uses local information from user-chosen neighborhoods, but it differs from them in that it uses ideas from the area of "graph layout". A common paradigm in graph layout is to achieve desirable drawings of graphs by minimizing energy functions that balance attractive forces between near points and repulsive forces between non-near points against each other. We approach the force paradigm by proposing a parametrized family of stress or energy functions inspired by Box-Cox power transformations. This family provides users with considerable flexibility for achieving desirable embeddings, and it comprises most energy functions proposed in the past.

Facing an embarrassment of riches of energy functions, however, one needs a method for selecting viable energy functions. We solve this problem by proposing a meta-criterion that measures how well the sets of K-nearest neighbors agree between the original high-dimensional space and the low-dimensional embedding space. This meta-criterion has intuitive appeal, and it performs well in creating faithful embeddings. The meta-criterion not only solves the problem of choice of energy function but also provides a measure for comparing embeddings generated by arbitrary dimension reduction methods, including principal components, LLE and Isomap. Lastly, the meta-criterion

can be used pointwise as a diagnostic tool for measuring the local adequacy of embeddings.

Our methods generally do a superior job in flattening the well-known Swiss roll and some of the other illustrative datasets used in the LLE and Isomap papers. Our work contributes to, and ties together, three areas: nonlinear dimension reduction, graph layout, and proximity analysis.

Key words and phrases: MDS, Local Linear Embedding, LLE, Isomap, Principal Components, PCA, Energy Functions, Force-Directed Layout, Cluster Analysis, Unsupervised Learning

1 INTRODUCTION

A well-chosen picture provides instant insight. This is why exploratory data analysis (EDA) and data visualization (DV) are such powerful data-analytic tools, and why they have been playing an increasing role in many areas of science. Since the human visual system only possesses the ability to see two- and three-dimensional shapes but many scientific data are intrinsically high-dimensional, visual understanding of data fundamentally involves dimension reduction, that is, the art of flattening data that fall near an intrinsically low dimensional manifold in high dimensional data space. High dimensionality of data is really one of two possible aspects of largeness of data: it simply means that the data have a large number of variables/attributes, as opposed to a large number of objects/cases. Such high-dimensional data have arisen naturally in recent times as one has moved from the analysis of single images or single signals to the analysis of whole databases of images and signals. Images and signals are then considered as objects/cases, and their pixels in different locations or amplitudes at different time points are then the variables. The correlations between nearby pixels or amplitudes make it plausible that the collections of images and signals are intrinsically low-dimensional, and this is where dimension reduction comes into its own.

In what follows we briefly describe common as well as newer dimension reduction methods, which will segue to our own proposals.

PCA The most well-known dimension reduction method is principal component analysis. Suppose we have multivariate data $\mathbf{x}_1, ..., \mathbf{x}_N \in \mathbb{R}^p$ and stack them as rows of an $N \times p$ matrix $\mathbf{X}$. Each column of $\mathbf{X}$ is called a feature/attribute/variable. PCA consists of finding linear combinations of the features that can capture most

variation of the data, that is, PCA emphasizes the dimensions that have large variances and ignores the dimensions that have small variances.

MDS Unlike PCA, multidimensional scaling (MDS) aims to preserve the pairwise distances between the data points. MDS was originally invented to study the proximity structure between pairs of objects when the data were not multivariate features of the objects, but similarities or dissimilarities (proximities) of pairs of objects. Given measurements of proximities, MDS maps the objects to low-dimensional points that best represent the proximities (similarities or dissimilarities). For the purpose of dimension reduction, the dissimilarities are the distances between the high-dimensional data points (p large), and, using only their pairwise distances, MDS maps them into low dimensional space ($d = 2$ or 3). There are two main approaches to MDS: classical scaling (Torgerson 1952) and distance scaling (Kruskal 1964a,b). Even though distance scaling has become the leading version of MDS, classical scaling is still frequently used because it can be solved with a simple eigendecomposition. In the situation where the distance matrix $\mathbf{D}$ contains Euclidean distances derived from an $n \times p$ matrix $\mathbf{X}$, classical scaling can be shown to be equivalent to principal component analysis and hence amounts to a linear dimension reduction method. Distance scaling, however, is a genuine nonlinear dimension reduction technique. It is computed by minimizing a so-called "stress function" that measures the discrepancy between the distances in high-dimensional space and those in the reduced dimensions. More details will be given in the following section.

Isomap This method makes the assumption that the data concentrate around a low dimensional manifold in high-dimensional data space. The (artificial) standard example is the well-known "Swiss Roll" (Figure 1), a two-dimensional manifold which is usually shown in three-dimensional space but could be embedded in arbitrarily high dimensions. Intuitively speaking, the goal of Isomap is to "flatten" such manifold data, thus eliminating the dimensions that are used up by curvature. If successful, a flattened manifold requires only a reduced number of dimensions equal to its intrinsic dimension. This, however, cannot be achieved with PCA and MDS: these methods try to preserve global structure and do not allow the kind of local deformation needed for successful flattening of a curved and intrinsically low-dimensional manifold. The interesting property of the "Swiss Roll" is that points far apart on the underlying manifold, as measured by their geodesic distances, may appear deceptively close in the high-dimensional input space, as measured by their straight-line or chordal distances.

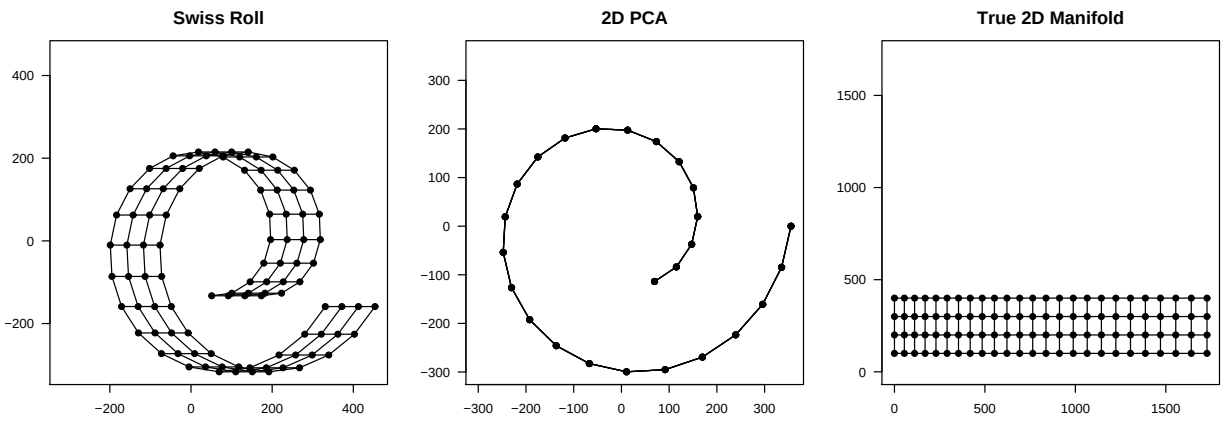

Figure 1: *Swiss Roll: The configuration from PCA (middle) does not recover the underlying 2D manifold (right).*

"Isometric feature mapping" or Isomap (Tenenbaum et al. 2000) tries to solve the problem of flattening a curved manifold by building on classical multidimensional scaling but aims to recover the intrinsic geometry of the data by preserving geodesic distance within the manifold. That is, *Isomap is classical scaling in which the distances have been replaced by estimates of intrinsic geodesic distance.* In practice, this is done as follows: for nearby points the geodesic distances are approximated by Euclidean distances; for distant points the geodesic distances are approximated by the length of the shortest path in a graph with edges connecting nearby points. Finally, Isomap applies classical scaling to the matrix of graph distances to obtain a representation of the data in low dimensions.

LLE Local linear embedding (Roweis and Saul, 2000) represents an entirely new idea. Like Isomap it uses localization, but unlike Isomap it does not attempt to match low-dimensional and high-dimensional distances. Instead, *LLE attempts to preserve local affine structure.* Assuming that each data point lies on a locally linear patch of a manifold, LLE characterizes local geometry by representing each data point as an approximate affine mixture of its neighbor points. LLE then achieves dimension reduction by constructing a point scatter in low dimensions that preserves as best as possible the affine mixture coefficients obtained in high-dimensional data space. — Saul and Roweis (2003, Sec. 5.1) also developed a version of LLE that takes local distances as inputs. The underlying idea is analogous to a localized version of classical scaling whereby inner products are reconstructed from squared Euclidean distances.

Local MDS, our new proposal that competes with Isomap and LLE, can be seen as a derivative of MDS on the one hand, and as method for graph layout on the other hand. As graph layout plays an important role in what follows, we introduce the customary definitions: A graph $G(V, E)$ is a pair consisting of a vertex set V and an edge set E. A graph is called "distance-weighted" if the edges are assigned real numbers that are interpreted as distances. An unweighted graph is the special case where each edge has the same weight, 1, say. Graphs are abstract structures that are used to model the idea of objects partially connected to each other. In practice, graphs are visualized as points connected by lines in what is called a "graph layout". Graph layout methods form an important specialty in the scientific visualization of relational data (Di Battista et al., 1999; Kaufmann and Wagner, 2001; Michailidis and de Leeuw, 2001). Though graph layout methods do not seem immediately related to dimension reduction due to the different type of data, there exist natural links:

- The idea of localization used by Isomap and LLE naturally generates a non-trivial graph consisting of those edges that correspond to local distances.
- Graph layout is similar to MDS in that most of its approaches rely on the minimization of an analog of stress functions, called "energy functions."
- In fact, MDS has been used early on as a graph layout method by defining distances between vertices in terms of shortest path lengths within the graph (Kruskal and Seery, 1980; Kamada and Kawai, 1989; Gansner et al., 2004). Therefore, Isomap is MDS-based graph layout using shortest path lengths computed from a distance-weighted graph of local edges.

Our approach to dimension reduction differs from Isomap in that it does not complete the graph of local distances; instead, it uses a physical paradigm common in graph layout, namely, the minimization of energy functions that decompose into "attractive forces" between local objects and "repulsive forces" between distant objects. [For general introductions to "force-directed" graph layout, see Di Battista et al. (1999, chap. 10), and Brandes (2001). Our critical reference, however, is Noack (2003).]

Localization has a history in MDS, albeit an unsuccessful one: Removing large distances from the stress function has probably been tried many times since Kruskal's seminal papers (1964a, 1964b). In distance scaling it is a rather obvious idea to remove those terms from the stress function that correspond to the large dissimilarities, in the hope that the small dissimilarities somehow add up to a globally meaningful configuration. This hope was dashed in a widely cited study by Graef and Spence (1979), who reported about simulations in which they removed the largest third and the smallest third of the dissimilarities from the stress function. While removal of the smallest third

was relatively benign, removal of the largest third had calamitous effects. Since then the received wisdom in MDS has been that the stability of MDS configurations stems from the large dissimilarities while the small dissimilarities are relatively unimportant. In this light, localized MDS does not appear to be a viable approach.

One way out of the problem is of course the imputation of large distances from local distances by replacing them with shortest-path lengths in a near-neighbor graph, as practiced by Isomap and much earlier by Kruskal and Seery (1980). Imputation with shortest-path lengths has problems, however: in our experience shortest paths in graphs zig-zag considerably in areas where the nearest neighbors are widely spread out. As a result, shortest-path lengths may accumulate a considerable degree of noise. In view of the fact that MDS relies mostly on large distances, Isomap may be driven largely by the noisy shortest-path extensions of the local distances, while the actual local distances themselves may play only a very minor direct role in the stress functions. A general tendency to noisiness is indeed visible in some embeddings generated by Isomap, especially compared to those generated by other nonlinear methods such as LLE or LMDS. Isomap embeddings tend to be fuzzier while LLE and LMDS embeddings tend to be crisper. The crispness of the latter may come at a cost: they may be systematically distorted in ways that Isomap avoids. The trade-off then may be one of classical bias versus variance: Isomap may suffer from more variance, LLE and LMDS from more bias.

We note, however, that LMDS has tuning parameters that allow us to generate a gamut of embeddings from very noisy with presumably less bias to very crisp with presumably more bias. In fact, we propose a new parametrized family of energy functions that has four parameters and that contains as special cases several previously proposed energy functions for graph layouts. Each parameter can be thought of as contributing a particular bias in the energy-minimizing configurations. For example, one parameter controls the balance of attractive and repulsive force, while another controls the tendency to show clusters crisply. The availability of four parameters to control configurations provides great flexibility on the one hand, but it also creates a quandary: how does one select appropriate parameters? In general, how does a practitioner know which among several configurations is most faithful to the underlying high-dimensional structure?

One way to answer this question is to propose measures of faithfulness other than the stress functions themselves. If such measures could be agreed on, they could be used to select viable configurations. Such measures would have to be fundamental and simple to be convincing. We propose one such measure, or family of measures: the average size of the overlap of K-nearest neighborhoods in low-dimensional configuration and in the

original high-dimensional space. This is a family of measures as there is one measure for every K. We call it the "Local Continuity" or "LC" meta-criterion. It is obviously simple, and it turns out to be practically useful for selecting good configurations.

With the LC meta-criterion in hand, we can draw an analogy of nonlinear dimension reduction to classification: In classification, the main measure of interest is the misclassification rate, yet classifiers are constructed as minimizers of surrogate criteria such as logistic loss or exponential loss. Similarly, in nonlinear dimension reduction, the main measure of interest is the LC meta-criterion, yet configurations are constructed as minimizers of energy functions. Similar to the surrogate criterion, energy functions are smooth and lend themselves for optimization, and similar to misclassfication rate, the LC meta-criterion is not smooth, hence difficult to optimize, yet of primary interest.

The article proceeds as follows. We first derive LMDS from Kruskal's distance scaling (Section 2). We then introduce the LC meta-criterion and its variants (Section 3), followed by illustrations of LMDS and the LC meta-criterion with two image datasets (Section 4). Next we introduce a rich four-parameter family of energy functions that vastly expands the search for desirable configurations (Section 5 and 6). Finally, we theorize about LMDS in particular cases that are analytically accessible (Section 7), and we give a few pointers to the minimization of energy functions with gradient descent and majorization (see the Appendix).

We conclude by noting that casting nonlinear dimension reduction as a graph layout problem for distance-weighted graphs entails a beneficial generality. The present work not only contributes to dimension reduction but also to graph layout. Furthermore, it contributes to "proximity analysis," the original target of MDS. Proximity analysis is concerned with mapping objects whose data consist of similarity or dissimilarity values for pairs of objects, and as such proximity analysis has had a long history in psychometrics. Our work contributes to proximity analysis as well by localizing multidimensional scaling and bringing graph layout ideas to bear on it.

Straddling three areas, we make liberal use of terminology from all three. In general, though, we adopt the term "configuration" from proximity analysis even though "embedding" or "graph layout" may be more appropriate depending on the context. We also adopt the term "dissimilarities" from proximity analysis to mean distances between high-dimensional feature vectors in the case of dimension reduction, and shortest-path-length distances in the case of graph layout. A correspondence of terminology between the three areas is given in Table 5 at the end of the article.

2 LOCAL MULTIDIMENSIONAL SCALING

2.1 Localization of MDS

The goal of MDS as a tool of proximity analysis is to map objects $i = 1, ..., N$ to "configuration" points $\mathbf{x}_1, ..., \mathbf{x}_N \in \mathbb{R}^k$ in such a way that the given dissimilarities $D_{i,j}$ between pairs of objects are well-approximated by the distances $\|\mathbf{x}_i - \mathbf{x}_j\|$. In so-called "metric distance scaling," one defines a measure of lack of fit, called "Stress," between dissimilarities $D_{i,j}$ and the configuration distances $\|\mathbf{x}_i - \mathbf{x}_j\|$. In the simplest case, Stress is a residual sum of squares:

$$\mathrm{MDS}^D(\mathbf{x}_1, ..., \mathbf{x}_N) \;=\; \sum_{i,j=1...N} \left(D_{i,j} - \|\mathbf{x}_i - \mathbf{x}_j\|\right)^2 \;. \tag{1}$$

We conceive of local MDS as a version of weighted metric distance scaling with a criterion of the form

$$\mathrm{LMDS}^D(\mathbf{x}_1, ..., \mathbf{x}_N) \;=\; \sum_{i,j=1...N} w_{i,j} \; \left(D_{i,j} - \|\mathbf{x}_i - \mathbf{x}_j\|\right)^2 \;.$$

For localization, let $\mathcal{N}$ be a symmetric set of nearby pairs (i, j). For example, $\mathcal{N}$ could be the symmetrized version of K-nearest neighbor pairs: a pair (i, j) is in $\mathcal{N}$ if object j is among the K nearest neighbors of object i, or object i is among the K nearest neighbors of object j (note that neither implies the other). Our proposal for a localized version of MDS is to use the dissimilarities $D_{i,j}$ for pairs in $\mathcal{N}$ with full weight $w_{i,j} = 1$ and replace the dissimilarities by a very large value D_∞ for all other pairs, but with small weight w:

$$\begin{aligned} \mathrm{LMDS}^D_{\mathcal{N}}(\mathbf{x}_1, ..., \mathbf{x}_N) \;&=\; \sum_{(i,j)\in\mathcal{N}} \left(D_{i,j} - \|\mathbf{x}_i - \mathbf{x}_j\|\right)^2 &(2)\\ &+\; \sum_{(i,j)\notin\mathcal{N}} w \cdot \left(D_\infty - \|\mathbf{x}_i - \mathbf{x}_j\|\right)^2 \;. &(3) \end{aligned}$$

The idea is to describe the "local fabric" of a graph or of a high-dimensional manifold in terms of the neighborhood set of pairs $\mathcal{N}$ and the corresponding small dissimilarities, and at the same time to introduce a bias for non-neighboring pairs of points by imputing them with a large dissimilarity with small weight. This bias should avoid the typical problem of MDS when the large dissimilarities are eliminated from the Stress: crumpling up of the configuration to a jumble. Crumpling means that non-near points are placed close together because of ignoring large distances; the imputation of a very large distance introduces a pervasive repulsive force throughout the configuration, similar to electric static that attempts to push points away from each other. The imputation of a

single large dissimilarity D_∞ is likely to introduce less noise than replacement of large dissimilarities with shortest-path estimates, as done by Isomap. In return, we accept a degree of bias to stabilize the configuration.

We call the above sum over $\mathcal{N}$ the "local stress" and the sum over the complement of $\mathcal{N}$ the "repulsion." The question is how to choose the weight w for the repulsion in relation to the imputed large dissimilarity D_∞. Our experience is that the weight w should be on the order of $1/D_\infty$:

$$\begin{aligned} \text{LMDS}^D_{\mathcal{N}}(\mathbf{x}_1, ..., \mathbf{x}_N) &= \sum_{(i,j)\in\mathcal{N}} (D_{i,j} - \|\mathbf{x}_i - \mathbf{x}_j\|)^2 \\ &+ \sum_{(i,j)\notin\mathcal{N}} \frac{c}{D_\infty} (D_\infty - \|\mathbf{x}_i - \mathbf{x}_j\|)^2 \end{aligned}$$

The constant c is a tuning parameter that adjusts the strength of the repulsion. In the following subsection we give a limiting argument that shows why w should indeed be on the order of $1/D_\infty$.

2.2 The Final Proposal of a Local MDS Criterion

In the tentative local MDS criterion (2) - (3) introduced in the previous subsection, there is still a question of how to choose D_∞. For an answer, we revert back to the stress with weight w shown in the repulsion term (3). We expand this term, ignoring pure functions of the dissimilarities that do not affect the minimization problem:

$$\begin{aligned} \text{LMDS}^D_{\mathcal{N}}(\mathbf{x}_1, ..., \mathbf{x}_N) &\sim \sum_{(i,j)\in\mathcal{N}} (D_{i,j} - \|\mathbf{x}_i - \mathbf{x}_j\|)^2 && (4) \\ &\quad -2wD_\infty \sum_{(i,j)\notin\mathcal{N}} \|\mathbf{x}_i - \mathbf{x}_j\| && (5) \\ &\quad +w \sum_{(i,j)\notin\mathcal{N}} \|\mathbf{x}_i - \mathbf{x}_j\|^2 && (6) \end{aligned}$$

Now, as D_∞ goes to infinity, w must go to zero at least on the order of $1/D_\infty$ in order to prevent term (5) from blowing up. The weight w cannot go to zero faster than $1/D_\infty$, though, because otherwise both terms (5) and (6) vanish. This leaves $w \sim 1/D_\infty$ as the only non-trivial choice. In this case, however, term (6) will disappear, which amounts to a simplification. In summary, it makes sense to pass to the limit $D_\infty \to \infty$ constraining w, for example, by $w = t/2D_\infty$, where t is a fixed constant:

$$\text{LMDS}^D_{\mathcal{N}}(\mathbf{x}_1, ..., \mathbf{x}_N) = \sum_{(i,j)\in N} (D_{i,j} - ||\mathbf{x}_i - \mathbf{x}_j||)^2 - t \sum_{(i,j)\notin N} ||\mathbf{x}_i - \mathbf{x}_j|| \qquad (7)$$

An obvious benefit of passing to the limit with a constraint is the replacement of two parameters (w and D_∞) with a single parameter t. In addition, the two remaining terms in (7) have intuitive meaning:

- $\sum_{(i,j)\in\mathcal{N}} (D_{i,j} - \|\mathbf{x}_i - \mathbf{x}_j\|)^2$: This term correlates $\|\mathbf{x}_i - \mathbf{x}_j\|$ with $D_{i,j}$ for $(i,j) \in \mathcal{N}$ and is responsible for preserving the intended local structure.
- $\sum_{(i,j)\notin\mathcal{N}} \|\mathbf{x}_i - \mathbf{x}_j\|$: This term contributes repulsion outside $\mathcal{N}$ and is responsible for pushing points away from each other if they are not locally linked.

The local stress (7) attempts to preserve the local fabric by driving down $\sum_{(i,j)\in N}(D_{i,j} - ||\mathbf{x}_i - \mathbf{x}_j||)^2$, while at the same time spreading the configuration out by driving up $\sum_{(i,j)\notin N} ||\mathbf{x}_i - \mathbf{x}_j||$. The relative strength of the two forces against each other is balanced by the tuning parameter t. Adjusting this parameter allows users to search for the most meaningful configurations.

A note on notation: We write $\text{LMDS}^D_{\mathcal{N}}$ when we allow a general graph $\mathcal{N}$, but we may write LMDS^D_K when the graph $\mathcal{N}$ is a symmetrized K-nearest neighbor graph.

2.3 Invariance Requirements for the LMDS Criterion

The tuning parameter t is unsatisfactory in some ways as the following discussion shows. The problems have to do with a lack of invariance under desirable transformations. These problems can be easily corrected, however, resulting in the final form of the tuning parameter.

- *Invariance under change of units:* If the measurement units of the dissimilarities $D_{i,j}$ are changed, then t needs to change along because it has the same units as $D_{i,j}$. One can eliminate the units of t by replacing it with $t = \text{median}_{\mathcal{N}}(D_{i,j})t'$, for example, where the new parameter t' is unit free.
- *Transformation of t under change of K in K-nearest neighbors:* The size of the first sum in (7) is $|\mathcal{N}|$, which is of the order of KN; the size of the second sum is about $N^2/2$, assuming $K \ll N$. This imbalance in the number of terms suggests that it may be useful to balance the repulsive force and the local stress by making t' a suitable function of K and N. A plausible proposal is $t' = \frac{K}{N}t''$. The new parameter t'' is expected to have comparable effects across different choices of the neighbor set size K.

These two considerations can be brought under one roof by reparametrizing the tuning constant t as follows:

$$t \;=\; \frac{K}{N} \cdot \text{median}_{\mathcal{N}}(D_{i,j}) \cdot \tau \;, \tag{8}$$

where τ is unit free and scales along with K. In our experience, a generally successful strategy is to start with a relatively high value $\tau = 1$ in order to obtain a stable configuration, and lower τ successively as far as 0.01, depending on stability and interpretability of the resulting configurations.

There remains, though, the natural question of how to choose tuning parameters such as τ and, more generally, how to compare configurations obtained from different methods. This is the topic of the following subsection.

