LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-583076

Final Report: MINDES - Data Mining for Inverse Design

C. Kamath

September 19, 2012

Disclaimer

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Final Report
MINDES: Data Mining for Inverse Design

Chandrika Kamath
Lawrence Livermore National Laboratory
kamath2@llnl.gov
September 14, 2012

1 Introduction

This is the final summary report on the work done as part of the ARRA-funded, *MINDES: Data Mining for Inverse Design* SciDAC-e project [10]. The goal of this project was to apply data mining techniques to data generated by the Center for Inverse Design [5], an Energy Frontier Research Center (EFRC) of the Office of Science, US Department of Energy. This center is pursuing a new approach to material science; rather than using the conventional direct approach ("Given the structure, find the electronic properties"), they are using a "materials by inverse design" approach ("Given the desired property, find the structure"). The specific target properties of interest include general semiconductor optical and electrical properties.

The analysis work falls in the broad area of design of computer experiments [7], where an ensemble of simulations is used to guide physical experiments and gain insights into the design space which maps the inputs of the simulations to the output(s). As the simulations are often computationally expensive, the ensemble must be carefully designed to obtain the greatest insights into the physical phenomenon of interest using as few simulations as possible. A possible solution is to consider an incremental approach where we analyze the input/output data from the simulations that have already been run to identify the next set of simulations such that these new simulations would add the greatest insights, by either refining the original data set in a region of interest, or exploring new regions in the design space.

There were two aspects to the project - the investigation of analysis techniques likely to be relevant to the task of identifying the inputs for the new simulations and the application of data mining techniques to EFRC data. The original plan was to analyze a dataset of spinel materials to determine if we could predict the formation enthalpy based on the properties of the elements used in the material and the impurities that were added to create the semiconductor compounds. At the request of the EFRC, we also considered another dataset of ternary compounds which had been generated for some other purpose, but was now being analyzed to determine if it was possible to identify the properties of the elements associated with band gap type 1 semiconductors.

2 Technical approach

For the two problems being addressed, the data can be written in the form of a table, where a row represents a compound and the columns represent the features describing the compound. Associated with each compound is an output, which, in the context of our problems, is the formation enthalpy and the band gap type, respectively. The former is a continuous value and can be positive or negative, while the latter is a discrete quantity, with integer values from 1 through 4.

Our technical approach for the analysis, which focused on the original problem of predicting the formation enthalpy and was formulated before we obtained the data, considered two categories of techniques to gain insights into the data. The first was to use dimension reduction methods, where we try to determine features which are relevant to the output. The second is model building using regression methods, where we build a model to predict the output. We next briefly describe the methods we considered.

2.1 Dimension reduction

Dimension reduction is the process of transforming a high-dimensional dataset into a reduced dimensional representation while preserving meaningful structures in the data. The dimension

here refers to the number of features. These methods allow us to identify important features or transform the data into another space where the we may be able to build better predictive models.

We considered several transform-based dimension reduction techniques: the linear Principal Component Analysis (PCA) [11], that preserves the largest variance in the data while decorrelating the transformed dataset, as well as four popular nonlinear dimension reduction (NLDR) techniques: Isomap [17] preserves pairwise geodesic distances between data points; Locally Linear Embedding (LLE) [14] preserves the reconstruction weights that are used to describe a data point as a linear combination of its neighbors; Laplacian Eigenmaps [2] provides a low-dimensional representation in which the weighted distances between a data point and other points within a neighborhood are minimized; Local Tangent Space Alignment (LTSA) [21] constructs the global coordinate system by aligning the tangent spaces generated by local PCA on the neighborhood of each data point. We also investigated feature selection techniques, including Relief [13] and a correlation-based method [8].