3 THE LC META-CRITERION FOR MEASURING LOCAL CONTINUITY

3.1 Motivation and Definition

Experimentation with tuning parameters such as τ has an element of subjectivity: whether an embedding is judged as meaningful may depend on the user. Subjective choices could be avoided if it were possible to propose a criterion that quantifies clearly desirable properties of configurations. Such properties would have to be rather minimal so as to grant a consensus among users. In general terms the intended criterion would have to measure some aspect of faithfulness of the embedding vis-à-vis the high-dimensional structure. In light of the earlier discussions, a particular aspect that is desirable here is local truthfulness of the configurations. The critical step we now take is to interpret the intuitive notion of local truthfulness to mean *preservation of neighborhoods*. That is, neighbors of an object in terms of the target dissimilarities $D_{i,j}$ should be mapped to neighbors of the object in the configuration. This formulation is of course reminiscent of the definition of continuity of maps in topology. Achieving continuity of mapping is a fundamental goal of localization, the principle shared by LLE, Isomap, and LMDS. Measuring the degree of continuity in terms of preservation of neighborhoods is the purpose of the LC meta-criterion to be introduced here. It provides an independent yardstick for judging the success of preserving the local fabric.

We now give the definition of a first version of the LC meta-criterion: For the i'th point we form the set $\mathcal{N}_K^D(i)$ of K nearest neighbors *with regard to the target dissimilarities* $D_{i,j}$; similarly, for the i'th point we form the set $\mathcal{N}_K^X(i)$ of K nearest neighbors *with regard to the configuration distances* $\|\mathbf{x}_i - \mathbf{x}_j\|$. The LC meta-criterion proposed here measures the preservation of local structure in terms of the overlap between $\mathcal{N}_K^D(i)$ and $\mathcal{N}_K^X(i)$. Ideally, $\mathcal{N}_K^X(i)$ would coincide with $\mathcal{N}_K^D(i)$, but this is

not usually achievable with d-dimensional configurations $\{\mathbf{x}_i\}_{i=1,\ldots,N}$. It is therefore sensible to measure the local preservation of K-nearest neighbor structure at the point i by $|\, \mathcal{N}_K^D(i) \,\cap\, \mathcal{N}_K^X(i) \,|$, the cardinality of the overlap:

$$N_K(i) \;=\; |\, \mathcal{N}_K^D(i) \,\cap\, \mathcal{N}_K^X(i) \,| \tag{9}$$

This quantity is a function of the i'th point and represents therefore the pointwise version of the LC meta-criterion. The pointwise quantities are turned into a global version of the LC meta-criterion by averaging over the sample:

$$N_K \;=\; \frac{1}{N} \sum_i N_K(i) \tag{10}$$

Both the local quantity $N_K(i)$ and global quantity N_K are bounded by K, and $N_K = K$ would therefore mean maximal faithfulness: $N_K(i) = K$ for all i, meaning perfect identity of K-nearest neighborhoods in terms of the data $D_{i,j}$ and in terms of the configuration.

As stated above, the main purpose of the LC meta-criterion is to guide the search for configurations that faithfully render local structure. It is not meant as a direct optimization criterion; this is the role of the local stress introduced above and more generally the local energy functions to be introduced below. The meta-criterion is not smooth and therefore difficult and unstable to optimize; by comparison, stress and energy functions are smooth and amenable to direct optimization. Given, however, a collection of configurations obtained from the optimization of various smooth criteria, one can use the LC meta-criterion to select among these configurations. It can be used to select among configurations obtained from the same method but with different values of tuning parameters, or from different methods altogether. Selection based on the LC meta-criterion can be quite effective, as will be illustrated in the next section. A successful example of its use is shown in Figure 5 where a trace of N_K is shown as a function of the repulsion parameter τ. A single global maximum points to a particular choice of τ that makes the corresponding configuration most successful in terms of rendering K-nearest neighborhoods. (Details for this particular data example will be given in the next section.)

We conclude this introduction by noting that the device of a meta-criterion is not uncommon in statistics; in fact, it has many precedents. In classification, for example, one optimizes smooth loss functions such as logistic loss (in logistic regression) or exponential loss (in boosting); for ultimate performance measurement, however, one uses misclassification error, which is a non-smooth and unstable meta-criterion and not generally used for optimization. Another large area that uses meta-criteria is

model selection: One fits models by optimizing likelihoods but selects among models by penalizing the likelihood with terms that measure the cost of model complexity. In a similar vein, smoothing splines are fitted with penalized likelihoods, but the size of the penalty may be selected with a generalized cross-validation criterion. The meta-criterion proposed here should be seen in the context of these widely accepted statistical approaches.

3.2 Normalization and Adjustment for Baselines

The LC meta-criterion N_K is useful for selecting a tuning parameter when it has been decided that the goal is faithfulness in terms of K-nearest neighborhoods. One may, however, want to compare LC meta-criteria for different values of K. If so, one may want to take advantage of the fact that N_K has K as an upper bound and normalize the LC meta-criterion to the 0-1 interval:

$$M_K \;=\; \frac{1}{K} N_K \tag{11}$$

In this form the LC meta-criterion expresses average overlap of $\mathcal{N}_K^D(i)$ and $\mathcal{N}_K^X(i)$ as a fraction of K. Thus an overlap of 3 for $K = 4$ would be compared with an overlap of 15 for $K = 20$. However one may feel about such equalization, it has the benefit that it rids a plot of M_K against K of the uninformative ascent that the non-normalized count version N_K would show. An example of an M_K trace is shown in Figure 6 (black curve). The shape of this trace makes it clear, however, that there is a problem with comparisons of fractional overlap because of its behavior for K near N: $M_{N-1} = 1$. This version of the LC meta-criterion is therefore not useful for large K.

In principle we could leave this problem with the recommendation that M_K be used only for K small relative to N. Yet, it may be worthwhile to pursue a solution by gauging M_K against some baseline that could be subtracted from M_K. An example could be derived from the idea of random overlap: if one assumed that the configuration were absolutely random and reflected nothing of the true K-nearest neighborhoods according to the dissimilarities $D_{i,j}$, there would still exist an overlap due to chance alone. In fact, its expected value is $K \cdot K/(N-1)$, and its fractional version is $K/(N-1)$ (the denominator being $N-1$ because the center point is not part of the randomization). It therefore makes sense to use the *adjusted LC meta-criterion*

$$M_K^{adj} \;=\; M_K \;-\; \frac{K}{N-1}\;. \tag{12}$$

An example of an adjusted trace is also shown in Figure 6 (red curve). As that example shows, it succeeds in that it creates a single absolute maximum of the trace. The

adjusted version of the LC meta-criterion can therefore be used to select a sensible value of K as well.

In practice, selection may work as follows: First, one chooses a stress or energy function for optimization; this function may have tuning parameters such as the relative weight of the attractive and repulsive forces. Second, one generates several configurations for different values of the tuning parameters. Third, one selects that configuration which generates the highest value of the meta-criterion M_K. Fourth, one repeats this procedure for various values of the neighborhood size K and selects the K that generates the highest value of the adjusted meta-criterion M_K^{adj}.

The selection procedure just described makes the implicit assumption that a configuration generated with LMDS with a K-nearest neighborhood graph should be judged by the LC meta-criterion corresponding to the same K. In other words, the K in LMDS_K^D is the same as the K in M_K. However, nothing prevents us from judging an LMDS configuration obtained as a minimizer of $\text{LMDS}_{K_0}^D$ with an LC meta-criterion M_K^{adj} for $K \neq K_0$. In fact, it might be informative to draw a trace of $K \mapsto M_K^{adj}$, where K runs through a relatively dense gird of values. If it were the case that a trace of one configuration dominated the trace of another configuration globally, one could say that the latter is an inadmissible configuration. An impression of such traces can be obtained from Figure 7, which shows both raw and equivalent adjusted traces (details to be given in the next section). The adjusted traces can be used to select an optimal nearest neighborhood size K_0, which in this case is for about $K_0 = 8$ (green trace).

As we will see, adjustment of M_K with random overlap as the baseline may not be sufficient. Other, more stringent baselines might be called for. For example, one could use configurations obtained from regular MDS (LMDS with $K_0 = N - 1$) as the baseline and subtract its trace $K \mapsto M_K$ from any other trace. The idea of this approach is that localization ($K_0 \ll N - 1$) should produce benefits over non-localized MDS in terms of M_K for smaller K. An example of adjustment with the MDS baseline is shown in Figure 13 (again, see the next section for details).

3.3 The Pointwise LC Meta-Criterion and Its Use for Diagnostics

We intentionally divided the introduction of the meta-criterion into two steps: a pointwise and a global version. While we just explained the use of the global version, it is quite evident that the pointwise version has merit also. In fact, the values $N_K(i)$ can be used to judge the local faithfulness of a configuration. We note that it is entirely conceivable that a configuration excels in faithfully rendering the local fabric for a ma-

jority of the objects but fails to do so for a minority. Detecting the existence of this minority and identifying it is of obvious importance. It can be achieved, for example, by making graphical use of the values $N_K(i)$ by coloring configuration points according to high and low values of $N_K(i)$. Illustrations of this procedure will be given below (for example, Figure 8). In summary, with the pointwise version of the meta-criterion we avail ourselves of a potentially powerful diagnostic tool for local faithfulness of configurations.

4 EXAMPLES

4.1 Highly Structured Facial Images

In this section we apply Local MDS to two sets of facial image data. We are interested in this kind of data for two reasons: First, facial image data are often intrinsically low-dimensional and hence promising for dimension reduction. We will discuss the details of this aspect below. Second, facial image data have been widely used in the recent literature on nonlinear dimension reduction. It is then of interest to compare the performance of LMDS with those of competing methods on the same data sets.

Images are essentially arrays of pixels. In the simplest case the pixels describe light intensities, that is, values on a grey scale. By treating each pixel as one dimension, an image becomes a single data point in a very high dimensional space, called image space. For example, an image of size 64×64 pixels is an element of a 4096-dimensional image space. These many dimensions, however, are highly redundant. Two sources of redundancy are as follows: First, in most images a majority of nearby pixel pairs have nearly equal light intensities; this fact translates to considerable correlations between the dimensions. Second, there are often far fewer degrees of freedom when a collection of images shows different aspects of the same object. The degrees of freedom in the facial image datasets we consider below include postures, facial expressions and lighting direction. For example, the dataset considered in the next subsection will have only three degrees of freedom: left-right pose, up-down pose and lighting direction (Figure 2). These intrinsic degrees of freedom should be found by any successful dimension reduction method.

4.2 Example 1: Sculpture Face Data

This data set includes 698 images of size 64×64 of a sculpture face and was originally published in the Isomap paper (Tenenbaum et al. 2000). For comparison, we use the

Figure 2: *Sculpture Face Data*

same size of neighborhoods, $K = 6$, as these authors, based on Euclidean distances in image space. This choice will be used both for LMDS and LLE.

Figure 3 shows that the 3-D configuration generated by Local MDS is close to a hyperrectangle. We selected a few points and labeled them in both views with the corresponding images. On the widest dimension (drawn horizontally) the images show the transition of the left-right pose. The second widest dimension (vertical axis of the upper view) shows the transition of the up-down pose, and the third dimension (vertical axis of the lower view) shows the transition of the lighting direction.

A nice feature of this data set is that the three obvious dimensions were measured physically and added to the data. Thus each image has three angular parameters attached for, respectively, the left-right pose, up-down pose, and lighting direction. We used these angles to color the data points of the configurations with yellow, orange and red. Such coloring would let us see the transitions if the dimension reduction method uncovered it in its configurations. Colored configurations for PCA, MDS, Isomap, LLE and LMDS are shown in Figure 4, with the 3-D configurations rotated to best reveal the color separations. We note that the three colors overlap in some of the plots; in particular the up-down transition is not well-captured by PCA, MDS and LLE. Interesting is also a wrap-around structure in the PCA configuration. The meeting

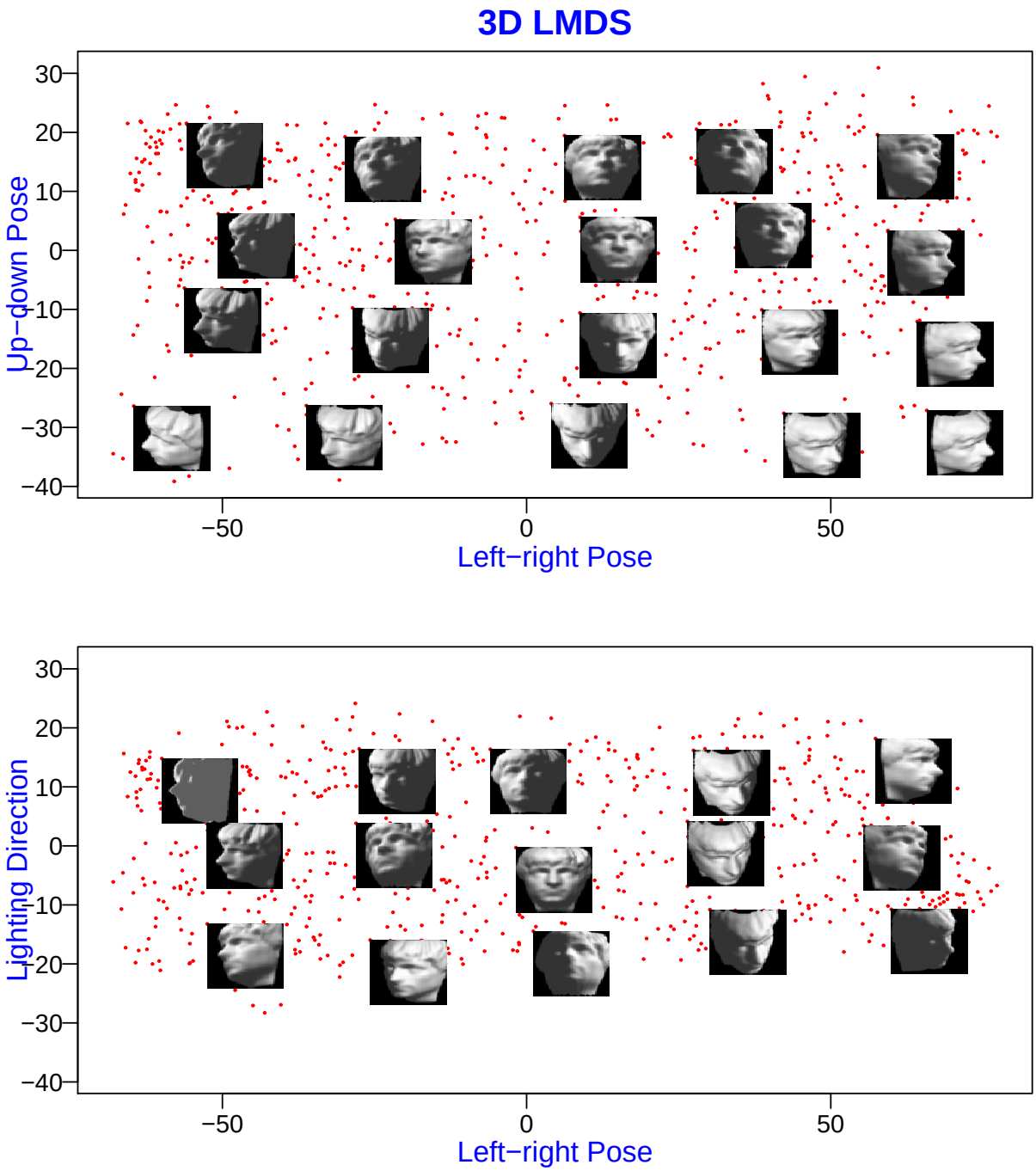

Figure 3: *Sculpture Face Data: 3-D LMDS configuration,* $K = 6$ *with optimized* τ.

PCA	MDS	Isomap	LLE	LMDS
2.6	3.1	4.5	2.8	5.2

Table 1: *Sculpture Face Data: The meta-criterion N_K.*

of the extremes (yellow and red) visible in the frame of PCA and 'Left-right Pose' is probably caused by the darkness of images showing the extreme left and extreme right poses. The LLE configuration shows characteristic spikes which we observed in most applications: LLE is prone to linear artifacts in its configurations. It may be the case that LLE is more difficult to fine-tune for competitive performance than Isomap and LMDS.

The Isomap and LMDS configurations in Figure 4 look quite similar and show clear color separations, which shows that they successfully find the nonlinear structure in this data set and map it to surprisingly straight directions in the configurations. Of the two, LMDS shows crisper boundaries in the configuration, which in our experience is a general observation that applies to other datasets as well. The fuzziness in the Isomap configurations can be explained by the noise that is inherent when adding up local edges in the calculation of shortest paths across the near-neighbor graph.

We also compared these methods according to the LC meta-criterion. Table 1 shows N_K . We see that Isomap and LMDS generate better configurations in terms of this criterion, with LMDS winning out by a small margin.

A remark on the optimization of the LMDS criterion: As we generate a configuration, we usually start with relatively large weight for the repulsion, say, $\tau = 1$, and lower it to values as small as $\tau = 0.01$. When we performed experiments with τ by trying a wider range of τ and decreasing τ by smaller stepsizes, we observed that the LC meta-criterion $N_K(\tau)$ first ascends and then descends, with a clear peak for a specific value of τ, as reflected in Figure 5. In this particular data example, the peak is achieved at $\tau = 0.006$, where the count meta-criterion is as high as 5.2 out of 6-nearest neighbors. The decrease of the meta-criterion to the left of the peak indicates that too weak a repulsion allows the local stress to take over and let the configuration degenerate into noise. Therefore, repulsion is essential for the stability of the optimization and for the quality of the configurations.

For comparison purposes, we used the same number of neighbors $K = 6$ as in the Isomap paper where the data set was originally published. The authors of this paper gave no discussion of the choice of K, though. Presumably they used trial-and-error to find a useful configuration. We can be more systematic in this regard as we can

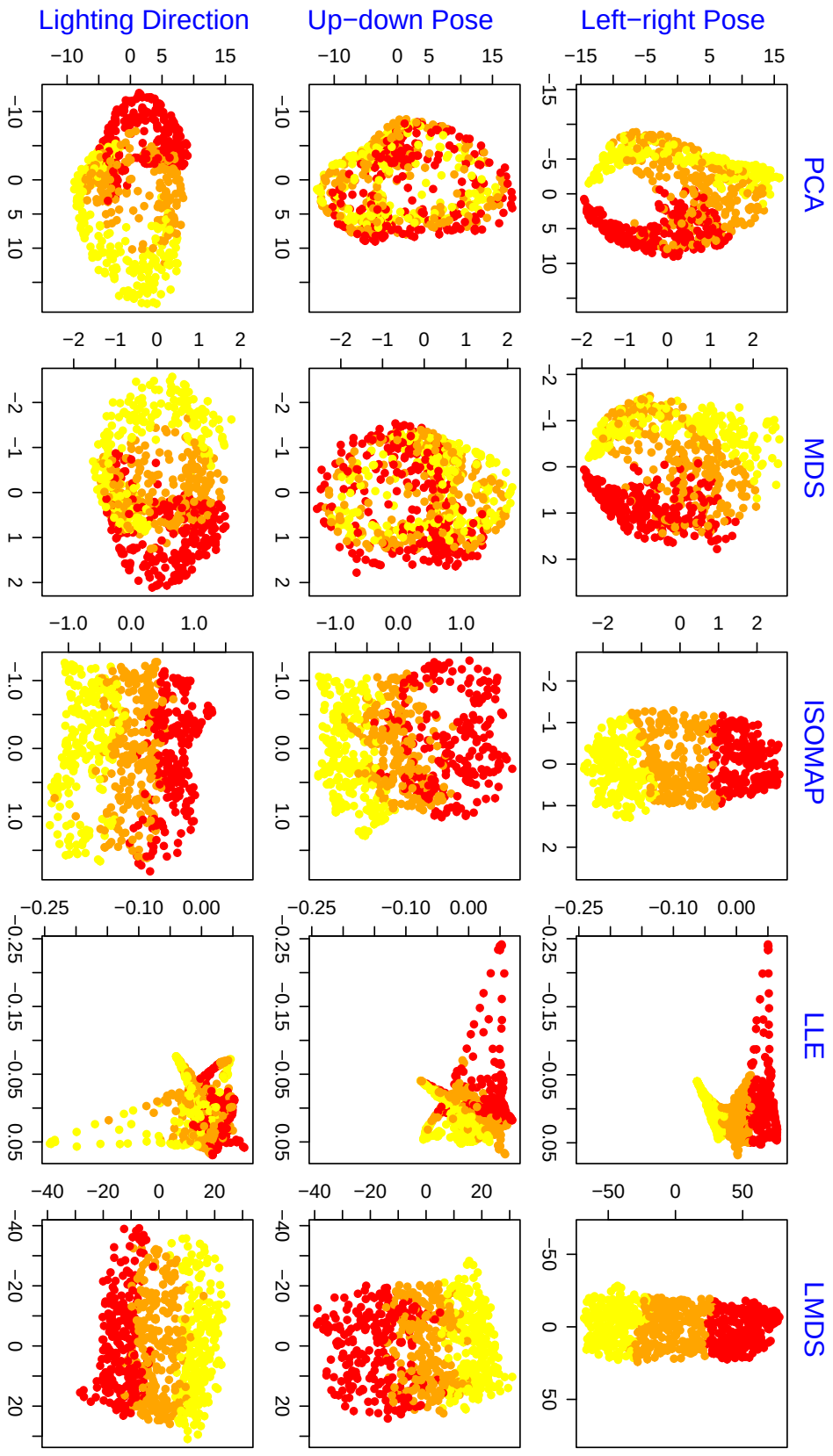

Figure 4: *Sculpture Face Data: 3-D configurations from five methods.*

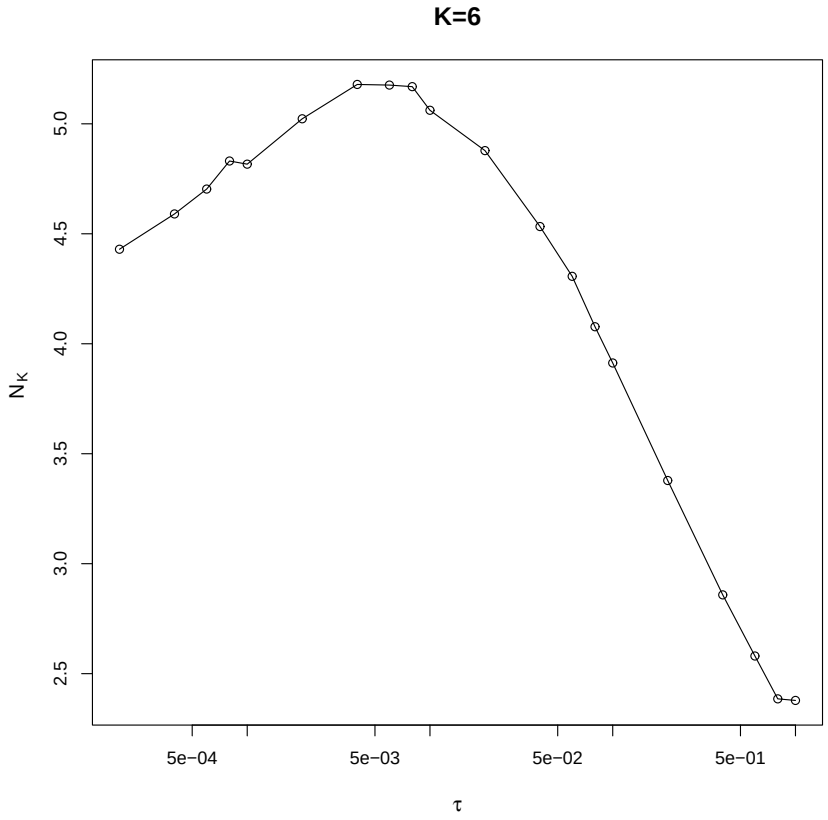

Figure 5: *Sculpture Face Data: Trace of the LC meta-criterion as a function of τ, the weight parameter for the repulsive force.*

use the meta-criterion to choose K. We tried a number of K's in the range of 4 to 650 by increasing K in increments of 2 in the lower range (4-20), 25 in the middle range (25-200), and 50 in the higher range (250-650). For each K, we optimized the meta-criterion over the repulsion weight τ, and we used this optimized value of M_K as the criterion for selecting K. As a side note, for different K's the highest M_K is achieved at different values of τ, and larger K usually requires a larger τ.