Since many dimension reduction methods require that the reduced dimension of the data be explicitly set, we investigated techniques to determine the intrinsic dimensionality of a dataset. For PCA, we can exploit the eigenvalue spectrum, but for nonlinear methods, this idea only works in rare cases where the data lie on a linear manifold [15]. Instead, we used the lack-of-fit measures that measure the deviation between a certain objective in the input space and in the low-dimensional space for Isomap and LLE [18, 15]. We also consider other classical approaches that do not require the setting of input parameters or any explicit assumptions on the underlying model. These include a robust version of the box-counting approach that determines the locally linear scale in the presence of noise in the data [9, 3] and a statistical approach based on hypothesis tests and nearest-neighbor information [19].

2.2 Building predictive models

In addition to dimension reduction techniques, we considered techniques to build predictive models for discrete and continuous outputs for the band-gap type dataset and the formation enthalpy dataset, respectively. We considered decision trees for discrete data and locally weighted regression and regression trees for continuous output. We briefly describe these methods below.

Decision trees [4, 12] belong to the category of classification algorithms wherein the algorithm learns a function that maps a data item into one of several pre-defined classes. Classification algorithms typically have two phases. In the training phase, the algorithm is "trained" by presenting it with a set of examples with known classification. In the test phase, the model created in the training phase is tested to determine how well it classifies known examples. If the results meet expected accuracy, the model can be put into operation to classify examples with unknown classification.

A decision tree is a structure that is either a leaf, indicating a class, or a decision node that specifies some test to be carried out on a feature (or a combination of features), with a branch and sub-tree for each possible outcome of the test. The decision at each node of the tree is made to reveal the structure in the data by dividing the feature space into regions where all data points are primarily of one class.

Locally weighted regression (LWR) [6] combines local models that are fit to nearby data. Unlike regression procedures using global models, which fit a single model to all data points, LWR fits a different regresion model everywhere, weighting the data points by how close they are to the point of interest. In addition to a regression function, LWR contains three critical parts: distance function, weighting function and smoothing parameters. The distance function determines the data around the point of interest that should be included in the fitting; the weighting function determines if

observations near the point of interest contribute more to the prediction than points which are far from it; and the smoothing parameter can be used to adjust the radius of the weighting function and reduce cross validation error when fitting the data. With right choices of these three elements, LWR can be quite successful at recovering the underlying nonliner regression function [1].

Regression trees are similar to decision trees, except the output (or class label) is continuous rather than discrete. Thus, in the creation of a regression tree, instead of evaluating a split based on how uniform the class labels are in each of the two groups resulting from a split, we consider how similar the values are in each of the two groups resulting from the split.

3 Accomplishments

We next present a brief summary of our work - first, the results of our investigation into dimension reduction methods, followed by the insights we obtained into the two datasets on formation enthalpy and the band-gap type.

3.1 Performance of dimension reduction methods

We first considered the performance of different dimension reduction methods using several real scientific datasets. Since we had experience with these datasets from past projects, we used them to understand the different methods better so we could select an appropriate method for the analysis of the two EFRC datasets. We briefly summarize our experiences below; more details are in a technical report (Section 6, item 4).

In our comparison of dimension reduction techniques, we considered both data transformation methods (linear and non-linear) and feature subset selection techniques. Using classification problems in five scientific datasets, we compared the classification error rates for the original dataset with those obtained for the reduced representations resulting from the application of the dimension reduction methods.

Our experiments indicate that, while the supervised feature subset selection techniques consistently improve the classification of all datasets, the data transformation methods do not. However, it is possible to use them to find properties of the data related to class labels. Our experiments show that both PCA and Isomap are able to find representations that improve data classification. Since both PCA and Isomap employ the eigenvectors corresponding to the largest eigenvalues, they seem to perform better than methods which use the eigenvectors corresponding to the smallest non-zero eigenvalues, such as LLE, Laplacian Eigenmaps and LTSA. Like PCA, when the data tend to have strong linear properties, Isomap can identify these properties. Isomap can also capture some kind of nonlinear properties that PCA can not find. Although there exists applications indicating that PCA is better than Isomap in terms of classification [20], our experiments indicate a different conclusion. We also observe that the ability to interpret the reduced dimension made by data tranformation methods is very limited.