In Figure 6 we plot the relation between K and the τ-optimized value of M_K (black) and the adjusted $M_K^{adj} = M_K - K/(N-1)$. The parameter K plays the role of both the K-nearest neighborhood size used in the LMDS stress and in the meta-criterion. The graphs show an unexpected shape: there are two local maxima at $K = 8$ and $K = 16$, the latter being quite minor; after these peaks, the unadjusted trace dips before it ascends to very high values, with a final upper limit of $M_K = 1$ and $M_K^{adj} = 0$ for $K = N-1$. It is obvious that the high values of M_K for K larger than 200, say, are

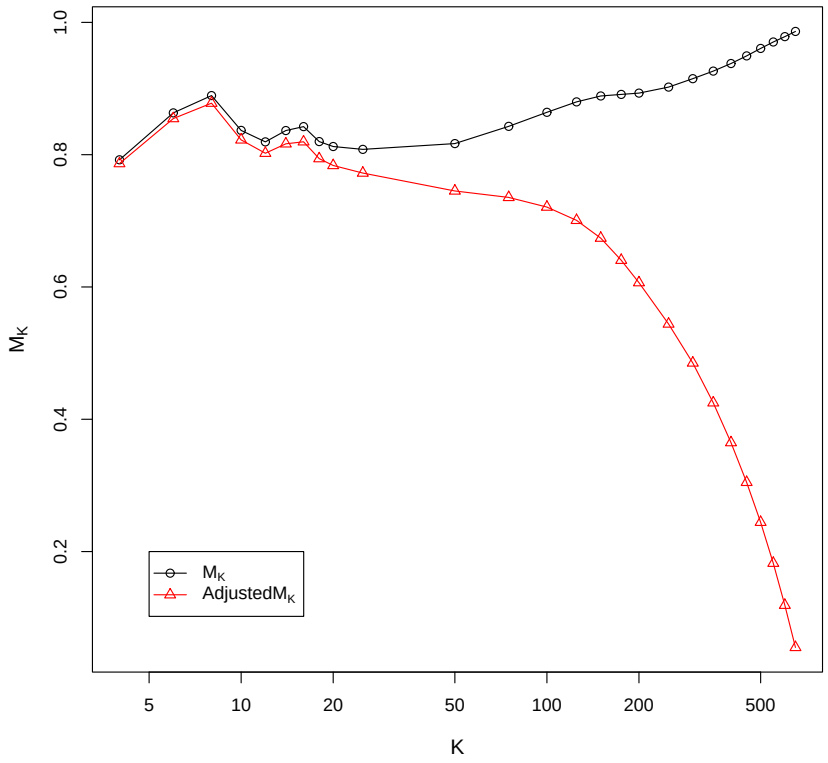

Figure 6: *Sculpture Face Data: Raw (black) and adjusted (red) traces of the LC meta-criterion. The horizontal axis represents the size parameter K for the K-nearest neighborhood in both the LMDS stress and the LC meta-criterion. Adjustment in the red trace is by subtracting the expected size of random overlap of K-nearest neighborhoods.*

not meaningful: K this large does not reflect local structure. Adjustment for random overlap has the desired effect of creating a single absolute maximum at $K = 8$. It seems therefore that the Isomap authors' choice $K = 6$ is not far from the maximum.

In the next exercise with the LC meta-criterion, we unlink the choice K_0 for the nearest neighbor graph in $\text{LMDS}^D_{K_0}$ from the K in the LC meta-criterion M_K. In Figure 7, we plot the relations between K and M_K for a number of τ-optimized configurations generated from $\text{LMDS}^D_{K_0}$, both raw (left) and adjusted (right). Interestingly, the M_K traces reach their maxima at or near $K = K_0$. This fact may lend support to the practice of equating the K's of LMDS^D_K and M_K. This fact also lends general support to the fundamental premise that LMDS^D_K is a suitable optimization criterion for getting local structure right as measured by M_K.

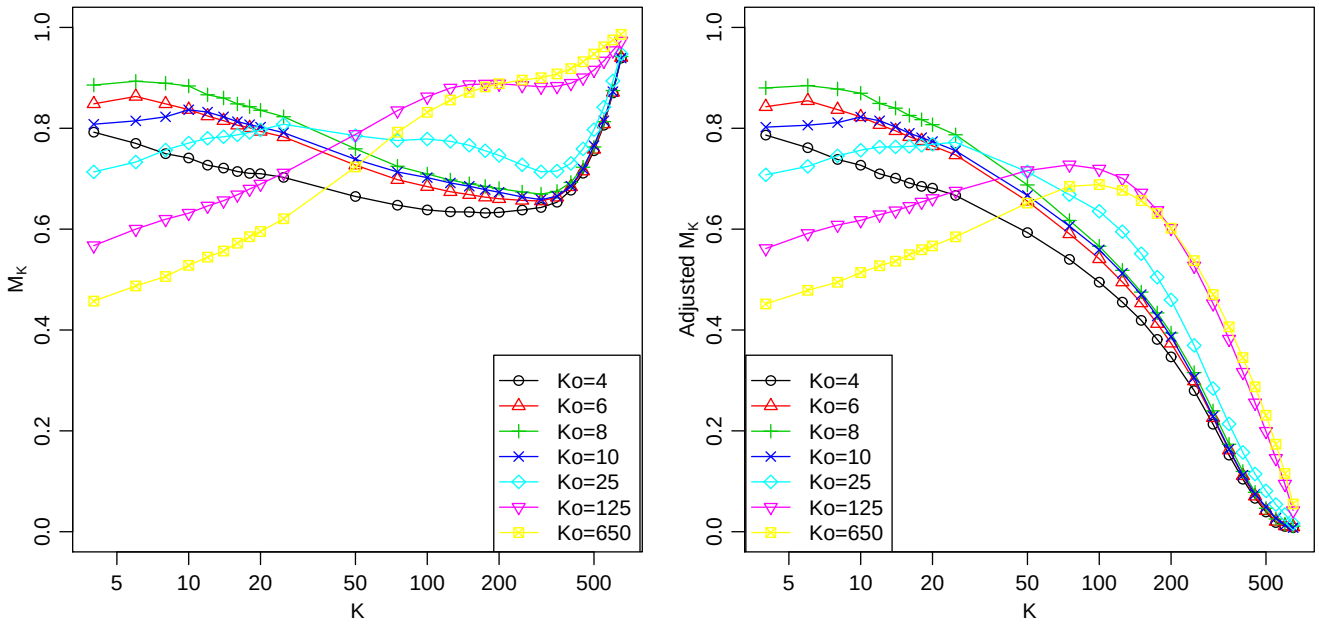

Figure 7: *Sculpture Face Data: Traces of the LC meta-criterion, raw (left) and adjusted (right). Each trace corresponds to one configuration obtained with LMDS for a given size parameter K_0 used in the LMDS stress. Each trace shows the values of the LC meta-criterion as a function of its own neighborhood size K, chosen independently of the neighborhood size of the LMDS stress.*

Finally, we use the pointwise version $N_6(i)$ of the LC meta-criterion as a diagnostic by color coding the configurations from PCA, Isomap, LLE and LMDS. The result is shown in Figure 8. The overall impression of the coloring reflects the average level of $N_6(i)$, namely, N_6. Correspondingly, the lower quality configurations from PCA and LLE are overall more red, those from Isomap and LMDS more green. It appears that Isomap's configuration is of lesser quality in the left and right extremes where the color is more densely orange, while LMDS' configuration is of lesser quality in the center where again orange color is more concentrated.

4.3 Example 2: Frey Face Data

This data set, from the LLE paper (Roweis and Saul 2000), includes 1965 images of the face of a single person, Brendan Frey, taken as sequential frames from a short piece of video footage. The size of the images is 20x28 and hence the image space is

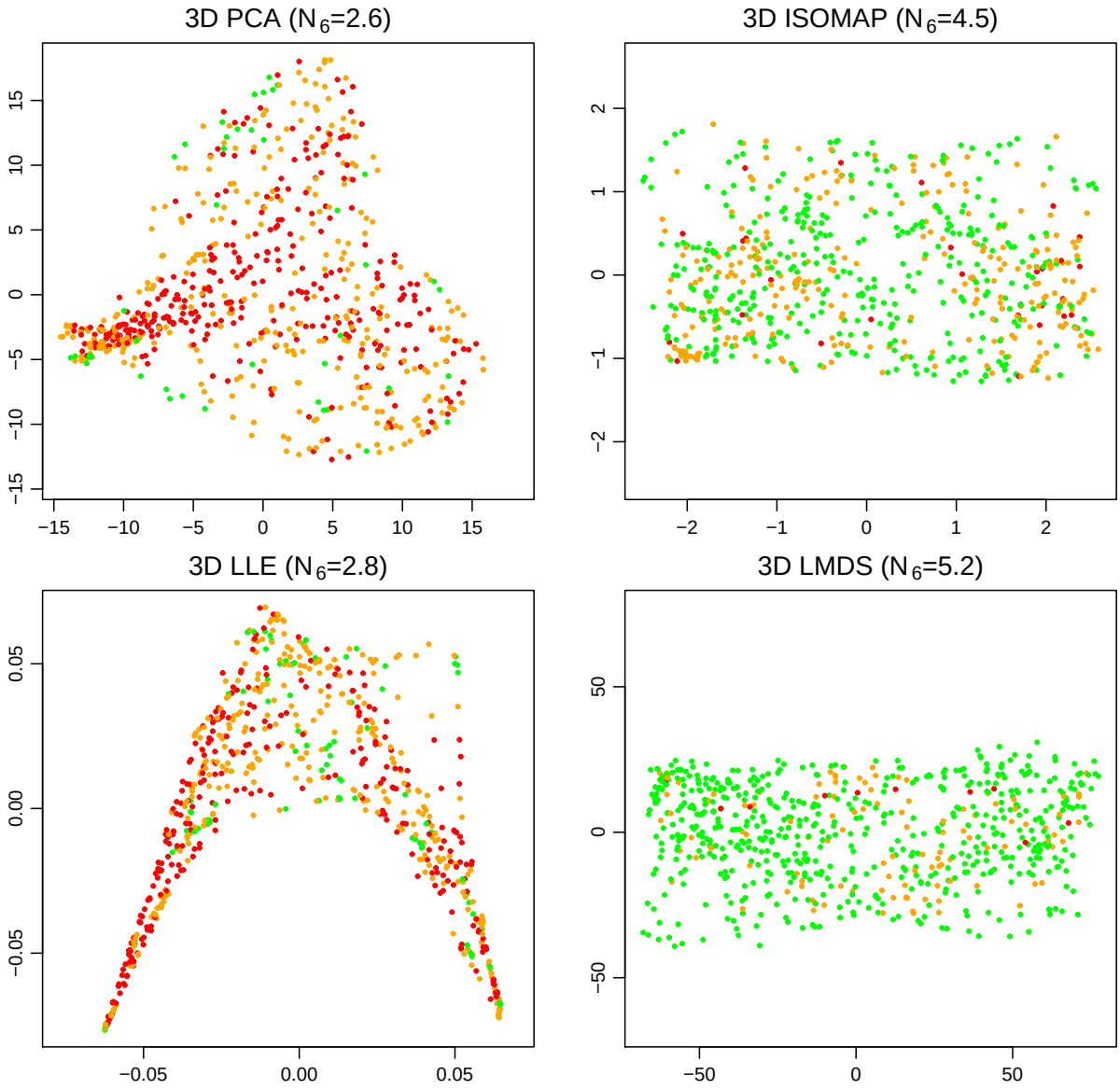

Figure 8: *Sculpture Face Data: The pointwise values of the LC meta-criterion $N_6(i)$ are used to color code the configuration points. Shown are 2-D views of 3-D configurations from four methods. The colors code $N_6(i)$ as follows: red $\sim$ 0 to 2, orange $\sim$ 3 to 4, green $\sim$ 5 to 6.*

560-dimensional.

As with the sculpture face data, we applied PCA, MDS, Isomap, LLE and LMDS to this data set. For the localized methods, nearest neighbor computations were based on plain Euclidean distances in 560-dimensional image space. We used $K = 12$ as in the LLE paper. With visualization in mind we chose again a reduction to three dimensions. The results are shown in Figures 9, 10, and 11, the last with two LMDS configurations, one from $K = 12$ and one from $K = 4$. For interpretation, we labeled a selection of points with the images themselves. The coloring of the points is the same in all plots; it is based on painting the clusters and substructures visible in the LMDS configuration for $K = 12$, which may be the most articulated of the six configurations. The joint coloring allows us to compare subsets of interest across configurations. The 2-D projection of each 3-D configuration was interactively chosen to reveal the most structure and the best color separations.

The overall impression is that the configurations generated by LMDS show more structure compared to the configurations generated by the other methods. In the LMDS configurations we can see two major clusters (top and bottom of each plot), some smaller sub-clusters and transition paths between clusters. As the labeling shows, the main clusters correspond to, respectively, smiling faces (top) and serious faces (bottom). This main division is visible in all configurations. A second feature shared by all configurations is the existence of a dimensions that correponds roughly to the left-right pose. All six plots were rotated within the projection to line up the serious-smiling division with the vertical dimension and the left-right pose with the horizontal dimension.

The aspect where the six configurations show differences the most is in the degree to which the major divisions are divided into subclusters and linked by transitions between them. While the PCA and MDS configurations show essentially only the two base dimensions described in the previous paragraph, they give virtually no indication of further subdivisions. By comparison, LLE produces a configuration that is quite structured, but it suffers again from a tendency to spikiness which is most likely artifactual. While Isomap comes closest to LMDS in regard to subclusters and transitions, its noisiness obscures in particular transitions between clusters. Finally, the LMDS configurations show transitions between subclusters that can be shown to be real. Our confidence in this statement is based on the fact that these images stem from video footage that has an underlying timeline. We therefore connected the points with lines according to the time order of the images, as in Figure 12. The LMDS configurations make it clear that there exist essentially four transitions between smiling faces at the top and serious faces at the bottom. One transition links the gray cluster below and the

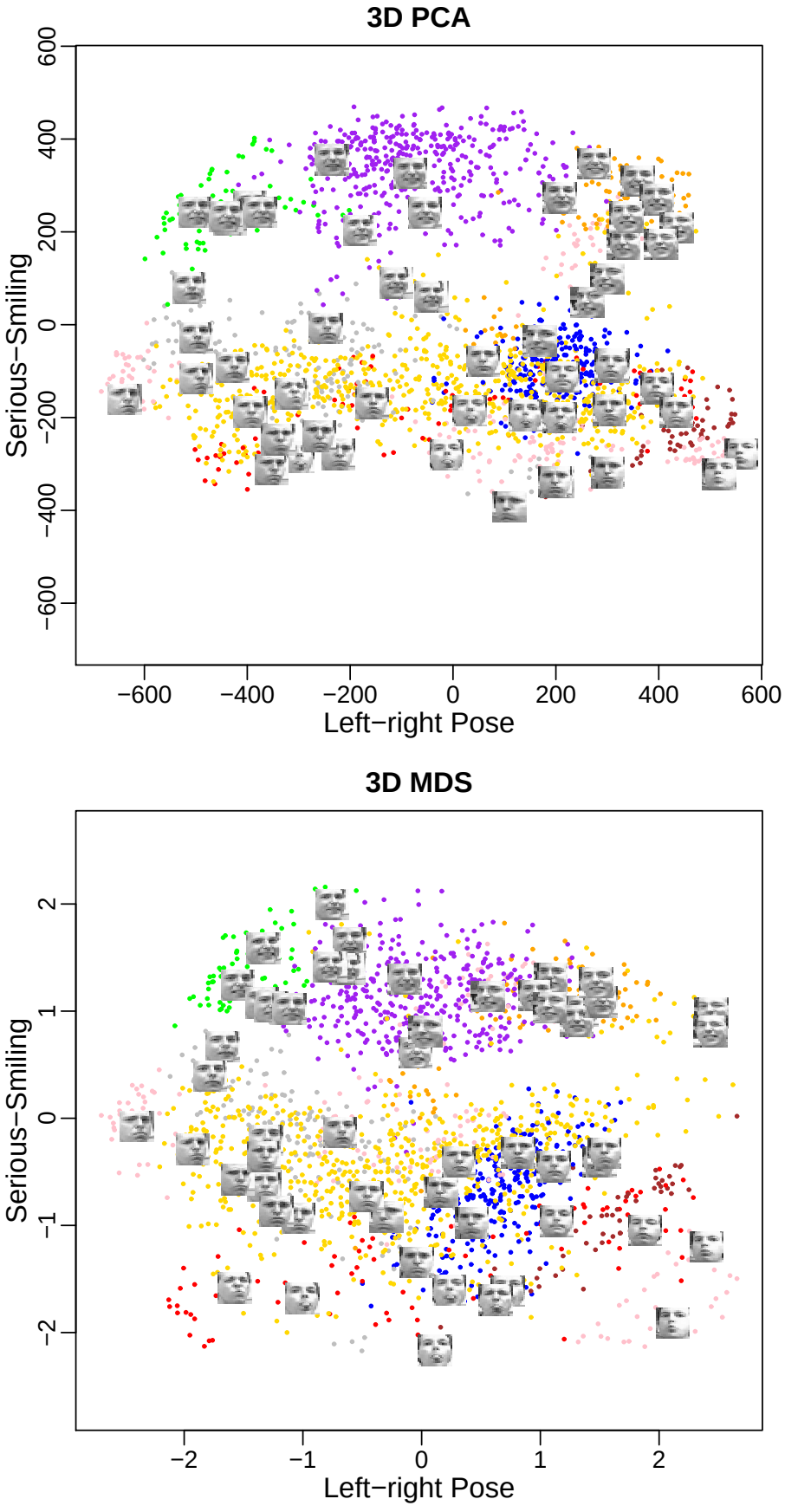

Figure 9: *Frey Face Data: Configurations from PCA and MDS.*

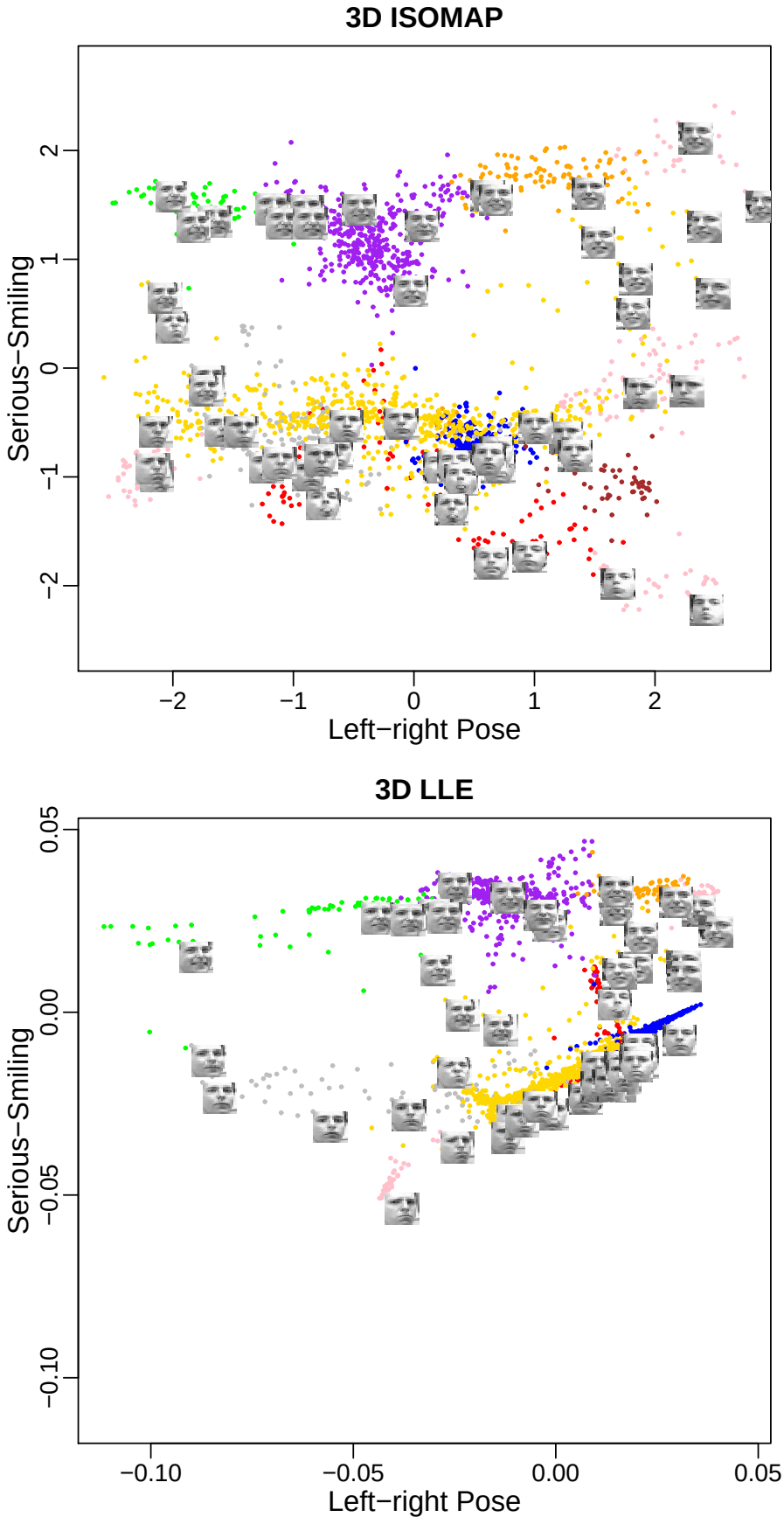

Figure 10: *Frey Face Data: Configurations from Isomap and LLE.*

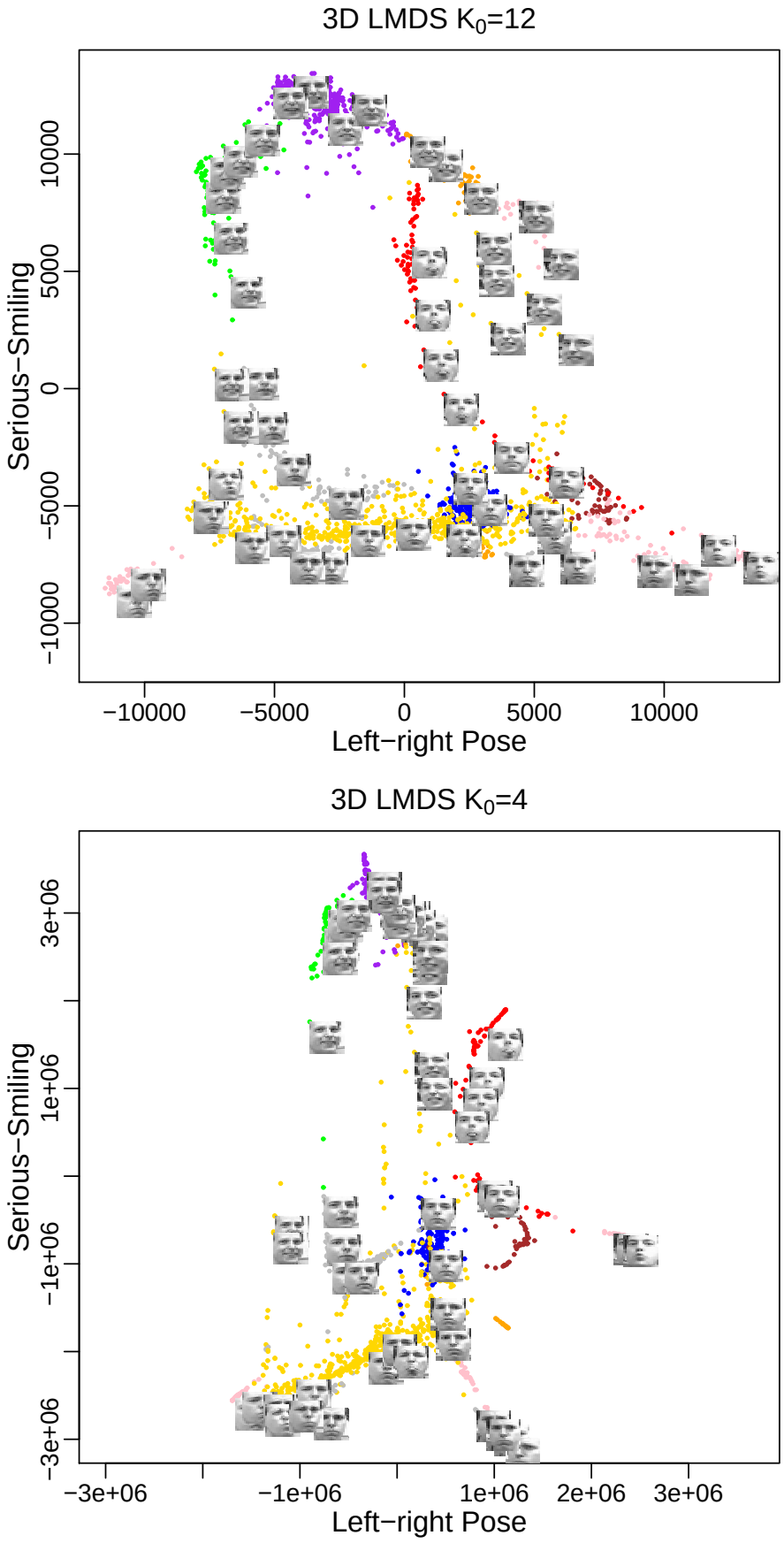

Figure 11: *Frey Face Data: Configurations from LMDS for* $K_0 = 12$ *and* $K_0 = 4$.

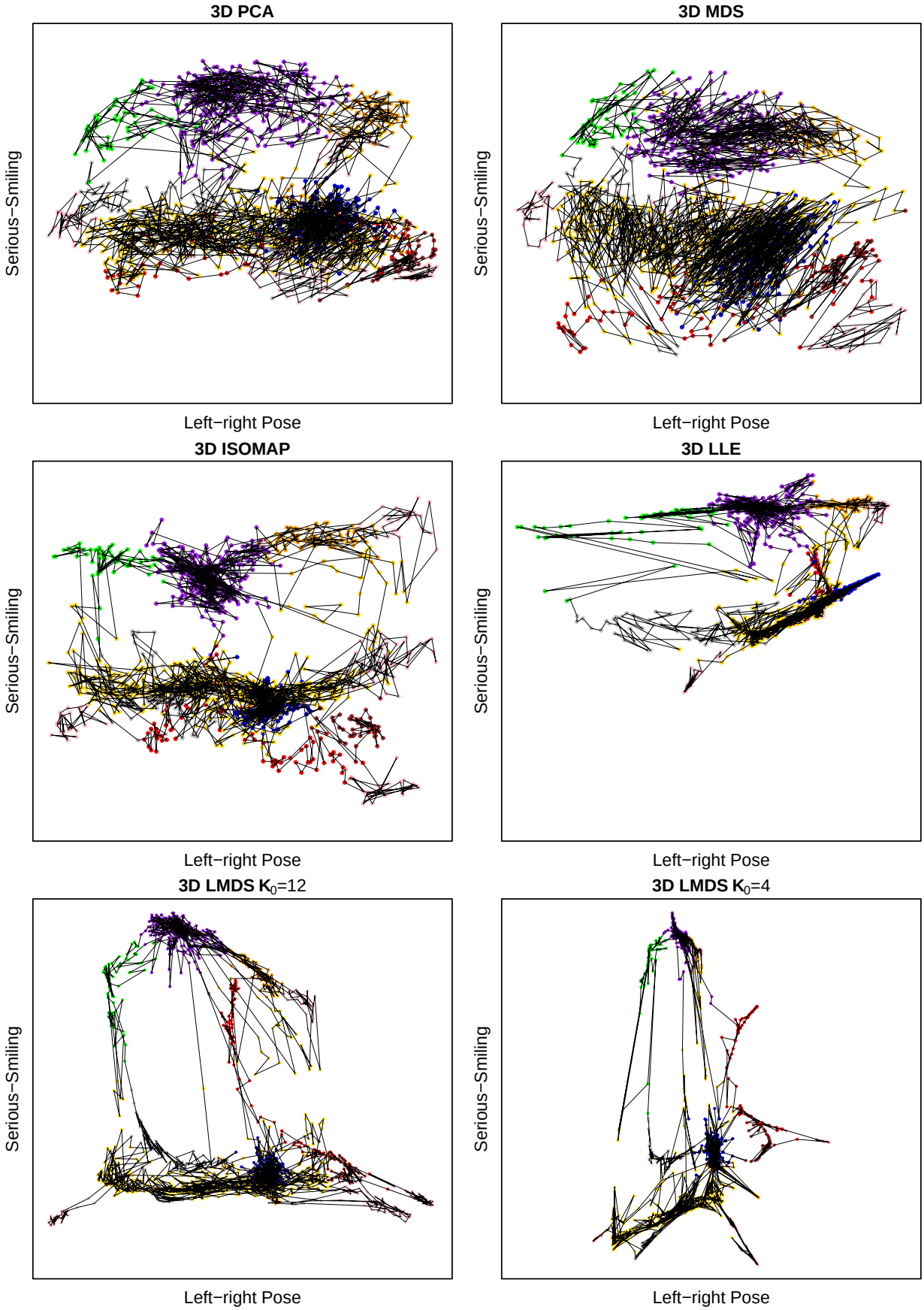

Figure 12: *Frey Face Data: The six configurations with connecting lines that reflect the time order of the video footage.*

green cluster above. Another transition is a near-jump between the purple cluster at the top and the gray and yellow striations at the bottom. Furthermore, there exists one transition through the red cluster (which contains the faces with protruding tongue; see Figure 11). The transitions are even more clearly distinguishable in the LMDS configuration for $K_0 = 4$ than for $K_0 = 12$, although the former's spikiness would have raised suspicions that it suffers from the same problem as the LLE configurations. The connecting lines, however, show that the spikes describe indeed paths taken by the video footage. This is most clearly visible in the red cluster which is beautifully articulated in a way that guides the paths of the video footage. The main bias of the LMDS configuration for $K_0 = 4$ is most likely in the extent to which it attempts to separate the serious and smiling faces; the "true" separations are most likely better reflected in the cruder configurations from PCA, MDS and Isomap.

PCA	MDS	Isomap	LLE	LMDS_{12}	LMDS_4
3.6	4.8	4.2	3.2	4.6	5.1

Table 2: *Frey Face Data: The LC meta-criterion N_{12} for five methods.*

We now compare the five methods according to the LC meta-criterion. The values of N_{12} for the six methods are shown in Table 2. Interestingly LMDS_4 is at the top, but LMDS_{12} is dominated by regular MDS. Of course, from the point of view of wealth of structure revealed, MDS is not a serious competitor of LMDS_{12}. We have here an indication that the LC meta-criterion should not be taken as the sole arbiter of quality of configurations; as important or even more so is the informal degree to which configurations are interpretable. Then again, as biased as the LMDS_4 configuration appears with its vertical elongation, it is the most congenial for the video path, and the meta-criterion appropriately singles it out.

We next consider M_K-traces for τ-optimized LMDS configurations for various choices of K_0 of the neighborhood graph in $\text{LMDS}^D_{K_0}$. Such traces, both raw and adjusted, are shown in Figure 13. The unadjusted traces in the right hand plot show a qualitatively different behavior from similar traces for the sculpture face data. All raw traces are ascending, hence no realistic maximization is possible. Even adjustment for random overlap is insufficient as it barely affects the traces in the range of K's shown here (4 to 150, out of $N - 1 = 1964$). We therefore chose a more drastic adjustment, with MDS as the baseline, as shown in the right hand plot of Figure 13, where the horizontal line at level zero marks MDS. It turns out that $K_0 = 4$ performs best in terms of MDS-adjusted M^{adj}_K for K up to 8, whereas $K_0 = 8$ and $K_0 = 12$ perform quite badly,

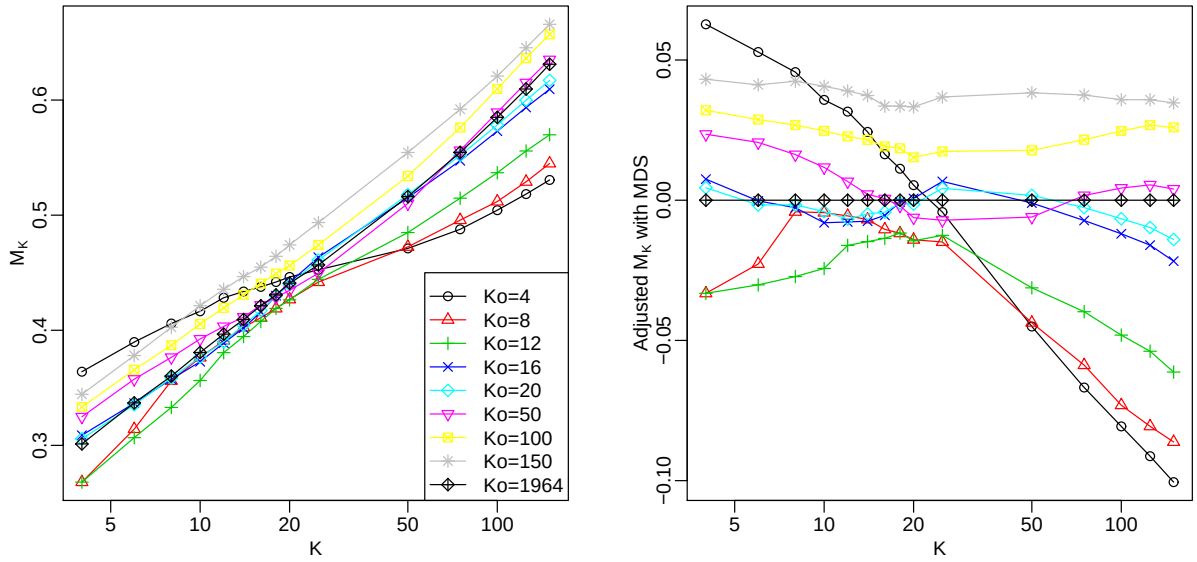

Figure 13: *Frey Face Data: Traces of the LC meta-criterion M_K for various choices of K_0 in the local stress* $\mathrm{LMDS}^D_{K_0}$.

being beaten even by MDS. Again, this contrasts with the qualitative wealth of the LMDS^D_{12} configuration compared to the plain MDS configuration. We note, however, that the overall level of the raw M_K is low for small K, in the order of 0.25 to 0.40.

At this point one may wonder whether averaging of the pointwise values $N_K(i)$ may be concealing local uneveness in the quality of the configurations. We therefore plot the configurations with color codes that show brackets of $N_{12}(i)$ values, as in Figure 14. It appears that in terms of this criterion the weak spots in the configuration for $K_0 = 4$ are in the centers of the heavy clusters, mostly of the serious faces at the bottom (red color). By contrast, some of the transitions and extreme spikes show excellent quality (blue color). Thus, in contrast to the LLE configuration, the spikes in the LMDS^D_4 configuration are real. One may now wonder whether increasing the dimension might allow LMDS to recover further detail inside the heavy clusters which seem to suffer from congestion, possibly caused by a lack of sufficient dimensions.

We conclude by mentioning a standard diagnostic in MDS: the so-called Shepard plot, which is a scatterplot of the configuration distances $d_{i,j}$ against the configuration distances $d_{i,j}$. As an illustration we show a sub-sampled version of the Shepard plot of the LMDS configuration for $K_0 = 12$ in Figure 15. Sub-sampling is necessary because the number of pairs of points is nearly two million, which would result in serious overplotting even with the smallest of point glyphs. Note that the Shepard plot uses the reverse convention of the statistical "whole model plot" where the response

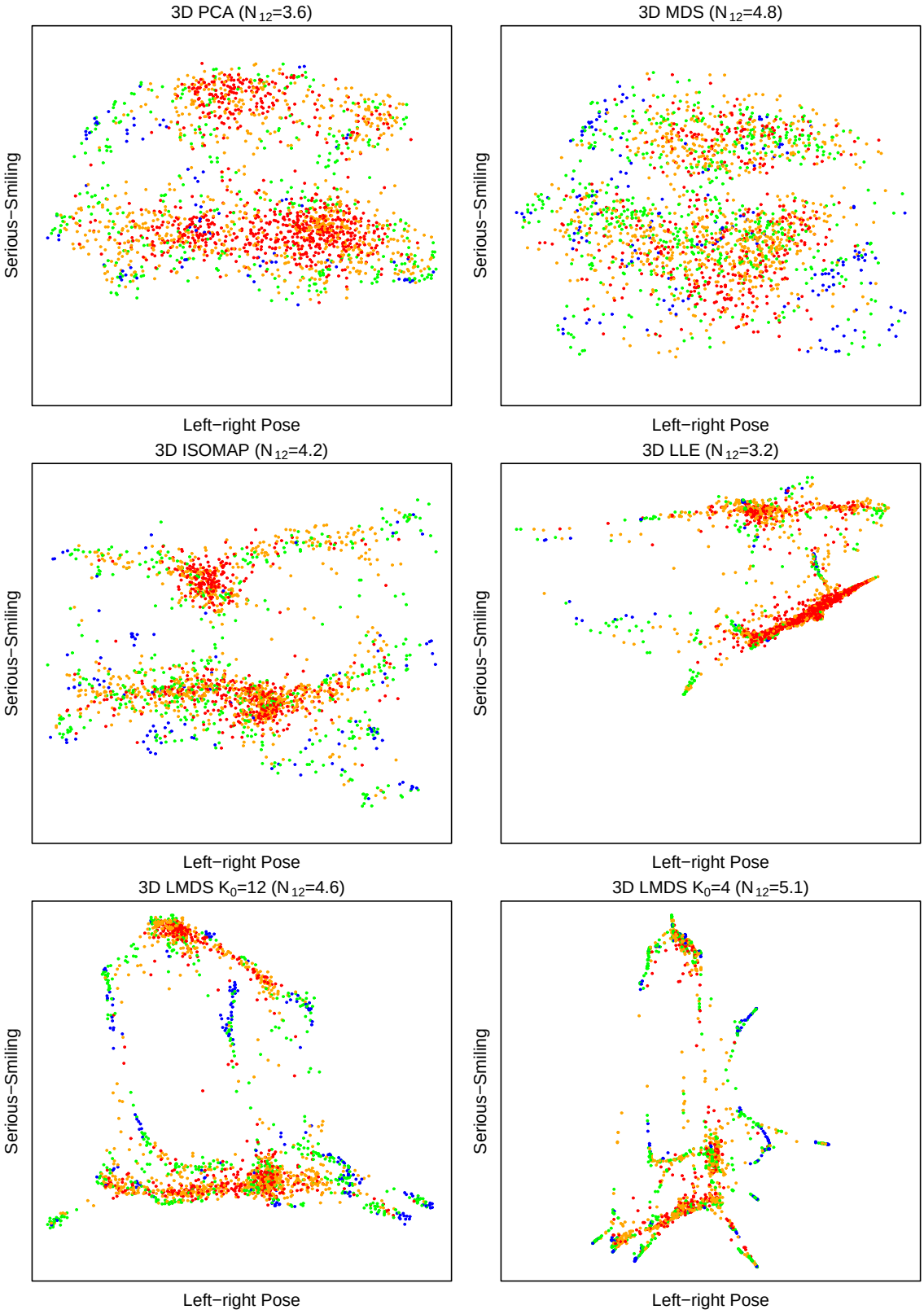

Figure 14: *Frey Face Data: The pointwise LC meta-criterion $N_{12}(i)$ as a diagnostic. Color coding: red $\sim$ 0 to 3; orange $\sim$ 4 to 6; green $\sim$ 7 to 9; blue $\sim$ 10 to 12. The most faithfully rendered parts are blue.*

values are on the vertical axis and the predictions on the horizontal axis; $d_{i,j}$ is the prediction, and $D_{i,j}$ is the response. Just the same, the Shepard plot provides a qualitative assessment of the goodness of fit. In MDS, we look for a roughly linear relation between $d_{i,j}$ and $D_{i,j}$. In local MDS, however, we must allow deviations from linearity due to the repulsion among non-neighboring points. Just the same, in light of Figure 15 we may wonder whether a log transformation of the dissimilarities $D_{i,j}$ wouldn't have been beneficial. On the other hand, this is not an MDS but an LMDS configuration for which one expects that the unfolding of the high-dimensional shape into three dimensions causes some tearing of chordal distances, reminiscent of the "Swiss Roll" whose intrinsic large distances are also much larger than the chordal distances. In light of these considerations, a convex and monotone increasing trend should exactly be the expected pattern.

The only anomalous feature in Figure 15 is then the oval cluster above the plot center. In order to gain some insight into the nature of this cluster, we identified the most extreme cases by placing a red rectangle in the top left. As each point in the plot corresponds to a pair of points in the configuration, we produced another view of the configuration in Figure 16 but with a subset of these point pairs connected by lines (again, an overplotting problem with the full 1500 lines forced us to show only a subsample of 200 lines). From these lines it becomes apparent that the oval cluster in the Shepard plot corresponds to distances between the two main features of all configurations: the two clusters for the smiling (top) and the serious (bottom) faces. Clearly, the LMDS configuration places these two main clusters too far apart compared to their chordal distances in image space. As suggested earlier, the PCA and MDS configurations give a more faithful view of the true distance between smiling and serious faces, but this prevents both methods from showing local detail when flattening the structure into a low-dimensional configuration.

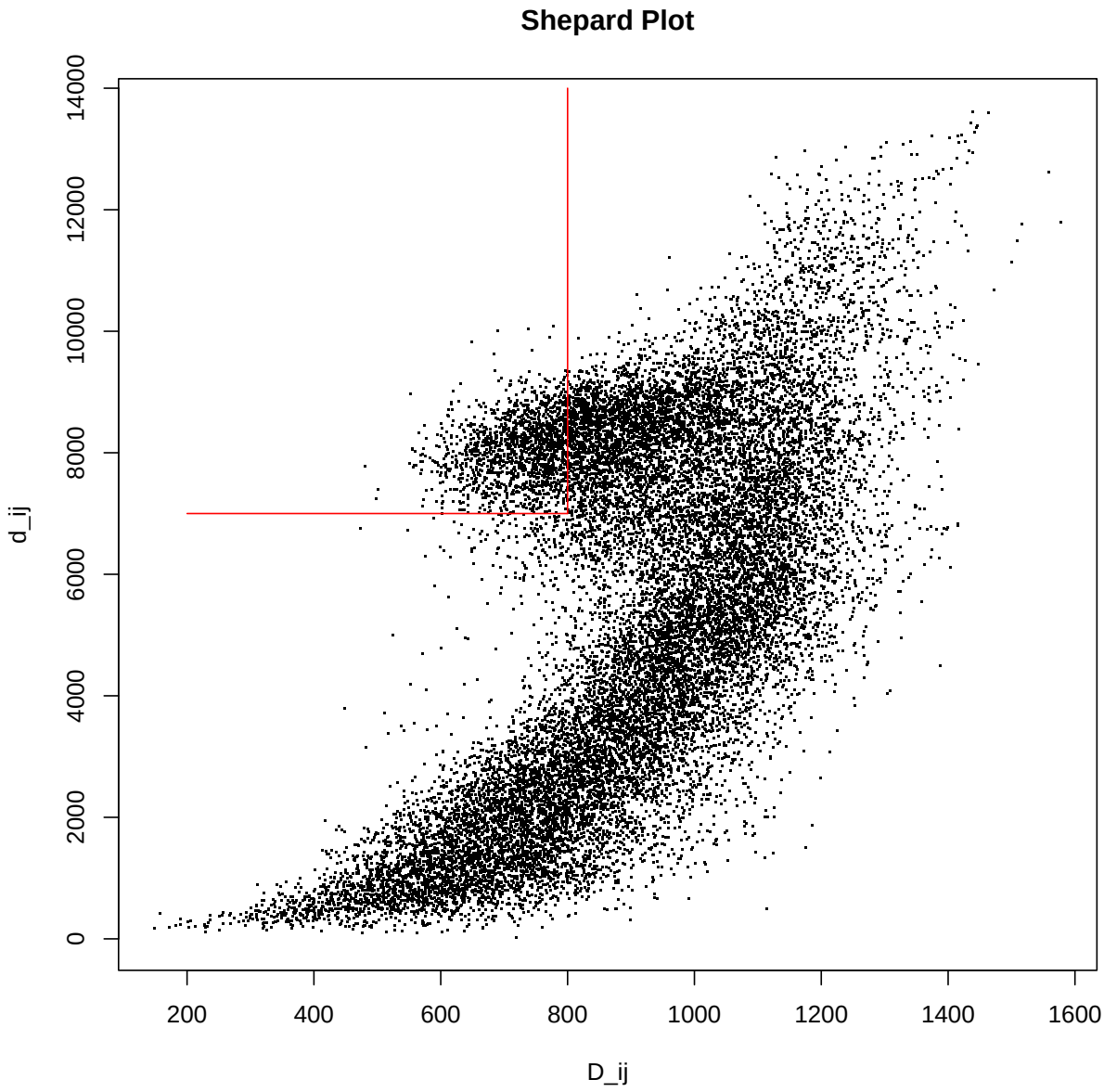

Figure 15: *Frey Face Data: Diagnostic plot for 200,000 pairs of d_{ij} vs D_{ij}, sub-sampled from a total of about 2 million pairs. The red rectangle in the top left contains point pairs that have small dissimilarities but fairly large distances in the configuration.*

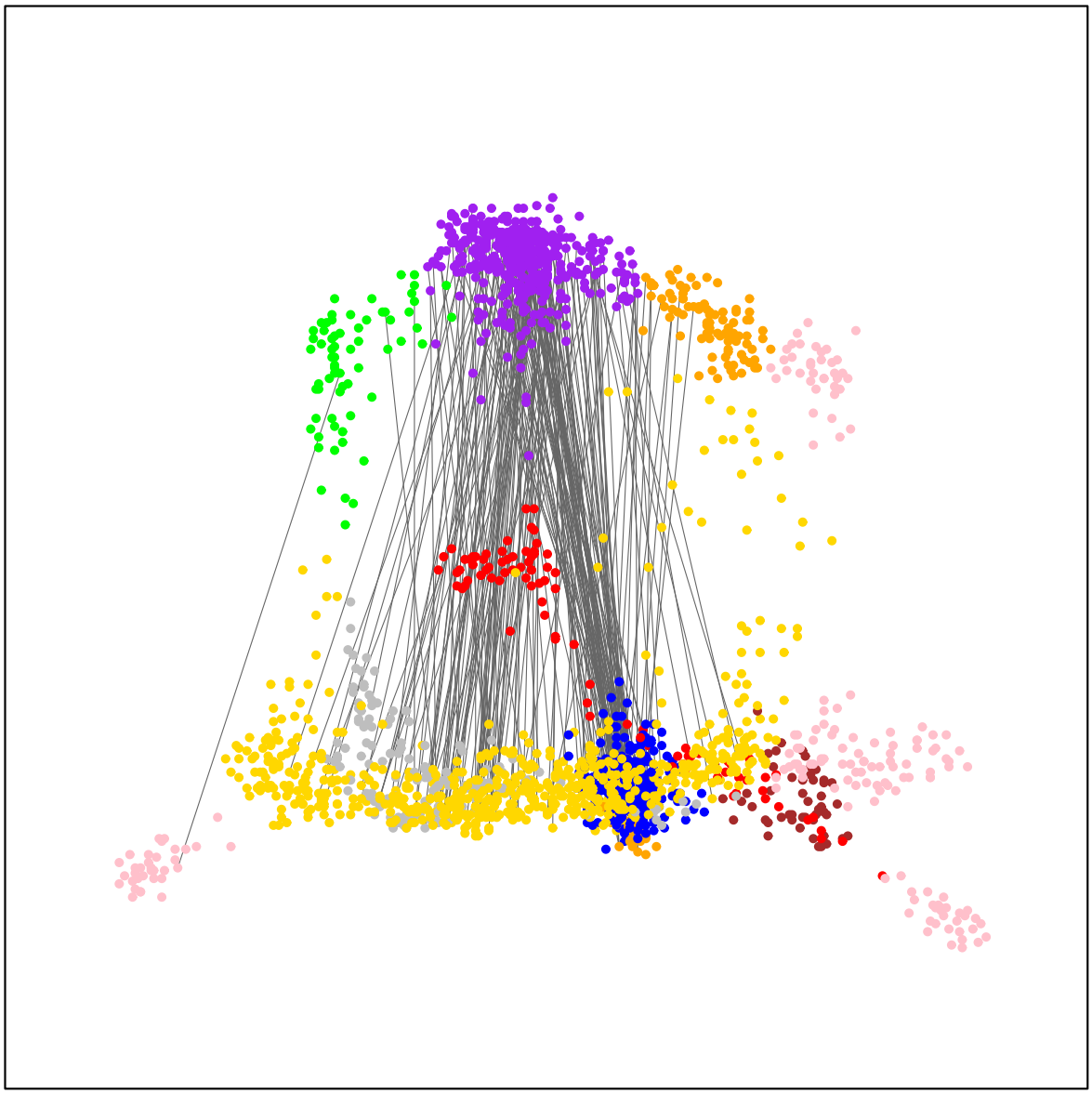

Figure 16: *Frey Face Data: The lines are linking a subset of 200 point pairs in the red rectangle of Figure 15.*

5 ENERGY FUNCTIONS BASED ON BOX-COX TRANSFORMATIONS

In this section, we introduce the energy model for graph layout and propose a large parametrized family of energy functions that comprises not only LMDS but also many energy functions previously proposed in the graph layout literature. A major theme is the "clustering postulate" proposed by Noack (2003) for determining the strength with which graph layouts respond to clustering in the form of loosely connected near-cliques. Noack intended his postulate as a desideratum that all good graph layout techniques should satisfy. We soften up his postulate by embedding it in a parametrized family of postulates where the parameter is an objective measure of clustering strength. This family of postulates in turn motivates a family of energy functions that are basically power laws but best expressed with Box-Cox transformations to include logarithmic energy functions.