A key finding of our experiments with dimension reduction techniques was the confirmation of the fact that we need to have sufficient number of observations (that is, data points) relative to the dimension of the problem. Specifically, the number of data points needed to accurately estimate the true dimension of a D-dimensional data set should be at least $10^{\frac{D}{2}}$ [16]. So, in practice, if the sample size of a dataset is small, we should try reducing the number of features using domain

information prior to determining its intrinsic dimensionality. As we shall see in Section 3.2, this finding has important implications in the analysis of the two datasets from the EFRC.

3.2 Analysis of the EFRC datasets

In this section, we describe our work on the two EFRC datasets we analyzed as part of this project. The first is the dataset on spinel materials, where we are interested in predicting the formation enthalpy based on the properties of the atomic species that form the spinel and the impurity (either a vacancy or a substitution) that is added to create the semiconductor. The second problem is finding properties relevant to semiconductors with band-gap type 1 from a dataset which included materials with band gap types 1-4.

3.2.1 Analysis of the formation enthalpy dataset

This dataset was composed of compounds derived from three spinel materials - Co2ZnO4, Rh2ZnO4, Mn2CrO4. In this problem, given a material, say Co2ZnO4, new compounds are created by introducing an impurity, which can be either a vacancy at one of the locations in the spinel structure or a substitution of one atom of an atomic species by another atomic species. These impurities can be at different charge states with a different formation enthalpy associated with each charge state.

There are three types of data provided for this problem:

- Formation enthalpy values associated with each charge state of a compound created from a spinel material by introducing an impurity.
- Structure data for the new compounds created from the three spinel materials. These provide the locations of the different atoms in 3-D space. Several of these files were missing for the compounds; so, the information on the structure was not used in the analysis.
- A file on material properties for 87 atomic species in the periodic table. These properties include information such as the row and column in the periodic table; atomic number and atomic weight; properties indicating the size, such as molar volume, single bond radius, and double bond radius; various electronegativity scales, such as Mulliken-Jaffe, Pauling and Allred-Rochow; density; boiling point; and melting point. Not all properties are available for all atomic species.

We observe that, for this problem, the data are not in a tabular form, that is, one of the tasks is to determine a representation of a compound for use in the analysis. This is challenging as we need to have a single representation for both vacancies and substitutions. We considered several possiblilities, and in consultation with the domain experts from the EFRC, focused on two:

- Periodic table (PT) dataset: In this smaller dataset, we consider just the periodic table information for each species to determine if we can learn any patterns from the data. The information for each species includes five quantities: its symbol, the number of atoms, and the group, row, and column of the periodic table for the species. Each compound is represented by five species: the species in the first location (Co in the case of Co2ZnO4); the impurity in the first location (none, if the compound was created by means other than the introduction of an impurity in the first location); the species in the second location (Zn in the case of

Co2ZnO4); the impurity in the second location (none, if the compound was created by means other than the introduction of an impurity in the second location); and the species in the third location, which is Oxygen for a spinel.

- Materials property (MP) dataset: Generating this dataset is more challenging. After discussing possibilities with NREL scientists, we considered two options. First, based on a suggestion by Haowei Peng (NREL), we considered the difference between the properties of a species and the one it was replacing (in the case of a substitution). The property was unchanged in the case of no impurity and set to 0.0 in case of a vacancy. However, this representation had several drawbacks, so we focused on using the ratios of properties instead of differences. Vacancies were represented by a property of zero and locations where the species did not change were represented by property values of 1.

Before we summarize our analysis results, we make an important observation on the datasets for this problem. As the data are generated using computationally expensive simulations, the number of compounds for each of the three materials is quite small. For Co2ZnO4, the dataset has 53 compounds, while the dataset for Rh2ZnO4 has 52 compounds and Mn2CrO4 has 49 compounds. The three datasets are analyzed separately as one of the questions we want to address is how much of the analysis results from one material carry over to the other. Given the small size of the datasets, we focused mainly on qualitative analysis as there were too few samples for a quantitative analysis. We also realized that based on the small number of compounds and the large number of features, our idea of using dimension reduction techniques or regression approaches would not have been successful, a fact that was confirmed when we tried to apply these techniques.

For the PT data for the three materials, we first ordered the materials based on increasing values of the formation enthalpy, and then tried to identify patterns that might lead to high or low formation enthalpy values. We made several observations on the three materials, including:

1. When the compounds are listed in increasing value of f-enthalpy, examples with higher f-enthalpy tend to have negative charge state, while those with lower enthalpy tend to have positive charge states

2. Vacancies tend to result in higher formation enthalpy than substitutions.

The analysis of the properties dataset for the three spinel materials is challenging, especially as it is unclear what is the best way to represent a compound in terms of the properties of its constituent atomic species; the sample size is quite small and the design space not adequately sampled; not all properties are available for all atomic species; and we need to represent compounds that differ only in the charge state, but may have very different formation enthalpy values. All this indicates that some information relevant to predicting the formation enthalpy may be missing. The crystal structure of a compound may play a role here, though converting the locations of the different atoms in 3-D space into relevant features is an open question.

Our attempts to find correlations between the features describing a compound and its formation enthalpy were inconclusive. We did find that, if we consider the compounds generated using a substitution, the difference between the formation enthalpy of two compounds with charge state 0 appears to be close to the difference between the formation enthalpy of the corresponding compounds in charge state -1. This may make it possible to predict one given the other three. The observation extends to other charge states as well, though there are too few examples to draw a

conclusion. However, the observation is not universally true and it may be worth investigating the compounds for which this is the case. In addition, the distance to the nearest neighbor for each compound could be used to add additional sample points as appropriate, resulting in a more complete coverage of the design space.

A detailed analysis of the three materials was communicated to the domain scientists and is available in a technical report (Section 6, item 5).

3.2.2 Analysis of the band-gap type dataset

The dataset considered in this analysis is from computer simulations of ternary compounds. Given the composition of a compound, that is, the three elements and their percentages in the compound, the simulations calculate various quantities, and associate with each compound a band gap type, which can take values 1, 2, 3, or 4. The primary focus of the analysis was to determine which properties of the elements are associated with band gap type 1 compounds. Any insights obtained on the other band gap types in the course of the analysis were also deemed to be of interest.

The data were provided in the form of a table, consisting of 487 compounds (also referred to as instances), each described by 83 features. Of these 83 features, 3 are the atomic species (i.e., elements) that make up the compound, referred to as "A", "B", and "C"; 3 are the compositions of each species (the values are all less than 1.0 and sum to 1.0), referred to as "p", "q", and "r"; and 3 are other variables ("E1", "E4", and "sg"). "sg" is the space group, reflecting the crystal symmetry and "E1" and "E4" are properties of the compound. Since our goal is to use the properties of the atomic species to predict the gap type of the compound, we ignore features E1 and E4. Of the remaining 74 features, 25 describe species A, 23 describe species B, and 26 describe species C. We also observe that the dataset does not include the same set of properties for each of the three species as some properties are unavailable for some elements.

Associated with each compound is the gaptype, which takes on values 1 through 4. Of the 487 instances, 76 are of gaptype 1 (15.6%), 100 of gaptype 2 (20.5%), 133 of gaptype 3 (27.3%), and 178 of gaptype 4 (36.6%). This is an unbalanced data set as the percentage of gap type 1 compounds is quite small. This can make it difficult to ascertain if an observation is a physics insight or simply the result of too few samples.

An analysis of the features indicated that, for a compound, each feature was generated by taking the value of a property of the element and weighting it by its composition. This meant that if a compound had a certain element, say Al, occurring at species "A", at a composition of 0.25, then the features corresponding to "A" for that compound would be the same as the "A" features for any other compound which also had Al occurring at species "A" at a composition of 0.25. This results in several repeating values, which can cause problems with some algorithms. It is also unclear if the features thus generated could represent the characteristics that determine the band gap type of a compound.