5.1 Graph Layout with Energy Functions

Let $G = (V, E)$ be a graph consisting of vertices V and edges E. The goal of graph layout is to draw a low-dimensional picture of points (vertices) and lines (edges) to best visualize the relational information in a graph.

Although the present narrative reads like a piece of work in pure graph layout, it should be kept in mind that the application to nonlinear dimension reduction and proximity analysis is immediate once the energy function framework is extended to distance-weighted graphs. In either of these applications the graph is a symmetrized K-nearest neighbor graph $\mathcal{N}$ and hence a special case of the general graph layout problem.

A widely used method for graph layout is force-directed placement. The idea can be best explained with Eades' metaphor (Eades 1984):

> To embed [lay out] a graph we replace the vertices by steel rings and replace each edge with a spring to form a mechanical system... The vertices are placed in some initial layout and let go so that the spring forces on the rings move the system to a minimal energy state.

Based on the heuristic of such physical systems, investigators in graph layout have proposed a fair number of algorithms. In Table 3 we list a few popular examples. Perusing this table, it is not a priori clear what the advantage would be of one proposal over the others. The last example was recently proposed by Noack based on his "clustering

postulate," which will be the starting point for our own proposals.

Eades (1984)	$\mathrm{U} = \sum_E d_{i,j} \cdot \left(\log(d_{i,j}) - 1 - {d_{i,j}}^{-2}\right) + \sum_{V^2} {d_{i,j}}^{-1}$
Fruchterman and Reingold (1991)	$\mathrm{U} = \sum_E {d_{i,j}}^3/3 - \sum_{V^2} \log(d_{i,j})$
Davidson and Harel (1996)	$\mathrm{U} = \sum_E {d_{i,j}}^2 + \sum_{V^2} {d_{i,j}}^{-2}$
Noack (2003)	$\mathrm{U} = \sum_E d_{i,j} - \sum_{V^2} \log(d_{i,j})$

Table 3: *Some well-known energy functions* U *for graph layout.*

We note that each energy function U is a function $\mathrm{U} = \mathrm{U}(G, X)$ of a graph G and a configuration $X = (\mathbf{x}_i)_{i=1\ldots N}$. Minimization of an energy function over all configurations produces the best layout as measured by the particular energy function. A commonality of the examples listed in Table 3 is the assumption that attractive forces exist among the vertices connected by edges, while repulsive forces exist between *all* vertices. Energy functions are written in a common form where the first term corresponds to the energy from attraction and the second term to the energy from repulsion:

$$\mathrm{U} = \sum_{(i,j)\in E} f(d_{i,j}) - \sum_{(i,j)\in V^2} g(d_{i,j}) \tag{13}$$

The symbol V^2 is shorthand for the full graph. Both $f(d)$ and $g(d)$ are monotone increasing to do justice to the terms "attraction" and "repulsion": as $d_{i,j}$ decreases, the attraction term decreases, and as $d_{i,j}$ increases, the repulsion term decreases. The idea is once again to balance graph structure implicit in the attractive force with a tendency to spread out implicit in the repulsive force. The above form of energy function is for unweighted graphs and therefore does not contain MDS or LMDS as special cases. To this end, one will have to allow the attraction terms $f()$ to be functions of dissimilarities $D_{i,j}$ as well.

5.2 Noack's Clustering Postulate

We now discuss a clustering postulate proposed by Noack (2003). This postulate derives from the consideration of an idealized two-cluster situation that is so simple that it lends itself to analytical treatment, yet is sufficiently general that it lends insight into the response of energy functions to clusters in graphs. Noack laid down his postulate as a precise requirement on how strongly energy functions should attempt to separate clusters. We find his requirement too rigid, however, which is why we soften it up

to describe an infinite range of responses to clustering. In essence, we propose a one-parameter power family of postulates where the parameter is a measure of clustering strength.

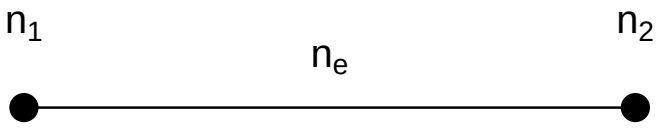

Figure 17: *Noack's simple graph with two subgraphs that have n_1 and n_2 internal vertices and n_e connecting edges.*

In the details, this works out as follows: According to Noack (2003), we consider an idealized graph consisting of two subgraphs, to be called "clusters." To justify this term, we assume the vertices highly connected within each cluster and loosely connected between them. As depicted in Figure 17, we assume n_1 and n_2 vertices in the two graphs and n_e edges between them. To simplify the model, we assume all the vertices of one cluster are drawn at the same position. As we are only interested in the distance between clusters, we omit the energy within clusters and only consider the energy between them. The relevant energy then simplifies to the following form, where d is the distance between the two clusters:

$$U(d) \;=\; n_e f(d) - n_1 n_2 \; g(d)$$

The minimal energy distance $d = d_{\min}$ is achieved when

$$U'(d) \;=\; n_e f'(d) - n_1 n_2 \; g'(d) \;=\; 0 \;, \qquad \text{that is,} \;\; f'(d) \;=\; \frac{n_1 n_2}{n_e} \; g'(d) \;. \tag{14}$$

At this point Noack introduces the concept of "coupling strength" to measure how strongly the two clusters are connected. He defines coupling strength C as the ratio of the number of existing edges to the number of all possible edges between the clusters. For the graph of Figure 17 this is

$$C \;=\; \frac{n_e}{n_1 n_2} \;.$$

This quantity is always between 0 and 1, with larger C implying stronger coupling strength. The minimum energy condition becomes

$$f'(d) \;=\; \frac{1}{C} \; g'(d) \;. \tag{15}$$

This minimum condition hints that there should exist constraints between the attractive and repulsive forces, but because the coupling strength is an unknown quantity, the constraint expressed by (15) is not actionable. Noack's next step is to link the minimizing distance d and the coupling strength C in his coupling postulate, thereby creating an actionable constraint between $f()$ and $g()$. He proposes the

Weak Clustering Postulate: *More strongly coupled clusters should be placed closer to each other.*

This implies that the energy-minimizing distance d in (15) should be a monotone descreasing function of the coupling strength C. He required more specifically that the minimizing distance be the reciprocal of C:

Strong Clustering Postulate: $\quad d \;=\; \frac{1}{C}$

As this should hold for all values $C \leq 1$ and > 0, the minimal energy condition (15) results in the following constraint between attractive and repulsive forces:

Strong Clustering Constraint:

$$f'(x) \;=\; x \cdot g'(x) \;\; \forall x \geq 1 \tag{16}$$

This is a powerful condition that reduces the arbitrariness in the choice of energy functions: if either the attractive or the repulsive force is given, the other force is essentially determined by (16).

Noack applies the constraint to two special cases: $g(x) = ln(x)$ and $g(x) = x$. The first case results in what he calls the *LinLog* energy:

$$U_{LinLog} \;=\; \sum_{(i,j)\in E} d_{i,j} - \sum_{(i,j)\in V^2} \ln d_{i,j} \;,$$

the second case in what he calls the *QuadLin* energy:

$$U_{QuadLin} \;=\; \sum_{(i,j)\in E} \frac{1}{2}\, {d_{i,j}}^2 - \sum_{(i,j)\in V^2} d_{i,j} \;.$$

This second example is reminiscent of the LMDS stress criterion, but it lacks distance-weighting. Motivated by the LinLog energy, Noack goes one step further and considers a one-parameter family of "PolyLog" energy functions:

$$U_{PolyLog} \;=\; \sum_{(i,j)\in E} {d_{i,j}}^r - \sum_{(i,j)\in V^2} \ln d_{i,j} \;.$$

where $r \geq 1$. He notes that in the idealized two-cluster situation they attain their minimum energy distance at

$$d \;=\; \left(\frac{1}{C}\right)^{1/r} \;,$$

thereby allowing for weaker clustering strengths for increasing r.

5.3 The B-C Family of Energy Functions for Unweighted Graphs

We take Noack's work as the starting point for an extended clustering postulate by requiring the distance between two clusters to be a power function of the inverse coupling strength:

Power Clustering Postulate: $\quad d \;=\; \left(\frac{1}{C}\right)^{\lambda} \qquad (\lambda > 0)$

As before, a constraint between attractive and repulsive forces results. The minimum energy condition (15) specializes to the following:

Power Clustering Constraint:

$$f'(x) \;=\; x^{\frac{1}{\lambda}} \cdot g'(x) \quad \forall x \geq 1 \tag{17}$$

The parameter λ is a measure of the strength of the clustering effect: the larger λ, the more loosely coupled clusters will be separated from each other. As λ goes to zero, the clustering effect tends to zero. We therefore call λ the "clustering power."

In Table 3, the energy functions by Fruchterman and Reingold (1991) and Davidson and Harel (1996) both satisfy the power clustering constraint with the clustering powers $\lambda = \frac{1}{3}$ and $\frac{1}{4}$, respectively. This fact explains why these energy functions have a weaker clustering effect than the LinLog and QuadLin energy functions, which both have the clustering power $\lambda = 1$ (compare Noack 2003, Table 3)

We now introduce a power family of energy functions, but because we wish to include logarithmic functions we use Box-Cox transformations instead of raw powers. A secondary benefit of Box-Cox transformations is that even negative powers are monotone increasing and hence qualify as attractive and repulsive forces. See Figure 18 for an illustration of Box-Cox transformations.

We also assume a power clustering constraint with a given clustering power λ. As a result, if either $f()$ or $g()$ is of the Box-Cox variety, the other will be, too. Thus, denoting the Box-Cox power by μ, we obtain the following matched pairs of repulsive and attractive energies:

$$g(d) \;=\; \begin{cases} \frac{d^{\mu}-1}{\mu} & (\mu \neq 0) \\ \ln(d) & (\mu = 0) \end{cases} \tag{18}$$

$$f(d) \;=\; \begin{cases} \frac{d^{\mu+\frac{1}{\lambda}}-1}{\frac{1}{\lambda}+\mu} & (\mu + \frac{1}{\lambda} \neq 0) \\ \ln(d) & (\mu + \frac{1}{\lambda} = 0) \end{cases} \tag{19}$$

Thus μ is the power law of repulsion, $\mu + \frac{1}{\lambda}$ the power law of attraction, while λ is the

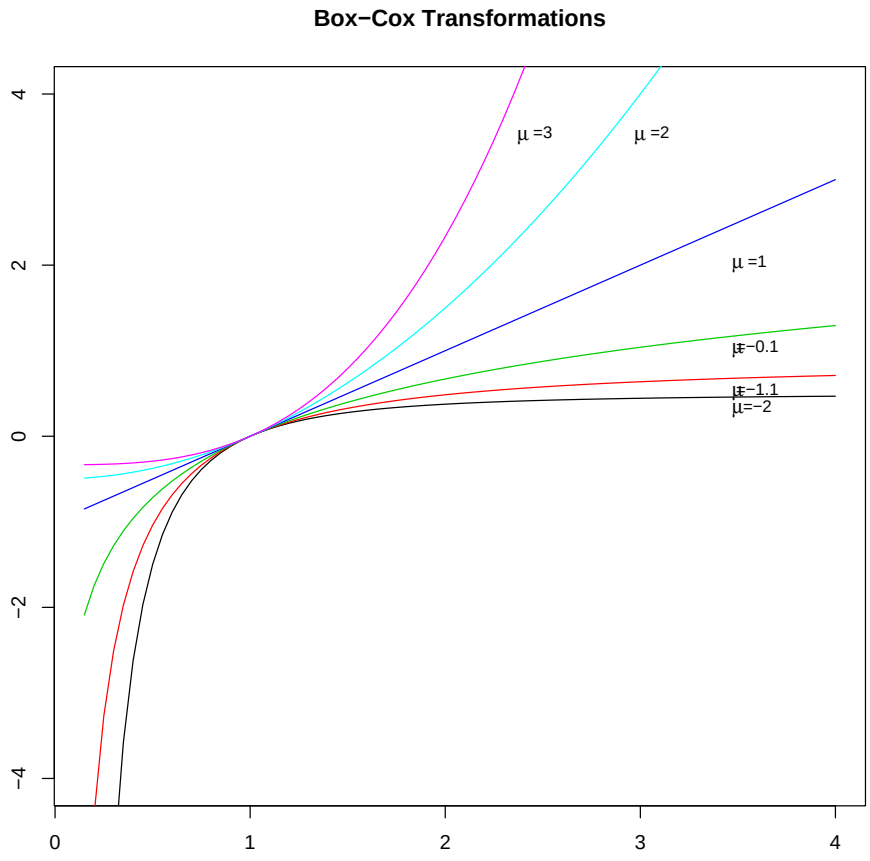

Figure 18: *Box-Cox Transformations:* $y = \frac{x^\mu - 1}{\mu}$

clustering power. The full energy function is therefore:

$$\mathrm{U} \;=\; \sum_{(i,j)\in E} \frac{{d_{i,j}}^{\mu+\frac{1}{\lambda}} - 1}{\mu + \frac{1}{\lambda}} \;-\; \sum_{(i,j)\in V^2} \frac{{d_{i,j}}^{\mu} - 1}{\mu} \;, \tag{20}$$

with logarithmic fill-in when either μ or $\mu + 1/\lambda$ is zero. This is the preliminary form of what we call the **B-C family of energy functions**. It includes the LinLog and QuadLin energy functions as special cases for $\lambda = 1$, $\mu = 0$ and $\lambda = 1$, $\mu = 1$, respectively.

We restate that the Box-Cox power μ is unconstrained in principle: $f()$ and $g()$ are increasing for all $\mu \in \mathbb{R}$. If physical analogies were relevant, though, one would wonder about repulsive forces that get stronger as the distance gets larger. The Coulomb force between equal charges, for example, has energy potential $1/d$ and repulsive force $-1/d^2$, which decays with greater distance. Energy power laws $g(d) \sim d^\mu$ with decaying force are those with monotone decreasing $g'(d)$, that is, $\mu < 1$. The QuadLin energy is therefore not physically motivated as it is a limiting case with a repulsive force that is constant irrespective of distance.

Just the same, we are currently exploring the counter-physical case of a repulsive power energy with $\mu = 2$ because it would be attractive from a computational point of view. Its total repulsive energy is $\sum_{i<j} {d_{i,j}}^2 = 2N \sum_i \|\mathbf{x}_i\|^2$. This algebraic identity implies that the force can be computed at a cost proportional to N as opposed to $N(N-1)/2$. The total cost of force computations is therefore no longer driven by the N^2 cost of the repulsive force but by the cost of the attractive force, which is proportional to the size of the graph. Thus, if the repulsive energy for $\mu = 2$ produce viable configurations, the computational savings would be considerable for sparse graphs.

5.4 The B-C Family of Energy Functions for Distance-Weighted Graphs

The graphs we have considered so far have uniform edge length 1. For dimension reduction and proximity analysis, however, we need distance weighted graphs, where the edges have lengths given by dissimilarities $D_{i,j}$. A natural way to extend the clustering postulate to distance-weighted graphs is to require that in Noack's two-cluster paradigm the minimum energy distance between clusters be achieved at the target dissimilarity D times the clustering factor $(\frac{1}{C})^\lambda$:

Power Clustering Postulate with Distance Weighting:

$$d \;=\; D \cdot \left(\frac{1}{C}\right)^{\lambda} \tag{21}$$

We note that in the special case of two connected points (each cluster containing just one point), we have $n_e = n_1 = n_2 = 1$ and hence $C = 1$. Therefore, the energy-minimizing distance is $d = D$. In other words, the energy balance between connected vertices is such that two connected points i and j attempt to be placed at a distance $d_{i,j} = D_{i,j}$ from each other.

Combined with the minimum energy condition (15) the new postulate (21) results in the following constraint between attraction and repulsion:

Power Clustering Constraint with Distance Weighting:

$$f'(d) \;=\; \left(\frac{d}{D}\right)^{\frac{1}{\lambda}} \; g'(d) \tag{22}$$

We can again prescribe $g(d)$ and obtain $f(d)$ from (22). The proposal for $g(d)$ we make

here is basically a (18), but multiplied with a power of the target D:

$$g(d) \;=\; D^{\nu} \begin{cases} \frac{d^{\mu}-1}{\mu} & (\mu \neq 0) \\ \ln(d) & (\mu = 0) \end{cases} \tag{23}$$

$$f(d) \;=\; D^{\nu-\frac{1}{\lambda}} \begin{cases} \frac{d^{\mu+\frac{1}{\lambda}}-1}{\frac{1}{\lambda}+\mu} & (\mu+\frac{1}{\lambda} \neq 0) \\ \ln(d) & (\mu+\frac{1}{\lambda} = 0) \end{cases} \tag{24}$$

The immediate problem with this attempt is that it produces only terms that apply to edges of the graph for which a target distance D exists. We therefore arrive only at an energy function for the full graph, that is, at a stress function:

$$\mathrm{U} \;=\; \sum_{(i,j)\in V^2} {D_{i,j}}^{\nu} \left({D_{i,j}}^{-\frac{1}{\lambda}} \, \frac{{d_{i,j}}^{\mu+\frac{1}{\lambda}}-1}{\mu+\frac{1}{\lambda}} \;-\; \frac{{d_{i,j}}^{\mu}-1}{\mu} \right) \tag{25}$$

We call this the *B-C family of stress functions*. (As always, we assume logarithmic fill-in for $\mu = 0$ and $\mu + 1/\lambda = 0$.) In order to arrive at energy functions for non-full graphs, we avail ourselves of the localization device that produced LMDS from MDS. We replace again the dissimilarities $D_{i,j}$ for $(i,j) \notin E$ with a single large dissimilarity D_{∞} that goes to infinity. We downweight these terms with a weight w in such a way that ${w D_{\infty}}^{\nu} = t$ is constant:

$$\begin{aligned} \mathrm{U} \;=\; & \sum_{(i,j)\in E} {D_{i,j}}^{\nu} \left({D_{i,j}}^{-\frac{1}{\lambda}} \, \frac{{d_{i,j}}^{\mu+\frac{1}{\lambda}}-1}{\mu+\frac{1}{\lambda}} \;-\; \frac{{d_{i,j}}^{\mu}-1}{\mu} \right) \\ +\; & w \sum_{(i,j)\notin E} {D_{\infty}}^{\nu} \left({D_{\infty}}^{-\frac{1}{\lambda}} \, \frac{{d_{i,j}}^{\mu+\frac{1}{\lambda}}-1}{\mu+\frac{1}{\lambda}} \;-\; \frac{{d_{i,j}}^{\mu}-1}{\mu} \right) \end{aligned}$$

As $D_{\infty} \to \infty$, we have ${D_{\infty}}^{-1/\lambda} \to 0$, hence in the limit we obtain what we now call the *B-C family of energy functions:*

$$\mathrm{U} \;=\; \sum_{(i,j)\in E} {D_{i,j}}^{\nu} \left({D_{i,j}}^{-\frac{1}{\lambda}} \, \frac{{d_{i,j}}^{\mu+\frac{1}{\lambda}}-1}{\mu+\frac{1}{\lambda}} \;-\; \frac{{d_{i,j}}^{\mu}-1}{\mu} \right) \;-\; t \sum_{(i,j)\notin E} \frac{{d_{i,j}}^{\mu}-1}{\mu} \tag{26}$$

(Again, we assume logarithmic fill-in for $\mu = 0$ and $\mu + 1/\lambda = 0$.) We have thus weighted the terms within the graph by ${D_{i,j}}^{\nu}$ and the terms not in the graph by t. The parameter t plays the same role as in the LMDS stress function for balancing the relative strength of the combined attraction and repulsion inside the graph with the repulsion outside the graph. For completeness, we list the assumed ranges of the parameters:

$$t \geq 0 \;, \quad \lambda > 0 \;, \quad -\infty < \mu < \infty \;, \quad -\infty < \nu < \infty \;.$$

The B-C families contain both MDS and LMDS as special cases: When $\mu = \lambda = \nu = 1$, (25) becomes the Kruskal-Shepard MDS stress function (1) and (26) the LMDS energy function (7). For other choices of parameter values, one obtains several energy functions previously proposed in the graph layout literature. Table 4 gives a summary. Isomap could be listed in the MDS row if Kruskal-Shepard distance scaling were used instead of Tenenbaum's et al. (2000) choice of classical scaling. The table shows that Noack's (2003) QuadLin is the specialization of LMDS to unweighted graphs with unit repulsion weight $t = 1$. The sense in which Noack's (2003) PolyLog power family is a subset of the B-C family will be explained in Subsection 5.5.

Choice of Parameters	Special Cases
$E = V^2$, $\lambda = \mu = \nu = 1$	MDS (Kruskal 1964; Kruskal & Seery 1980)
$E = V^2$, $\lambda = \mu = 1,\ \nu = -1$	Kamada & Kawai (1989); Sammon (1969)
$\lambda = \mu = \nu = 1$	LMDS
$\lambda = 1/3$, $\mu = \nu = 0$, $D_{i,j} = 1$, $t = 1$	Fruchterman & Reingold (1991)
$\lambda = 1/4$, $\mu = -2$, $\nu = 0$, $D_{i,j} = 1$, $t = 1$	Davidson & Harel (1996)
$\lambda = 1$, $\mu = \nu = 0$, $D_{i,j} = 1$, $t = 1$	Noack's LinLog (2003)
$\lambda = \mu = \nu = 1$, $D_{i,j} = 1$, $t = 1$	Noack's QuadLin (2003)
$\lambda > 0$, $\mu = \nu = 0$, $D_{i,j} = 1$, $t = 1$	Noack's PolyLog family (2003; his $r = 1/\lambda$)

Table 4: *Some special cases of energy functions.*

5.5 An Irrelevant Constant: Weighting the Attraction

Noack (2003, Sec. 5.5) observed that for his LinLog energy function the relative weighting of the attractive energy relative to the repulsive energy is irrelevant in the sense that such weighting would only change the scale of the minimizing layout but not the shape. A similar statement can be made for all members of the B-C family of energy functions. To demonstrate this effect, we introduce B-C energies whose attraction is weighted by a factor $a^{\frac{1}{\lambda}}$ where $a > 0$:

$$\mathrm{U}_a(d) \;=\; \sum_{(i,j)\in E} {D_{i,j}}^{\nu} \left(a^{\frac{1}{\lambda}} {D_{i,j}}^{-\frac{1}{\lambda}} \, \frac{{d_{i,j}}^{\mu+\frac{1}{\lambda}} - 1}{\mu + \frac{1}{\lambda}} \;-\; \frac{{d_{i,j}}^{\mu} - 1}{\mu} \right) \;-\; t \sum_{(i,j)\notin E} \frac{{d_{i,j}}^{\mu} - 1}{\mu}$$

We denote by $d = (d_{i,j})$ the set of all configuration distances for all edges (i, j), including those not in the graph E. Note that the repulsion terms are still differentially

weighted depending on whether (i, j) is an edge of the graph E or not, which is in contrast to most energy functions proposed in the graph layout literature where invariably $t = 1$. — In analogy to Noack's argument, we observe the following form of scale invariance:

$$\mathrm{U}_1(a\,d) \;=\; a^{\mu}\,\mathrm{U}_a(d) \;+\; \text{const} \tag{27}$$

This relation holds also for $\mu = 0$ (logarithmic repulsion) and $\mu + 1/\lambda = 0$ (logarithmic attraction). As a consequence, if d is a minimizing set of configuration distances for $\mathrm{U}_a(\cdot)$, then the distances $a\,d$ of the configuration scaled by a factor a minimize the original unweighted B-C energy $\mathrm{U}_1(\cdot)$. This makes sense as an example shows: for $\lambda = 1$ and $a = 2$, a set of configuration distances d is minimal for $U_a(\cdot)$ iff $2d$ is minimal for $U_1(\cdot)$, that is, the increased attraction of $U_a(\cdot)$ shrinks the solution by a factor of $a = 2$ compared to $U_1(\cdot)$.

It is in this sense that Noack's PolyLog family of energies can be considered as a special case of the B-C family: PolyLog energies agree with B-C energies for $\mu = \nu = 0$ and $t = 1$ up to a multiplicative factor of the attractive force.

5.6 Scale Invariant Forms of the Repulsion Weight

The remarks of Section 2.3 on invariance requirements on t apply here as well, and possible reparametrizations look similar:

$$t \;=\; \frac{|E|}{|V^2| - |E|} \cdot \left(\mathrm{median}_{(i,j)\in E}\; D_{i,j}\right)^{\nu} \cdot \tau \;, \tag{28}$$

where τ is now a unit free tuning parameter. We have made a modification by replacing K-nearest neighbor graphs with general graphs with edge sets E. In order to balance the different numbers of terms in the two sums of (26), we reweight the second term by a factor $|E|/(|V^2| - |E|)$ where the numerator and denominator are the exact counts of terms in the first and the second sum, respectively. For K-nearest neighbor graphs, this ratio is approximately K/N when K is small compared to N, reducing to (8).

The median of the values $D_{i,j}$ can obviously be replaced by any other symmetric first order homogeneous function of the values, including the mean and any order statistic.

The repulsion weight τ (or t) plays a special role in the optimization of the B-C energy functions because larger values produce more stable minima. Therefore, as earlier with LMDS, we start minimization with larger values of τ to spread out the vertices and achieve stability, and then we lower τ to obtain more truthfulness to the local structure.

5.7 Clustering Properties of the B-C Family

A consequence of the manipulations of Subsection 5.4 is that we violated the power clustering postulate. We can, however, examine how the B-C family of energy functions performs in Noack's two-cluster paradigm, and this will lead to some insights. The inter-cluster energy for a B-C energy function is:

$$\mathrm{U} \;=\; n_e\, D^{\nu} \left(D^{-1/\lambda}\, \frac{d^{\mu+1/\lambda} - 1}{\mu + 1/\lambda} \;-\; \frac{d^{\mu} - 1}{\mu} \right) \;-\; (n_1 n_2 - n_e)\, t\, \frac{d^{\mu} - 1}{\mu}$$

Given a target dissimilarity D, the energy-minimizing distance d satisfies the following stationarity condition:

$$0 \;=\; n_e\, D^{\nu} \left(D^{-1/\lambda}\, d^{\mu+1/\lambda-1} \;-\; d^{\mu-1} \right) \;-\; (n_1 n_2 - n_e)\, t\, d^{\mu-1}$$

This can be solved for d:

$$d \;=\; D \left(1 + \frac{t}{D^{\nu}} \left(\frac{1}{C} - 1 \right) \right)^{\lambda}$$

As always, $C = n_e/(n_1 n_2)$ is the coupling strength. In spite of the convoluted appearance of this solution, it is basically simple:

- For maximal coupling strength, as for a simple connected pair ($n_1 = n_2 = n_e = 1$), the clusters are placed at a distance that matches the target dissimilarity:
  $C = 1 \;\Rightarrow\; d = D$.
- As the coupling strength decreases, the clusters are placed farther apart:
  $C \searrow 0 \;\Rightarrow\; d \nearrow \infty$, assuming $t > 0$.
- As the repulsion weight increases, the clusters are placed farther apart:
  $t \nearrow \infty \;\Rightarrow\; d \nearrow \infty$, assuming $C < 1$.
- For vanishing repulsion weight, the distance of the clusters matches the target dissimilarity:
  $t = 0 \;\Rightarrow\; d = D$.
- As the clustering power increases, the clusters are placed farther apart:
  $\lambda \nearrow \infty \;\Rightarrow\; d \nearrow \infty$, assuming $C < 1$ and $t > 0$.
- The power law parameter μ does not affect clustering properties.
- Consideration of the weight power ν requires the invariant version τ of the repulsion weight (28) and the assumption that the two-cluster graph is part of a larger graph E with median dissimilarity $m = \mathrm{median}_{(i,j)\in E} D_{i,j}$. Then the effect of ν depends on $(m/D)^{\nu}$. For example:
  $\nu \nearrow \infty \;\Rightarrow\; d \searrow D$ if $D > m$, assuming $C < 0$ and $\tau > 0$.

5.8 Interpretation of the B-C Energy Functions

The previous subsection explains much of the role of the parameters of the B-C energy family. We summarize their roles in general terms as follows.

- λ The clustering power λ adjusts the clustering effect in the graph. The larger λ, the more strength there is for coupling to come into play and hence the stronger the clustering effect is in the graph.
- μ The parameter μ describes the power law of the repulsive energy, and in combination with λ also the power law of the attractive energy. The larger μ is, the more sensitive the energy is to increases in configuration distances $d_{i,j}$.
- ν The term D^ν describes the relative weight with which the edges of the graph contribute to total energy as a function of $D_{i,j}$. In particular, when $\nu > 0$, large dissimilarities want to be matched more strongly; when $\nu < 0$, small dissimilarities want to be matched more strongly.
- t The parameter t adjusts the strength of repulsion outside the graph. The reparametrizaton (28) maps t to a unit free tuning parameter τ whose meaning is invariant under unit changes of the data $D_{i,j}$.

6 B-C ENERGIES APPLIED

6.1 Olivetti Faces Data

This dataset, published on Sam Roweis' website, has facial images of 40 persons. There are 10 images for each of them and a total of 400 images in the dataset (see Figure 19). Each image is of size 64 by 64. Again for visualization purposes we reduce the dimension from 4096 to 3. As the 40 faces form natural groups, we would expect effective dimension reduction methods to show clusters in their configurations. If this expectation is correct, then this dataset provides an excellent test bed for the effect of the clustering power λ.

Figure 20 shows 2D projections of 3D configurations generated by different energy functions (26) for different clustering powers, $\lambda = 0.25,\ 0.5,\ 1,\ 1.5$, while the other parameters are fixed at $\mu = \nu = 1$ and $\tau = 0.01$. We colored the data points according to the last configuration with $\lambda = 1.5$. With small λ, we see fussy clusters and as λ increases the clusters become clearer. These clusters are not artifactual but real, representing the images of the same person. In fact, the configurations produced fewer than the known number of 40 clusters: the big grey cluster in the center lumps

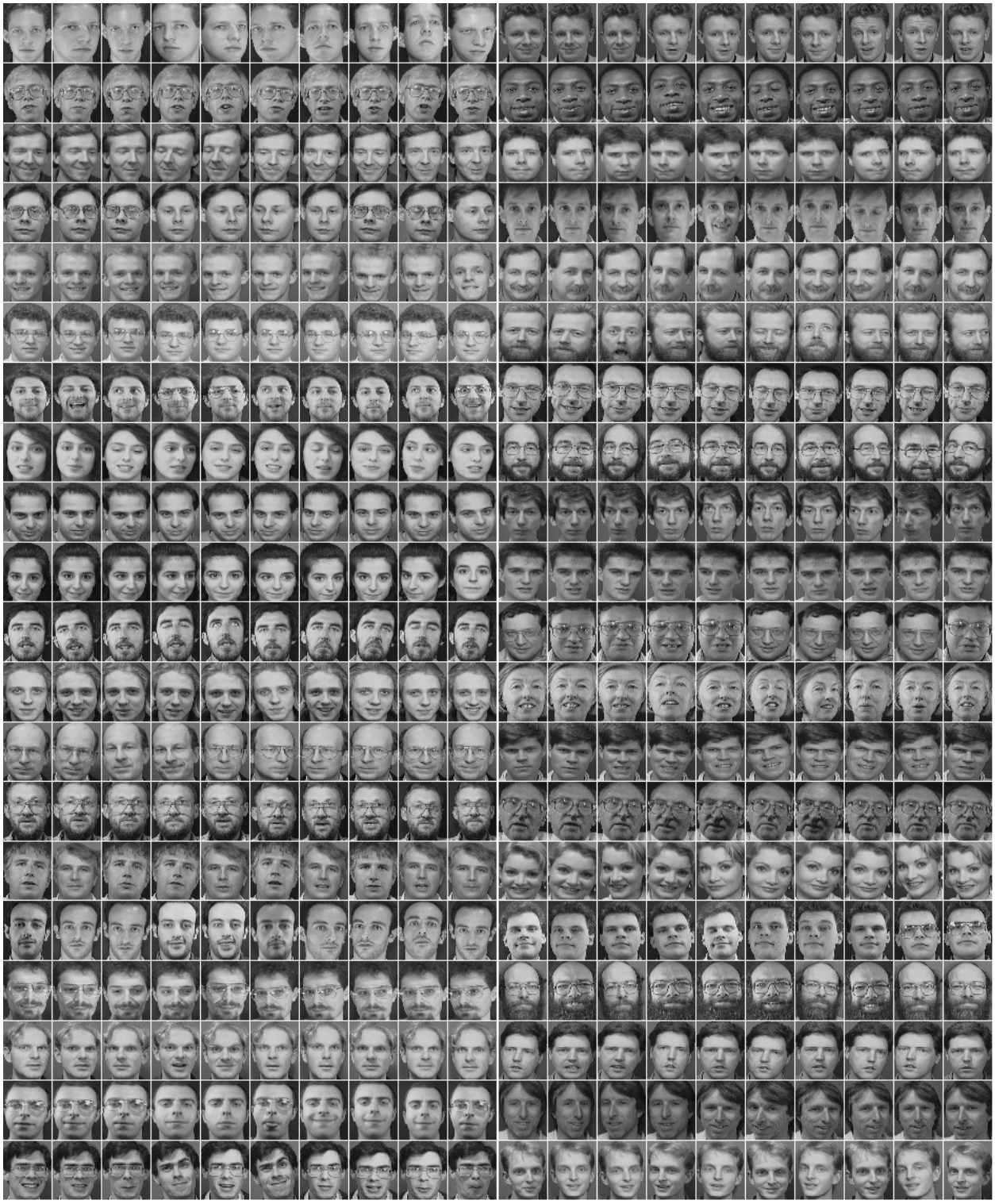

Figure 19: *Olivetti Faces*

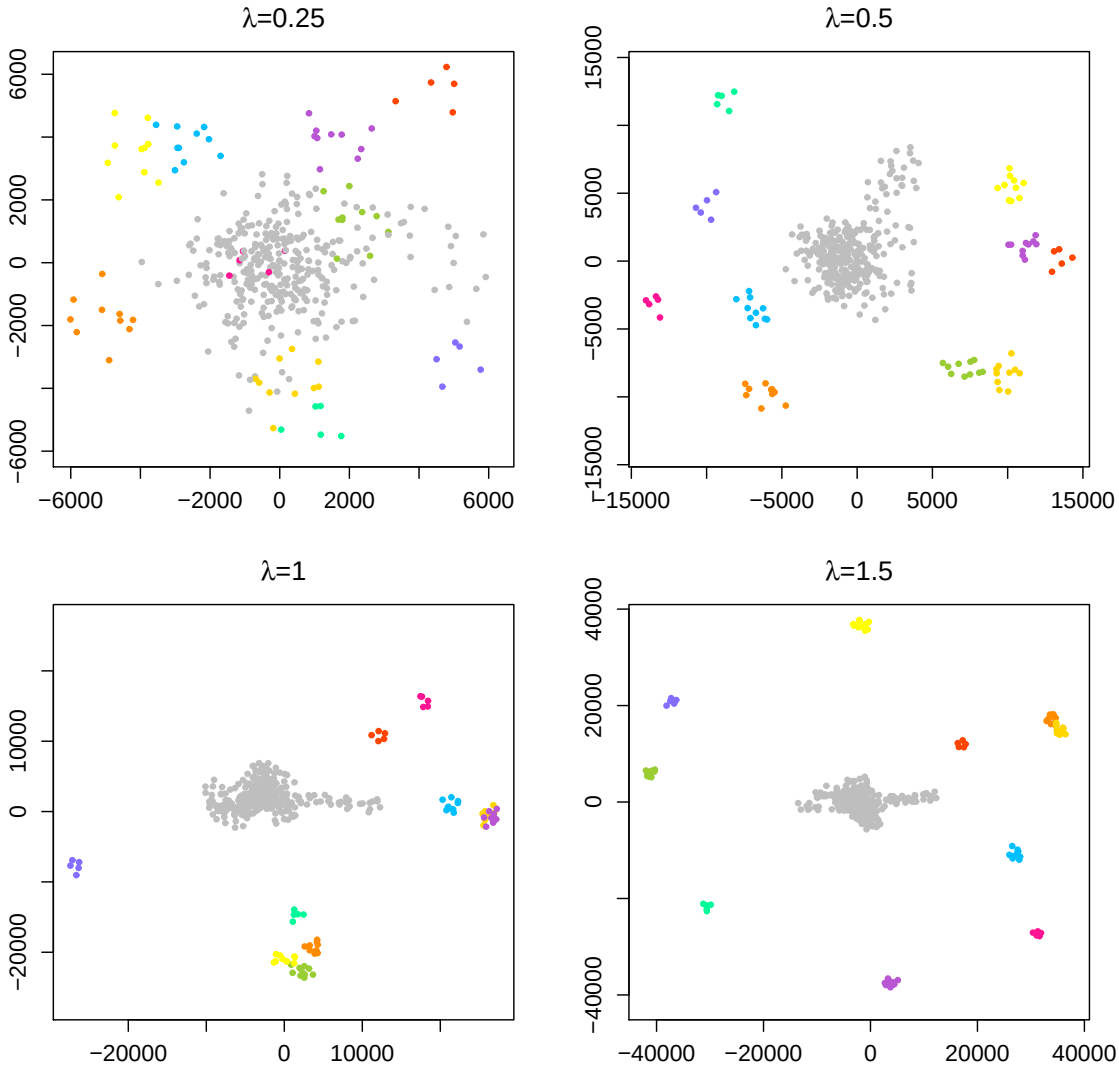

Figure 20: *Olivetti Faces: Configurations for four choices from the B-C family of energy functions.*

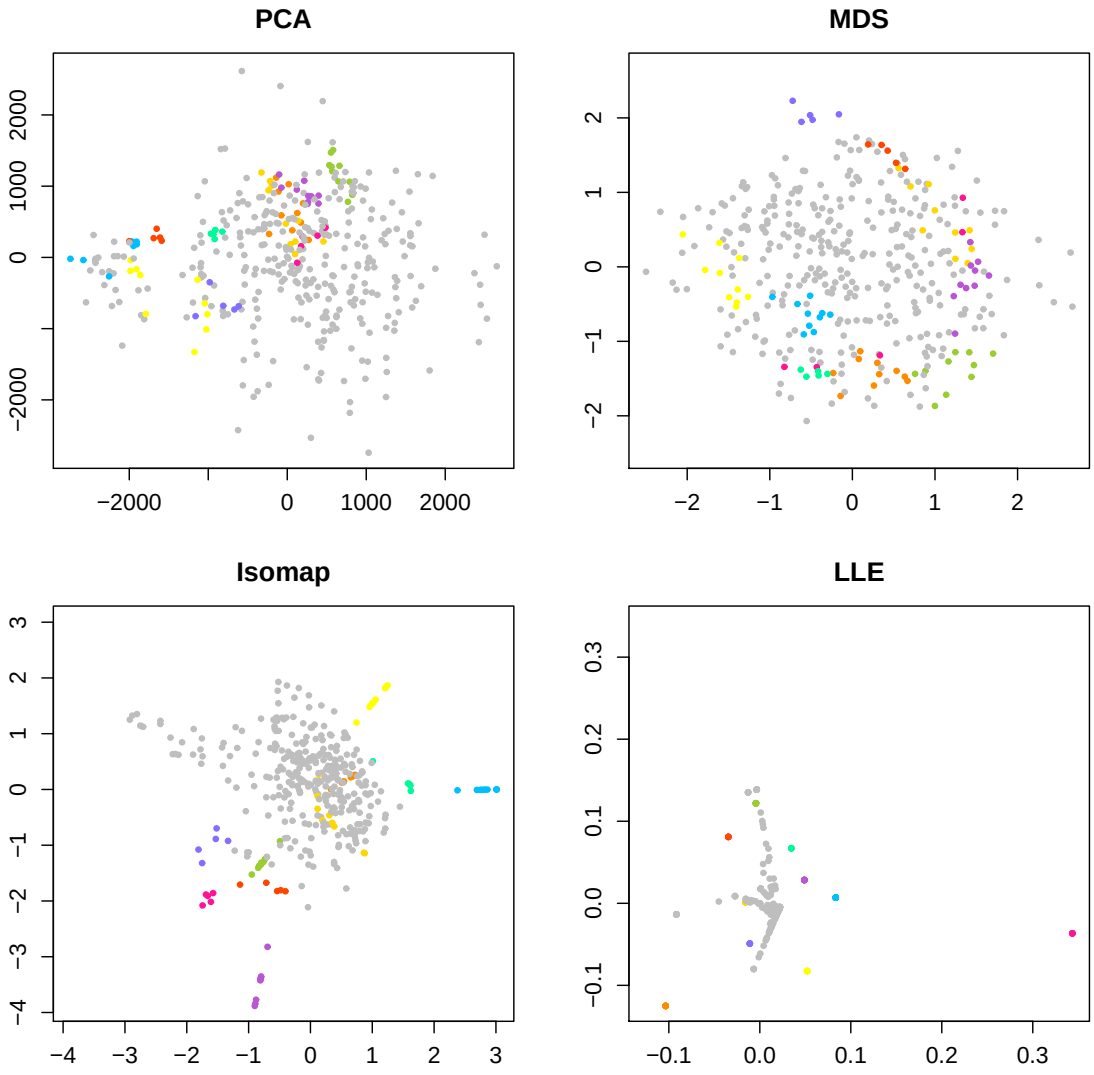

Figure 21: *Olivetti Faces: Configurations from PCA, MDS, Isomap and LLE.*

facial images of different persons together. The reason that these images are not distinguishable might be the crudeness of the Euclidean distance measure applied to the pixel dimensions. One should expect that some pixel dimensions are more informative than others; for example, the grey scale in the corner areas of an image does not change much and therefore contains less information, yet we treated these dimensions equally with the other more informative dimensions.

Figure 21 shows 2D projections of 3D configurations generated from the other four methods: PCA, MDS, Isomap and LLE. PCA and MDS did not find any clusters. Isomap did find a couple of clusters, but not as many nor as clearly as LMDS. LLE is quite successful on this dataset. It found even more clusters than LMDS (see the grey dots separated from the large grey part). The data points within each cluster collapse into one point, hence the colored points are in fact overplotted multiple points.

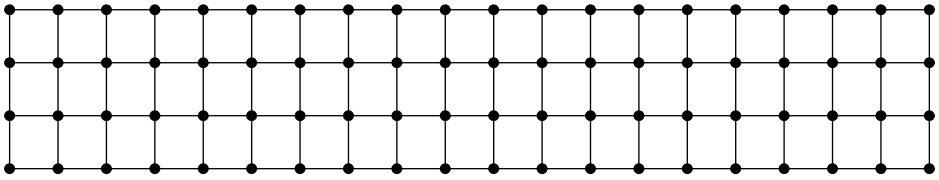

Figure 22: *Original Configuration*

6.2 Simulated Data

We introduced four parameters in the B-C energy functions, namely λ, μ, ν and t, and we tried to interpret them according to physical intuition. In this subsection, we will see how these parameters influence the configurations in an artificial example. The data set (Figure 22) has 80 data points that are equally spaced on a rectangular grid of 4 rows by 20 columns. The distance between adjacent points is 1. Since this is a small data set, for each data point we only define its adjacent points as its neighbors, that is, we use metric neighborhoods with radius 1. For interior points, this amounts to 4 nearest neighbors, for corner points 2 nearest neighbors, and for the other peripheral points 3 nearest neighbors.

Parameter λ: We set μ, ν and t equal to 1 and let λ vary to examine its effect. The configurations for $\lambda = 0.01,\ 0.2,\ 0.5,\ 1,\ 1.5$ and 3 are shown in Figure 23. Both small and large λ's fail to give reasonable results. When $\lambda = 0.2, 0.5$ and 1, the rectangular shape is recovered, but there is over-stretching on the long side: the larger λ, the more severe the over-stretching is. This observation is consistent with the interpretation of λ as the coupling strength, that is, the power on the coupling $\frac{n_e}{n_1 n_2}$. Image we split the 80 points into two groups in two ways, along the center of the long side and along the center of the short side, respectively. For both splits, $n_1 = n_2 = 40$ but $n_e = 4$ in the first case and $n_e = 20$ in the second case. This means that coupling is weaker in the direction of the long side, implying that the long side will experience much greater stretching than the short side, and this effect will increase as λ increases. A similar argument using splitting the long side on different locations explains the unequally spaced layout along the long side: points are more spread out around the center because this is where coupling is weaker compared to the extremes.

Parameter μ: In Figure 24, we examine the effect of varying μ as λ, ν and t are set to 1. Negative and small positive values of μ do not yield reasonable configurations. It appears that $\mu \geq 1$ is a good choice. We can explain this phenomenon with a property of Box-Cox transformations: Figure 18 shows that as μ decreases, the curve beyond 1 becomes flatter, implying that a change in configuration distance d affects the total

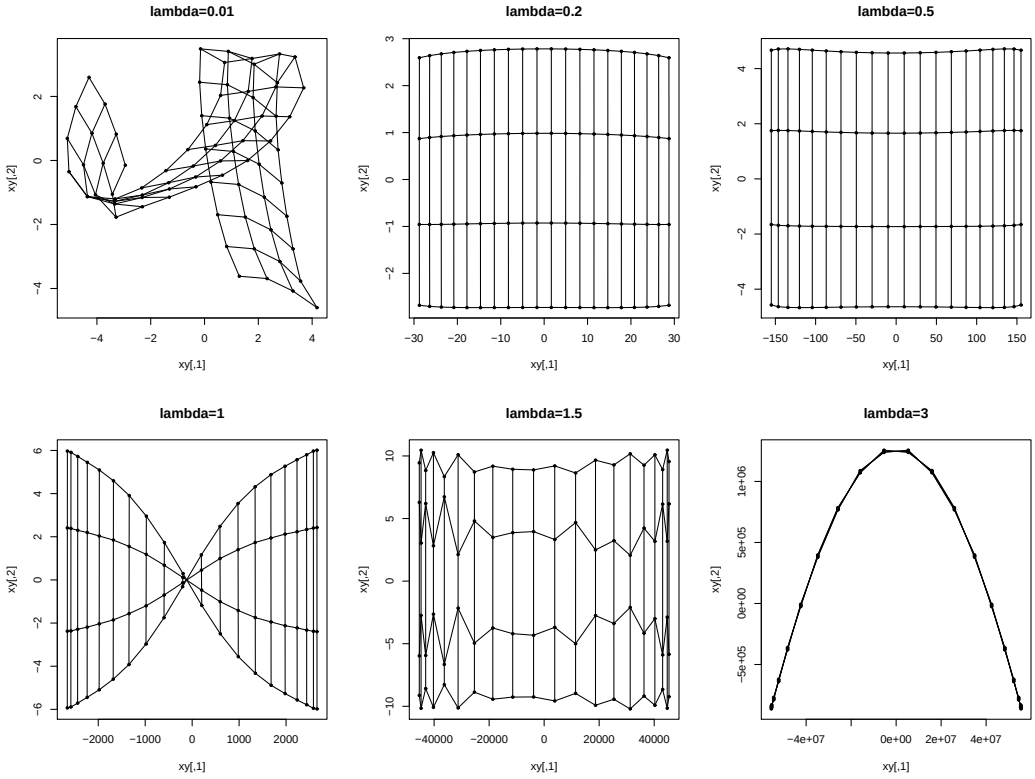

Figure 23: *Configurations with* $\lambda = 0.01, 0.2, 0.5, 1, 1.5$ *and* 3. *(The axes are not to scale.)*

energy very little. In other words, with small μ, the points are subject to very weak forces and are therefore more easily trapped to the local minima. Interestingly, in the quite reasonable configurations with $\mu = 1, 2, 3$, larger values of μ reduced the stretch bias in the direction of the long side. For example, with $\mu = 3$, the points are almost equally spaced. We will make the same observation in a one-dimensional problem.

Parameter ν: To see the effect of ν, we need to expand the neighborhood so that the distances between neighboring points are not all equal (ν has no differential effect when all $D_{i,j}$ are equal). We set the metric threshold to be $\sqrt{2}$ so the distances between neighboring points can be either 1 or $\sqrt{2}$. From (26), we know that larger values of ν give more weight to match larger distances between connected points. The configurations in Figure 25 give evidence in favor of this argument. The larger ν, the smaller scale is the configuration.

Parameter t: This is the parameter for adjusting the strength of the repulsive force between non-connected pairs of points. From Figure 26 we see that the larger t, the larger is the overall scale of the configuration. Since a larger number of points introduces more repulsive force in the direction of the long side, the ratio of the width to the height increases as t increases.

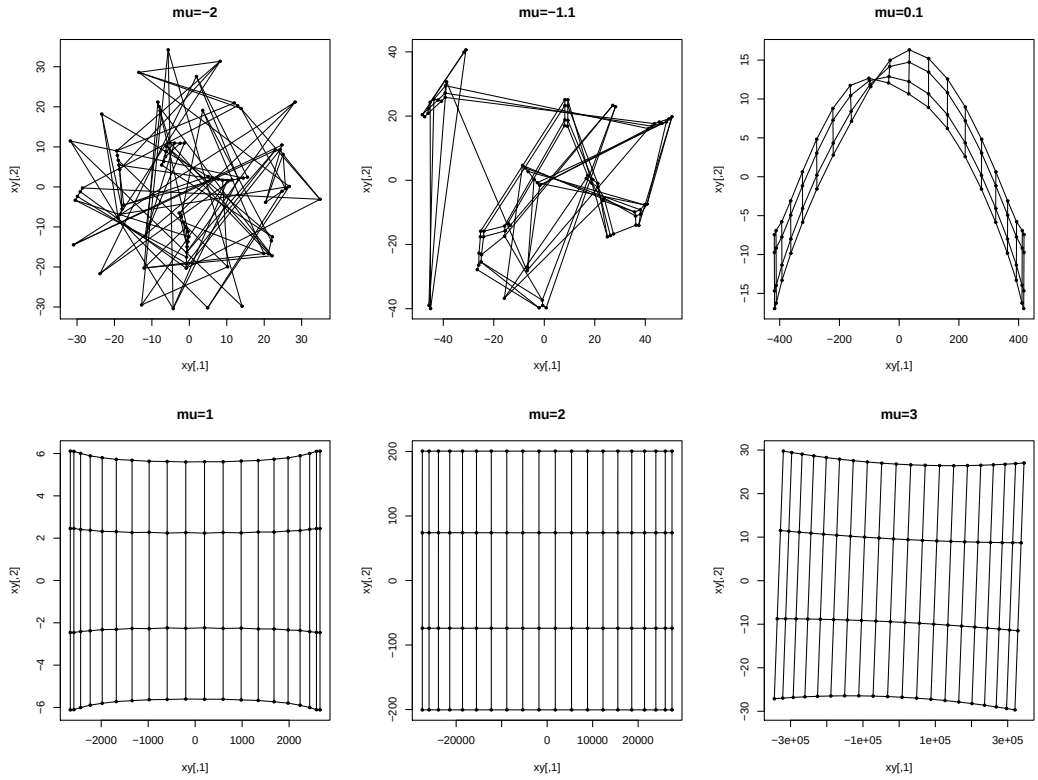

Figure 24: *Configurations with* $\mu = -2, -1.1, -0.1, 1, 2$ *and 3. (The axes are not to scale.)*

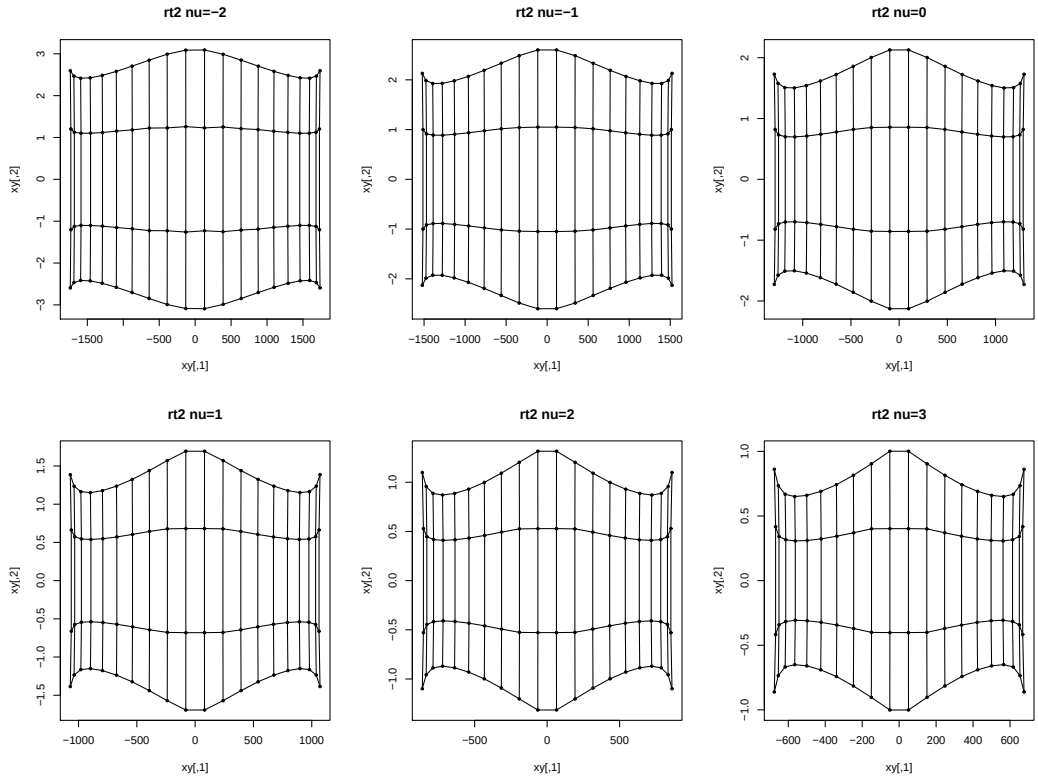

Figure 25: *Configurations with* $\nu = -2, -1, 0, 1, 2$ *and 3. (The axes are not to scale.)*

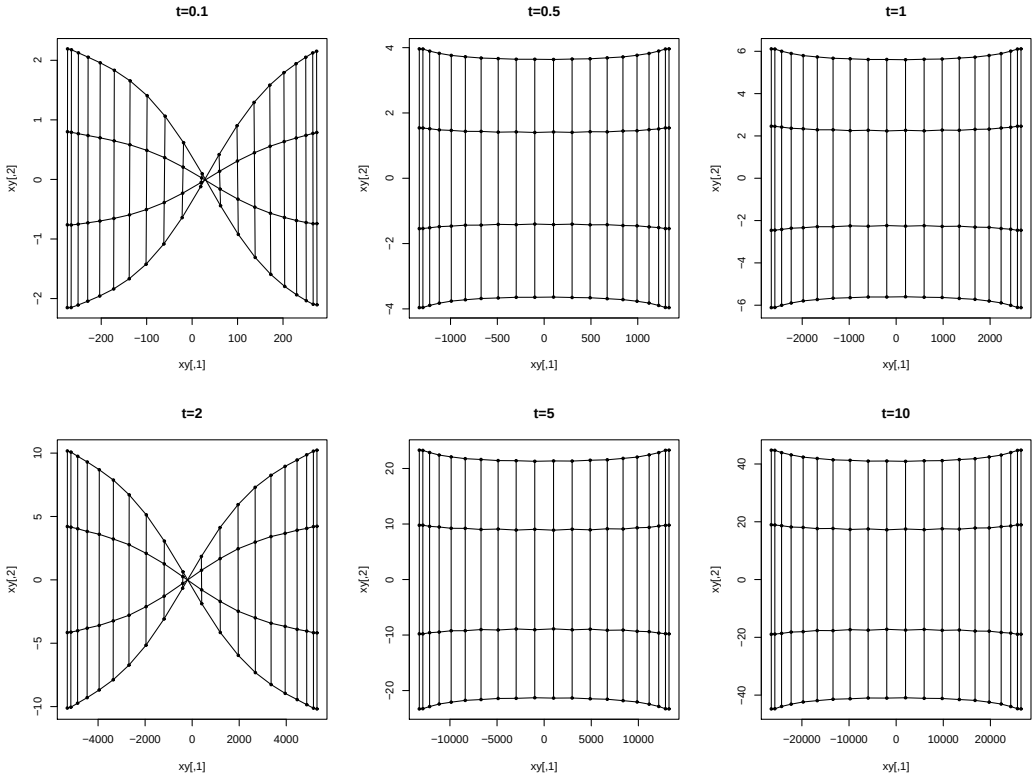

Figure 26: *Configurations with* $t = 0.1, 0.5, 1, 2, 5$ *and* 10. *(The axes are not to scale.)*

7 SOME THEORY

7.1 Four Equispaced Points on a Line

To explore further the meaning of the parameters of the B-C energy functions, we consider a simple one-dimensional problem with four equispaced points. For simplicity but without loss of generality, we assume the distance between two adjacent points is one. We define the neighborhood of each point to be the adjacent points, which makes this a 2-nearest neighborhood for the inside points and a 1-nearest neighborhood for the extreme points. Then the B-C energy functions become:

$$\begin{aligned} U(d_{12},\ d_{23},\ d_{34}) \ = \ & \frac{d_{12}^{\mu+\frac{1}{\lambda}} + d_{23}^{\mu+\frac{1}{\lambda}} + d_{34}^{\mu+\frac{1}{\lambda}}}{\mu+\frac{1}{\lambda}} \ - \ \frac{d_{12}^{\mu} + d_{23}^{\mu} + d_{34}^{\mu}}{\mu} \\ & - \ t\ \frac{(d_{12}+d_{23})^{\mu} + (d_{23}+d_{34})^{\mu} + (d_{12}+d_{23}+d_{34})^{\mu}}{\mu} \end{aligned} \tag{29}$$

Due to the symmetric structure of the original data, we expect $d_{12} = d_{34}$ in the configuration and therefore the energy (29) function simplifies to

$$U(d_{12},\ d_{23}) \ = \ \frac{2d_{12}^{\mu+\frac{1}{\lambda}} + d_{23}^{\mu+\frac{1}{\lambda}}}{\mu+\frac{1}{\lambda}} \ - \ \frac{2d_{12}^{\mu} + d_{23}^{\mu}}{\mu} \ - \ t\ \frac{2(d_{12}+d_{23})^{\mu} + (2d_{12}+d_{23})^{\mu}}{\mu}$$

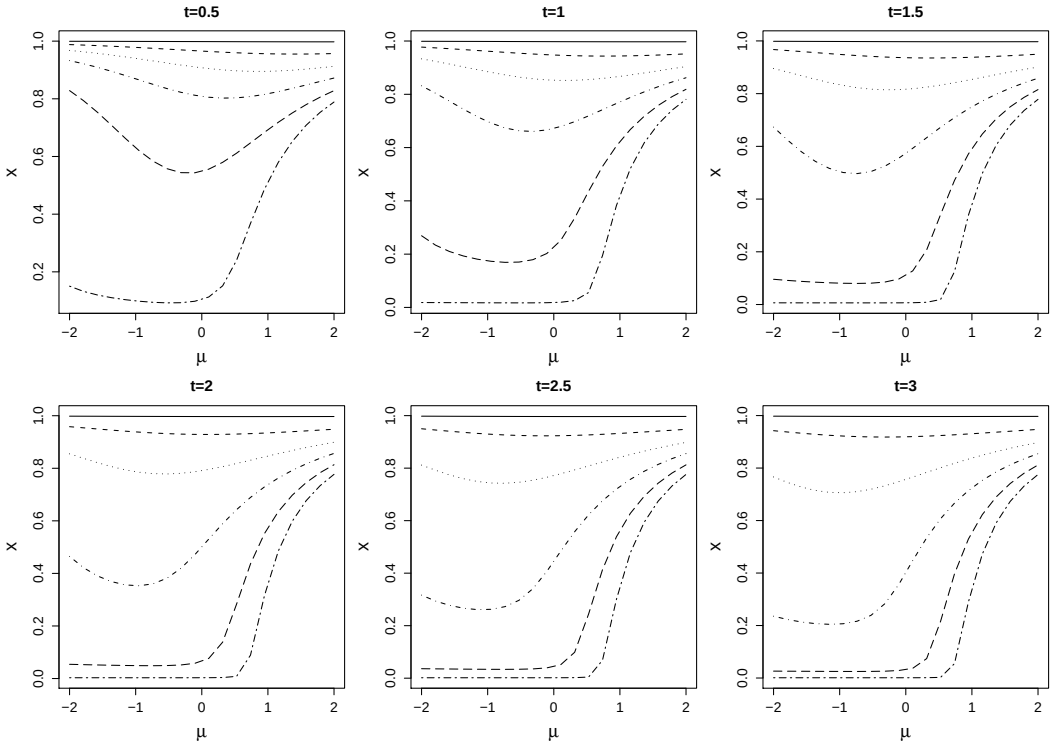

Figure 27: *The six curves in each panel correspond to different λ's. Top to bottom:* $\lambda = 0.010, 0.201, 0.495, 0.903, 1.646, 3.000$.

In this form of the energy function one can optimize the size of the configuration:

$$\begin{aligned} \mathrm{U}^*(d_{12},\ d_{23}) &= \min_{s>0}\ \mathrm{U}(s\cdot d_{12},\ s\cdot d_{23}) \\ &= \left(\frac{1}{\mu+\frac{1}{\lambda}}-\frac{1}{\mu}\right)\ \frac{(2d_{12}^{\mu}+d_{23}^{\mu}+2t\,(d_{12}+d_{23})^{\mu}+t\,(d_{23}+2d_{12})^{\mu})^{\mu\lambda+1}}{(2d_{12}^{\mu+\frac{1}{\lambda}}+d_{23}^{\mu+\frac{1}{\lambda}})^{\mu\lambda}} \end{aligned} \tag{30}$$

Equation (30) is a scale-invariant version of energy, and it can be written as a function of the length ratio of the end segment to middle segment $x = \frac{d_{12}}{d_{23}}$ by dividing both the numerator and the denominator by $d_{23}^{\mu^2\lambda+\mu}$:

$$U^*(x) = \left(\frac{1}{\mu+\frac{1}{\lambda}}-\frac{1}{\mu}\right)\ \frac{(2x^{\mu}+1+2t(x+1)^{\mu}+t(1+2x)^{\mu})^{\mu\lambda+1}}{(2x^{\mu+\frac{1}{\lambda}}+1)^{\mu\lambda}}$$

Minimizing $U^*(x)$ over x in the range from 0 to 1 gives the length ratio of the end segment to the middle segment from the optimal configuration.

Figure 27 shows the relation between the optimal ratio x and the power law of repulsion μ for six different values of repulsion weights t from 0.5 to 3 in increments of 0.5. The ratio x is in the range between 0 and 1 and the closer x is to 1, the smaller the

configuration bias is compared to the target dissimilarities. We experimented with μ's from -2 to 2 with 18 values in between. In each panel, six curves correspond to six values of the clustering power λ, from top to bottom, $\lambda = 0.01,\ 0.20,\ 0.50,\ 0.90,\ 1.65,\ 3.00$. Here are the relevant comparisons:

- The six panels show the same pattern in λ: larger values of λ lower x and hence move the two middle points toward the ends, which is consistent with the function of λ as the clustering power.
- Consider x as a function of μ as t and λ are fixed. For smaller clustering power λ, say $\lambda < 1$, the largest bias is found when μ is between -1 and 0. The bias decreases (x increases) when μ moves towards both ends and it decreases faster towards the positive end. For larger λ, x drops fast from 1 to 0 as μ moves from 2 to around 0 and it does not go up but stays near 0 as μ moves from 0 to -2.
- When we compare six curves with the same value of λ but different values of t, we find they move down in parallel, indicating larger t does introduce more bias.

We note that bias (here: low values of x) is not bad in itself. Often bias is beneficial in order to show local detail. Yet it is interesting to understand what combinations of parameter values in the B-C family cause low or high bias.

7.2 Equispaced Points on a Line, $N \to \infty$

We now consider a more general one-dimensional problem. Suppose we have $N+1$ equispaced points on a line of length 1. The distance between two adjacent points is $1/N$. The neighborhood is set as adjacent points. Assuming that the configuration is also of length 1, we consider the limit as $N \to \infty$. In the limit, the configuration can be described as a density function $f(s), s \in [0,1]$. To this end we write $||x_i - x_{i+1}|| = \frac{1}{N} f(s_i)$ for $s_i = \frac{i}{N}$. The energy function in this setup is:

$$\begin{aligned} U \;=\; & \sum_{i=1}^{N} \left(\left(\frac{1}{N}\right)^{\nu-\frac{1}{\lambda}} \frac{(f(s_i)\frac{1}{N})^{\mu+\frac{1}{\lambda}}}{\mu+\frac{1}{\lambda}} \;-\; \left(\frac{1}{N}\right)^{\nu} \frac{(f(s_i)\frac{1}{N})^{\mu}}{\mu} \right) \\ & -\; t^{\nu} \left(\sum_{i=1}^{N} \sum_{j=i}^{N} \frac{(\sum_{k=i}^{j} \frac{1}{N} f(s_k))^{\mu}}{\mu} \;-\; \sum_{i=1}^{N} \frac{(\frac{1}{N} f(s_i))^{\mu}}{\mu} \right) \\ \approx\; & \frac{1}{\mu+\frac{1}{\lambda}} \left(\frac{1}{N}\right)^{\mu+\nu-1} \int_0^1 f^{\mu+\frac{1}{\lambda}}(s)\, ds \;-\; \frac{1}{\mu} \left(\frac{1}{N}\right)^{\mu+\nu-1} \int_0^1 f^{\mu}(s)\, ds \\ & -\; \frac{t^{\nu}}{\mu} N^2 \int_0^1 \int_y^1 \left(\int_y^z f(s)\, ds \right)^{\mu} dz\, dy \;+\; \frac{t^{\nu}}{\mu} \left(\frac{1}{N}\right)^{\mu-1} \int_0^1 f^{\mu}(s)\, ds \end{aligned} \tag{31}$$

For simplicity we assume $\mu > -1$ since the fourth term in (31) is dominated by the third term and hence can be neglected as $N \to \infty$. We next make the first three terms

comparable by requiring

$$t^{\nu} N^2 \;=\; c^{\nu} \left(\frac{1}{N}\right)^{\mu+\nu-1} \quad , \tag{32}$$

where c is a constant that can be interpreted as an asymptotic repulsion weight. This interpretation becomes more evident if the condition (32) is rewritten as follows:

$$t \;=\; c \left(\frac{1}{N}\right)^{\frac{\mu+\nu+1}{\nu}} \tag{33}$$

This condition together with the assumption $\mu > -1$ entails that the first three terms in (31) are of the same order in N. Hence in the limit the energy function becomes an *energy functional* as follows:

$$\begin{aligned} U \;\propto\; & \frac{\mu}{\mu+\frac{1}{\lambda}} \int_0^1 f^{\mu+\frac{1}{\lambda}}(s)\, ds \;-\; \int_0^1 f^{\mu}(s)\, ds \\ & -\; c^{\nu} \int_0^1 \int_y^1 \left(\int_y^z f(s)\, ds\right)^{\mu} dz\, dy \end{aligned} \tag{34}$$

We should note the interesting fact that for such asymptotics to hold, the repulsion weight t has to shrink according to (33) as a function N.

Minimization of the functional (34) over $f(s)$ is a variational problem. If we let $F(s) = \int_0^s f(r),\ s \in [0,1]$, be the length of the configuration corresponding to the original segment $[0,\, r\,]$, the energy functional (34) can be rewritten as follows:

$$\begin{aligned} U \;\propto\; & \frac{\mu}{\mu+\frac{1}{\lambda}} \int_0^1 F'^{\mu+\frac{1}{\lambda}}(s)\, ds \;-\; \int_0^1 F'^{\mu}(s)\, ds \\ & -\; c^{\nu} \int_0^1 \int_y^1 (F(z) \;-\; F(y))^{\mu}\, dz\, dy \end{aligned} \tag{35}$$

By definition $F(0) = 0$. If we constrain the length of the configuration to be 1, we have $F(1) = 1$. In this case minimization of (35) becomes a two-boundary variational problem. For the special case $\mu = 1$, we can solve this problem by applying Euler's equation:

$$f(s) \;=\; (c^{\nu} s(1-s) + C_1)^{\lambda} \tag{36}$$

The constant $C_1 > 0$ is chosen such that $\int_0^1 f(s)\, ds = 1$. The form of (36) indicates the configuration is symmetric about $\frac{1}{2}$ and spreads out more in the center.

So far we have solved the problem under the constraint $F(1) = 1$. To loosen the constraint but still apply Euler's equation, we introduce a scale parameter a. That is, the configuration is described by $\tilde{F}(r) = aF(r),\ r \in [0,1]$, which has two parts: $a > 0$ describes the scale of the configuration, while $F(r)$, normalized to $F(1) = 1$, describes the shape of the configuration. With these assumptions the energy functional can be

rewritten as follows:

$$U(F,a) \;=\; \frac{\mu}{\mu+\frac{1}{\lambda}}\, a^{\mu+\frac{1}{\lambda}} \int_0^1 F'^{1+\frac{1}{\lambda}}(s)\, ds \;-\; a^{\mu} \int_0^1 F'^{\mu}(s)\, ds$$
$$- \; c^{\nu} a^{\mu} \int_0^1 \int_y^1 (F(z)-F(y))^{\mu} dz\, dy$$

The energy functional should be minimized over both the scale a and the shape function $F(s)$, $s \in [0,1]$. This setup allows us to maintain the constraint $F(1) = 1$ without loss of generality. Again, we know how to solve this problem when $\mu = 1$. First we minimize the energy functional over a: The optimal solution a_* for fixed $F(s)$ is

$$a_* \;=\; \left(\frac{c^{\nu} \int_0^1 F(s)(2s-1)ds + 1}{\int_0^1 F'^{\mu+\frac{1}{\lambda}}(s) ds} \right)^{\lambda} .$$

Then we get the solution for $f(s)$ by applying Euler's equation,

$$f(s) \;=\; (a_*^{-\frac{1}{\lambda}} c^{\nu} s(1-s) + C_1)^{\lambda} \tag{37}$$

where C_1 is chosen such that $\int_0^1 f(s)\, ds = 1$. By comparing (36) and (37), we see that the shape of the configuration does not change much by allowing the arbitrary scale.

8 DISCUSSION AND CONCLUSIONS

We conclude by discussing non-linear dimension reduction, graph layout and proximity analysis with the benefit of hindsight of the ideas presented in this work. We will distill some general principles, describe our contribution, and point to opportunities for future research. We start with a discussion of dimension reduction.

8.1 From Non-Linear Dimension Reduction to Graphs: Contiguity, Distances, and Localization

Nonlinear dimension reduction has the general goal to map high-dimensional data to low dimensions in such a way that contiguity between data points is preserved as best as possible. Obviously this goal can be attained the better the lower the intrinsic dimensionality of the high-dimensional data. Ideally, dimension reduction would amount to flattening of a curved sheet with little distortion. Realistically, though, dimension reduction will have to contend with non-linearity that cannot be flattened without distortion, with higher intrinsic dimension than provided in the reduction, and with noise in high dimensions. This latter difficulty — noise — seems especially formidable because with increasing dimension of the dataspace noise may drown out any systematic

structure. This consideration may be at the root of statisticians' general scepticism towards reduction of data spaces whose dimension is in the thousands. Yet, as the image data examples of the machine learning literature show, there do exist very high-dimensional datasets that are so highly structured that noise is not really an issue and nonlinear dimension reduction is successful.

When we named preservation of contiguity as central to dimension reduction, we only described half of the goal. A simple example illustrates the point: a bent sheet of paper in three dimensions can be flattened by crumpling it and pressing it on a table (without tearing it), thus preserving contiguity: nearby points on the sheet remain nearby. Crumpling, however, introduces contiguities that did not exist in the original dataspace. Therefore, dimension reduction should also preserve non-contiguity. We will refer to this second requirement below, and we note here that preserving non-contiguity can be interpreted in a lax manner: distant pairs of objects should be mapped to distant pairs of points, but exactly how distant is not much of a concern.

In order to judge contiguity, most dimension reduction methods rely on measures of distance or dissimilarity between pairs of high-dimensional objects as well as between pairs of their low-dimensional counterparts.[1] The task of dimension reduction can then be seen as finding maps of the objects in low dimensions that reflect the distances of the high-dimensional objects as best as possible, which us just the task of proximity analysis. Distilling high-dimensional objects in inter-object distances can be criticized on various grounds. For one thing, there is great leeway in forming distances and dissimilarities, and the use of simple Euclidean or L_p distances seems crude and ripe for innovation. At a minimum, future research should investigate variable selection and variable scaling in the formation of distances.

Yet, the use of distances is natural for nonlinear dimension reduction because distances give quantitative meaning to the notion of contiguity. One could define pairs of points to be contiguous when their distance is below a fixed threshold or if at least one of the two points is among the K-th nearest neighbors of the other. The importance of contiguity for nonlinear dimension reduction is that typical approaches attempt to reconstruct only local structure between contiguous points — following what we call the *principle of localization*. Such localization in effect generates a distance-weighted graph consisting of the pairs of contiguous points as edges and the interpoint distances as weights, what we may call the "local graph".

The success of any dimension reduction depends on getting localization right. As

[1]This interpretation applies even to LLE as inner products and Euclidean distances are largely interchangeable.

the Swiss roll example illustrates, choosing localization parameters too large may cause the local graph to contain cordal bridges between parts of the manifold that are distant in terms of intrinsic topology. On the other hand, choosing the localization parameters too small may introduce tears in the local fabric of the manifold. These problems call for research into "sufficiency" of the local graph for the underlying manifold. A truly convincing methodology would give principled guidance for choosing the parameters of the local graph. Future research might also investigate whether local choice of the neighborhood size would allow more faithful representation of local structure (for example, the K for K-nearest neighborhoods might be chosen locally). Finally, it would be useful to find ways of estimating local intrinsic dimension. For these problems it might be necessary to draw on current research into applied algebraic topology.

8.2 Graph Layout: Two Fundamental Approaches

By reducing high-dimensional data to a local graph and assuming that the local graph is a faithful description of the local intrinsic fabric of the data, we can now conceive of nonlinear dimension reduction as a graph layout problem. Just like dimension reduction, graph layout has the twin goals of preserving contiguity *and* non-contiguity (in a lax manner), where contiguity is now represented by the graph's edges and distance weights. Approaches to graph layout differ as much in how they deal with non-contiguity as with contiguity. Preservation of *non*-contiguity has been approached in at least three ways:

a) do nothing,

b) complete the distance weights to the full graph with a shortest-path-length algorithm so as to estimate large intrinsic distances, or

c) apply repulsive forces to pairs of points that are not edges.

Approach a) was tried and dismissed in MDS, but it has surprisingly resurfaced in LLE.[2] Approach b) goes back as far as Kruskal and Seery (1980) and is now the distinguishing feature of Isomap (Tenenbaum et al. 2000). Approach c) is standard in graph layout based on energy functions. Approaches b) and c) could be combined by filling in shortest-path-distances only partly, for paths two, three or five edges out, say, and applying repulsive forces to the remaining pairs of points. Such combinations

[2]LLE seems to avoid the crumpling problem by reconstructing locally affine as opposed to locally metric structure while entirely ignoring non-contiguity. In light of the failure of naively localized MDS the relative success of LLE poses something of a riddle. On the other hand, one might wonder whether the apparent tendency of LLE to create linear artifacts could be mitigated with a non-contiguity penalty. These questions are obviously material for future research.

would allow new trade-offs between the biases (and variances) of either approach, but their merits are unknown and should be investigated in future research. At this point we note that we contribute both to approaches b) and c): we propose a parametrized family of stress functions for use on complete distance matrices as well as a related parametrized family of energy functions for use on proper distance-weighted graphs.

Graph layout with energy functions has a long tradition, and many proposals can be found in the literature. The wealth of proposed energy functions poses problems, however. First, there is a need to systematize energy functions and describe their properties. Second, there is a need for criteria in choosing among energy functions. For both problems we proposed solutions: we systematized energy functions with a multi-parameter family that contains a number of previously proposed energy functions as special cases, and we proposed a method for selecting energy functions based on a data-driven meta-criterion. The latter proposal is based on the fact that there does not exist a single best energy function for all datasets simultaneously; instead, one needs to evaluate the performance of energy functions on any given dataset with a meta-criterion, that is, with a loss function that is not itself part of the energy functions used for optimization. Assuming that the meta-criterion measures desirable features of layouts, one then hopes that the proposed multi-parameter family of energy functions povides a search space for the meta-criterion that is sufficiently rich to produce satisfactory solutions for a large class of data problems. The two questions are therefore how rich the class of energy functions is and how desirable and practical the meta-criterion is. We discuss each question in turn.

8.3 B-C Energy and Stress Functions

The parametrized family of energy functions proposed here applies in its general form to distance-weighted graphs, that is, edges are assigned target distances or dissimilarities. Like most proposals in the literature, our energy functions are sums of terms that create attractive and repulsive forces between pairs of points. For a pair that is an edge of the graph, the two forces cancel out when the distance of the two points in the layout matches the target distance for that edge. For a pair that is not an edge, there exists only repulsive force, that is, there does not exist a finite proper placement of the two points. In what follows, we discuss the case of energy functions for a proper graph, but some of the discussion applies to the case of stress functions for full graphs as well.

While most proposals in the literature assume that the repulsive forces are the same in pairs that are edges and in pairs that are not, we only assume that repulsion

for graph-external pairs (non-edges) is of the same form but that it may differ by a weighting factor. This differential treatment of graph-internal and -external pairs may break metaphors of physics, but it moves us closer to paradigms of statistics: Consider a graph for which there exists a layout that perfectly matches the target distances in the given dimension; if the weighting factor for graph-external repulsion is zero, then this layout will have minimum energy (attractive and repulsive forces in the graph cancel out as all target distances are matched). Thus, vanishing graph-external repulsion is somewhat akin to unbiased or consistent estimation. By comparison, non-vanishing graph-external repulsion is analogous to Ridge penalties in linear regression: while a Ridge penalty biases an estimator to zero, graph-external repulsion pulls non-connected pairs of points toward infinite distance from each other, a bias one may call "stretching" in analogy to Ridge's "shrinking". In this view, allowing only identical and equally weighted repulsion externally and internally to the graph is akin to allowing only Ridge penalties with a unit coefficient. From a statistical point of view such a limitation is clearly unsatisfactory as data-dependent control of bias is an essential device in any complex estimation problem.

A major consideration in graph layout is the stability of optimization, as exemplified by the failure of MDS after removal of large target distances and the success of our local MDS after the addition of a repulsive energy. Thus "stretching" introduces a bias that makes stable layouts possible in the first place.

The weighting factor for graph-external repulsion is the first of several parameters in our parametrized family of energy functions. A second parameter was inspired by Noack's (2003) work on clustering properties of energy functions. In a qualitative formulation his clustering postulate says that clusters formed by tight subgraphs should be placed the farther apart the fewer the edges that connect them. A measure of connectedness is the so-called coupling strength, a number between 0 and 1 corresponding to, respectively, a zero and a maximal number of edges between the subgraphs. Noack's strict clustering postulate asks that target distances between clusters be upwardly biased by the reciprocal of the coupling constant, a fact that has him favor very specific energies. Even he, though, introduces for illustration a larger family of energies that differ in the degree of their upward bias. In a similar vein, we allow the upward bias to be any positive power of the reciprocal coupling strength, and it is this power that constitutes the second parameter of our family of energies. By controlling this power one can obtain layouts that vary in the crispness with which they show clustering: the larger the parameter value, the tighter the clusters are laid out in relation to the distance from each other.

Two more parameters complete the parametrization of our family of energies: a

power parameter for the layout distances and a power parameter for the target distances. The former parametrizes the power law of repulsion in the sense of physics, and the latter assigns a weight to each edge as a function of the target distance. In statistical terms, these parameters control the robustness to large and small layout distances and large and small target distances, respectively. A need for power laws as well as logarithmic laws motivated our use of Box-Cox transformations rather than raw powers. The convenient logarithmic fill-in and the fact that Box-Cox transformations are always ascending enriches our family of energies in such a way that numerous previously proposed energies become special cases: full-graph energies such as Kruskal's (1964) stress function and Sammon's (1969) method which coincides with Kamada and Kawai's (1989), as well as proper graph energies such as our newly proposed stress for local MDS and energy functions by Fruchterman and Reingold (1991), Davidson and Harel (1996, after some simplification), and finally all of Noack's (2003) proposals, including his LinLog, QuadLin and PolyLog energies.

In a sense there exists a further parameter, both in graph layout and dimension reduction: essentially, the size of the graph. In graph layout one may partly extend graphs by adding edges with the shortest-path-length algorithm (as mentioned above), and in dimension reduction one may size the neighborhood graph for example by choosing the number of nearest neighbors. Thus, the relevant graph is rarely a given; it is constructed. In general one can say: the more connected the graph, the more stable is the layout problem and the less stretching bias is needed.

8.4 The Selection Problem: LC Meta-Criteria

With this wealth of energy functions, we have a promising domain in which to search for useful and insightful graph layouts and dimension reductions, but without further help the search would remain an art rather than a science. Our remaining step is therefore the construction of a meaningful meta-criterion for data-dependent selection of energies. The inspiration for the meta-criterion stems from topology and its notion of continuity: nearby should be mapped to nearby. There are many ways in which agreement of nearness can be measured, but in general terms they will boil down to this: compute for each object two types of neighborhood, one in terms of the graph or the high-dimensional data, and one in terms of the layout or embedding; then tally the overlap of the two types of neighborhoods and average the result over all objects. This general concept can be realized both for graphs and for high-dimensional data. In our execution we worked largely in the framework of non-linear dimension reduction, but the underlying assumption of a complete distance matrix is available for graphs as well

(as always through shortest-path-length completion). The result is a local continuity or LC meta-criterion.

The plausibility of such a meta-criterion cannot be contested as it measures a minimal requirement of local faithfulness of layouts and embeddings vis-à-vis the data, be it a distance-weighted graph, a dissimilarity matrix, or high-dimensional data. The practical success in selecting layouts and embeddings is another matter. It is a theoretical possibility that LC meta-criteria give high ratings to layouts of most methods and do not usefully discriminate among them. That this is not so is shown empirically in our work: different layout methods do fare differently on different data examples, and LC meta-criteria do provide useful yardsticks for telling those that are locally faithful from those that are not in a collection of layouts or embeddings. As LC meta-criteria embody minimal requirements and are not smooth functions of the layout or embedding, they cannot be directly optimized and cannot replace energy functions, but they can serve as selection criteria.

	Proximity Analysis	**Energy-Based Graph Layout**	**Dimension Reduction**
Data	full matrix of target distances/ dissimilarities	distance-weighted proper graph	full matrix of high-dimensional target distances
Output	configuration	layout	embedding
Criteria	stress	energy	either
Family of Criteria	3-parameter B-C	4-parameter B-C with graph-external repulsion	either
Selection	LC meta-criteria		
Potential Pre-Processing	completion of distance matrix with shortest-path-lengths	localization by removing edges with large distances	either

Table 5: *An overview of terminology and correspondences between proximity analysis, graph layout and non-linear dimension reduction. The last row describes potential pre-processing steps that may be needed before proximity analysis or graph layout can be applied.*

8.5 Summary

The toolbox for graph-based non-linear dimension reduction, for energy-based graph layout, and for stress-based proximity analysis is summarized in Table 5. The connections between the three are simple, but the tools are rich and the present work barely touches their wealth of possibilities. What it contributes is briefly summarized in this final list:

- We exploited a uniform framework of proximity analysis and graph layout that includes MDS-style stress functions for complete distance/dissimilarity matrices and repulsion-based energy functions for proper distance-weighted graphs.
- We constructed parametrized families of energy and stress functions that serve as search spaces for useful layouts, configurations, embeddings.
- The parameters common to energy and stress functions are two power parameters for robustness to small and large distances in the data and in the embeddings, and a clustering parameter that controls cluster separations (Noack 2003).
- The parameter specific to energy functions is a weight factor for graph-external repulsion. It controls the stretching bias needed for stable energy minimization.
- The instability problem of localizing MDS by removing large distances is solved with the derivation of a weighted repulsive energy.
- The proposed families of stress and energy functions contain many previously proposed cases, including stresses by Kruskal (1964), Sammon (1969), Kamada and Kaway (1989), and energies by Fruchterman and Reingold (1991), Davidson and Harel (1996, after some simplification), Noack (2003, all proposals).
- In all cases we replaced the trial-and-error method of selecting one of the published energy or stress functions with a principled method for selecting parameters with meta-criteria for local faithfulness.
- In addition, the meta-criteria yield graphical diagnostics for local faithfulness.
- Non-linear dimension reduction can be cast as graph layout of a graph of local edges, or as proximity analysis of a distance matrix obtained with shortest-path-length completion. An example of the former is LMDS proposed here, and an example of the latter is Isomap (Tenenbaum et al. 2000) if classical MDS is replaced with Kruskal-Shepard MDS.

APPENDIX: ENERGY MINIMIZATION

8.6 Gradient Descent

Let the $n \times d$ matrix $X = (x_1, \cdots, x_n)^T$ represent the set of points in d dimensions. The D_{ij}'s are the target dissimilarities. The B-C energy function is

$$U(x_1, \cdots, x_n) = \sum_{(i,j)\in E} \left(D_{ij}^{\nu - \frac{1}{\lambda}} \frac{\| \mathbf{x}_i - \mathbf{x}_j \|^{\mu + \frac{1}{\lambda}} - 1}{(\mu + \frac{1}{\lambda})} - D_{ij}^{\nu} \frac{\| \mathbf{x}_i - \mathbf{x}_j \|^{\mu} - 1}{\mu} \right) - t \sum_{(i,j)\in E^C} \frac{\| \mathbf{x}_i - \mathbf{x}_j \|^{\mu} - 1}{\mu} \tag{38}$$

Let $\bigtriangledown U = (\bigtriangledown_1, \cdots, \bigtriangledown_n)$ be the gradient of the energy function with respect to X:

$$\bigtriangledown_i = \frac{\partial U}{\partial \mathbf{x}_i} = \sum_{j \in N(i)} \left({D_{ij}}^{\nu - \frac{1}{\lambda}} ||\mathbf{x}_i - \mathbf{x}_j||^{\mu + \frac{1}{\lambda} - 2} - {D_{ij}}^{\nu} ||\mathbf{x}_i - \mathbf{x}_j||^{\mu - 2} \right) (\mathbf{x}_i - \mathbf{x}_j) - t \sum_{j \in N^c(i)} ||\mathbf{x}_i - \mathbf{x}_j||^{\mu - 2} (\mathbf{x}_i - \mathbf{x}_j)$$

Define a $N \times N$ matrix M as follows:

$$M_{ji} = \begin{cases} {D_{ij}}^{\nu - \frac{1}{\lambda}} ||\mathbf{x}_i - \mathbf{x}_j||^{\mu + \frac{1}{\lambda} - 2} - {D_{ij}}^{\nu} ||\mathbf{x}_i - \mathbf{x}_j||^{\mu - 2} & \text{if } j \in N(i) \\ -t ||\mathbf{x}_i - \mathbf{x}_j||^{\mu - 2} & \text{if } j \notin N(i) \end{cases}$$

Note that M is symmetric. The gradient can be simplified to

$$\begin{aligned} \bigtriangledown_i = \frac{\partial U}{\partial \mathbf{x}_i} &= \sum_j M_{ji} (\mathbf{x}_i - \mathbf{x}_j) \\ &= (\sum_j M_{ji}) \mathbf{x}_i - \sum_j M_{ji} x_j \ , \end{aligned}$$

and

$$\bigtriangledown U = X * (E \cdot M) - X \cdot M \ ,$$

where E is a $d \times n$ matrix with all elements being 1. The symbol '$*$' represents elementwise multiplication of the two matrices, and the symbol '$\cdot$' stands for regular matrix multiplication

8.7 Majorization

Equation (38) can be written as follows:

$$U = \sum_{(i,j)\in E} D_{ij}^{\nu - \frac{1}{\lambda}} \frac{\| \mathbf{x}_i - \mathbf{x}_j \|^{\mu + \frac{1}{\lambda}} - 1}{\mu + \frac{1}{\lambda}} - \sum_{(i,j)\in V^2} \tilde{D}_{ij}^{\nu} \frac{\| \mathbf{x}_i - \mathbf{x}_j \|^{\mu} - 1}{\mu} \tag{39}$$

where $\tilde{D}_{ij} = D_{ij}$ if $(i,j) \in N$, $\tilde{D}_{ij} = t^{1/\nu}$ if $(i,j) \in N^c$. First we consider the majorizing functions for

$$f_\alpha(x) = \frac{x^\alpha - 1}{\alpha}, \quad x > 0 \ .$$

When $\alpha \leq 2$, it can be majorized by a quadratic function,

$$\begin{aligned} f_a(x) \leq & a_\alpha(z)x^2 + c_\alpha(z) \\ a_\alpha(z) = & \frac{1}{2}\alpha z^{\alpha-2} \\ c_\alpha(z) = & \frac{z^\alpha - 1}{\alpha} - \frac{1}{2}\alpha z^\alpha \end{aligned}$$

Now we consider the majorizing function for

$$f_\mu(x) = -x^\mu, \mu > 0, \quad x > 0 \ ,$$

when $\mu \geq 1$, $-x^\mu \leq -x$. When $0 < \mu < 1$,

$$\begin{aligned} f_\mu(x) \leq & b(z)x^2 + d(z)x \\ b(z) = & (1-\mu)z^{\mu-2} \\ d(z) = & (\mu-2)z^{\mu-1} \end{aligned}$$

Hence (39) is majorized by

$$\sum_{(i,j)\in E} D_{ij}^{\nu-\frac{1}{\lambda}} \left(a_{\mu+\frac{1}{\lambda}}(||z_i - z_j||)||\mathbf{x}_i - \mathbf{x}_j||^2 + c_{\mu+\frac{1}{\lambda}}(||z_i - z_j||) \right) - \sum_{(i,j)\in V^2} \tilde{D}_{ij}^{\nu} \frac{1}{\mu} ||\mathbf{x}_i - \mathbf{x}_j||$$

when $\mu \geq 1$, and by

$$\begin{aligned} & \sum_{(i,j)\in E} D_{ij}^{\nu-\frac{1}{\lambda}} \left(a_{\mu+\frac{1}{\lambda}}(||z_i - z_j||)||\mathbf{x}_i - \mathbf{x}_j||^2 + c_{\mu+\frac{1}{\lambda}}(||z_i - z_j||) \right) \\ & + \sum_{(i,j)\in V^2} \tilde{D}_{ij}^{\nu} \frac{1}{\mu} \left(b(||z_i - z_j||)||\mathbf{x}_i - \mathbf{x}_j||^2 + d(||z_i - z_j||)||\mathbf{x}_i - \mathbf{x}_j|| \right) \end{aligned}$$

when $0 < \mu < 1$. For computational simplicity, we vectorize the above majoring functions. As functions of the $n \times d$ matrix X, the terms are either in the form of

$$\sum_{(i,j)} w_{ij}||\mathbf{x}_i - \mathbf{x}_j||^2 = tr(X^T A X)$$

where $A = \sum_{(i,j)} w_{ij}(e_i - e_j)(e_i - e_j)^T$, or in the form of

$$\begin{aligned} -\sum_{(i,j)} w_{ij}||\mathbf{x}_i - \mathbf{x}_j|| \leq & -\sum_{(i,j)} w_{ij} \frac{1}{||z_i - z_j||} < z_i - z_j, \mathbf{x}_i - \mathbf{x}_j > \\ = & -tr(Z^T B X) \ , \end{aligned}$$

where $B = \sum_{(i,j)} \frac{w_{ij}}{||z_i - z_j||}(e_i - e_j)(e_i - e_j)^T$. Since $\frac{\partial tr(X^T A X)}{\partial X} = 2AX$ and $\frac{\partial tr(Z^T B X)}{\partial X} = B^T Z$, the minimized X^* can be obtained by solving a linear system, but the calculation involves the inverse of a $n \times n$ matrix which is extremely computationally expensive.

REFERENCES

Brandes, U., 2001, Drawing on Physical Analogies, in: *Drawing Graphs*, Kaufmann, M. and Wagner, D., eds., Springer: Berlin, 71-86.

Davidson, R. and Harel, D., 1996, Drawing graphs nicely using simulated annealing, *ACM Transactions on Graphics*, **15**(4), 301-331.

Di Battista, G., Eades, P., Tamassia, R., Tollis, I. G., 1999, Graph Drawing: Algorithms for the Visualization of Graphs. Prentice Hall.

Eades P., 1984, A heuristic for graph drawing, In *Congressus Numerantium*, **42**, 149-160.

Fruchterman, T. M. J. and Reingold, E. M., 1991, Graph drawing by force-directed placement, *Software-Practice and Experience*, **21(11)**, 1129-1164.

Gansner, E., Koren, Y., and North, S., 2004, Graph Drawing by Stress Majorization, *Graph Drawing*, 239-250.

Graef J., and Spence, I., 1979, Using Distance Information in the Design of Large Multidimensional Scaling Experiments, *Psychological Bulletin*, **86**, 60-66.

Kaufmann, M. and Wagner, D. (eds.), 2001, Drawing Graphs. Springer: Berlin.

Kamada, T., and Kawai, S., 1989, An algorithm for drawing general undirected graphs. *Information Processing Letters*, **31**, 7-15.

Kruskal, J. B., 1964a, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika*, **29**, 1-27.

Kruskal, J. B., 1964b, Nonmetric multidimensional scaling: a numerical method, *Psychometrika*, **29**, 115-129.

Kruskal, J. B. and Seery, J. B., 1980, Designing Network Diagrams, Technical Memorandum, Bell Laboratories, Murray Hill, NJ.

Michailidis, G. and de Leeuw, J., (2001), Data visualization through graph drawing, *Computation. Stat.*, **16**, 435-50.

Noack, A., 2003, Energy models for drawing clustered small-world graphs, Computer Science Report 07/2003, Brandenburg Technical University at Cottbus, Germany.

Roweis S. T. and Saul, L. K., 2000, Nonlinear dimensionality reduction by local linear embedding, *Science*, **290**, 2323-2326.

Sammon, J., 1969, A Non-Linear Mapping for Data Structure Analysis. *IEEE Transactions on Computers*, **C-18**(5), 401-409.

Saul, L. K. and Roweis, S. T., 2003, Think Globally, Fit Locally: Unsupervised Learning of Low Dimensional Manifolds, *J. of Machine Learning*, **4**, 119-155.

Tenenbaum, J. B., Silva, V. d., and Langford, J. C., 2000, A global geometric framework for nonlinear dimensionality reduction, *Science*, **290**, 2319-2323.

Torgerson, W.S., 1952, Psychometrika, **17**, 401-419.