An initial exploratory analysis indicated that one compound had an incorrect space group - this was removed from the dataset. There were two pairs of instances which were exact duplicates - these were left in the dataset. We also found one inconsistency - there were five pairs of compounds which had the same values of "A", "B", "C", "p", "q", "r", and space group, but different band gap types (as well as different E1 and E4 values). Since "A", "B", "C", "p", "q", and "r" determine the features that are derived from the properties of the elements, this inconsistency indicates that some important features, which determine how the properties of the elements in a compound are related to the band gap type, are missing from the dataset.

Given the small number of instances with band gap type 1, the inconsistency in the dataset, and the approach used to generate the representation of each compound, we expect that a straightforward application of analysis tools is unlikely to yield insights into band gap type 1 compounds. This was confirmed when we analyzed the data using parallel plots - we did not find any of the features discriminating and there appeared to be not much difference between the compounds belonging to the four gap types. Further, as expected, a decision tree classifier was not be able to automatically classify compounds of band gap type 1 with a low error rate.

Given this, we considered three avenues for exploration - we reduced the number of features by identifying those which were correlated; we derived alternate representations by taking ratios and difference of properties that are available for all three species in a compound; and we tried to determine if, using the new representations, we could identify parts of feature space where a large percentage of points were likely to be of band gap type 1.

Our results indicated that we could not find a region of the space spanned by the original dataset which had a high percentage of type 1 compounds. This agrees with our observation that the original features are not very discriminating. However, if we remove the correlated features and features that are available for only one or two of the three species, we find one region where 45% of the compounds are of type 1. When we use ratios or differences of the original scaled or the unscaled properties, we obtain additional insights into regions with a greater than normal percentage of type 1 compounds. The rules identifying these regions also indicate which features are relevant to band gap type 1 compounds. As an example, the rule

```
BCRmelting_point < 0.581793 and  ACRmolar_volume < 0.396468
```

returns 50 compounds, of which 28 (=56%) are of band gap type 1. The two variables in the rule are the ratios of the melting points for species in locations "B" and "C" and the ratios of the molar volume for species in locations "A" and "C". Note that this is much higher than the 15% type 1 compounds in the full data set.

In this dataset, we also found that there is a correlation between the space group and the band gap type. For example, space group 122 usually results in type 1 or 2 compounds, while space group 198 leads to type 3 compounds and space group 2 favors type 4 compounds.

A detailed analysis of this dataset for all four band-gap types and the rules for each were communicated to the domain scientists and is summarized in a technical report (Section 6, item 6).

4 Conclusions and ideas for future work

In this project, we considered data mining techniques in the context of the inverse design of materials with desirable properties as semiconductors. The idea is to run an ensemble of computationally expensive simulations to guide physical experiments and gain insights into the design space which maps the inputs of the simulations to the output(s). To gain the greatest insights with a small number of simulations, we consider an incremental approach where we analyze the input/output data from the simulations that have already been run to intelligently identify the next set of simulations

There were two parts to the project. The first was the investigation of analysis techniques likely to be relevant to the task of identifying the inputs for new simulations and the second was to apply

these techniques to real datasets from the Center for Inverse Design EFRC. We considered two datasets - one related to the formation enthalpy and the other to the bad gap type.

Our investigation into dimension reduction techniques indicated that feature selection techniques tended to perform better on practical problems and resulted in lower dimensional representations that could be interpreted easily. However, our analysis also confirmed the fact that for these techniques to be useful, we need to have a reasonably large number of data points relative to the features that are used to represent each material.

The two datasets we analyzed were of relatively small size. In the case of the formation enthalpy dataset, there were few compounds created from each of the three spinel materials, while in the band gap type dataset, there were few examples of compounds with band gap type 1, even though the full dataset was moderate in size. The small sizes of the data were to be expected as the EFRC was in the early stages when this project was started. In light of the small sizes, we considered more qualitative approaches to gain insights into the data, which were communicated to the EFRC scientists.

Our overall conclusions from this study are:

- It is important to check the dataset for quality issues. Sometimes, we may find inconsistencies, indicating that some critical information has not been included.
- The representation of the data is important. This should be refined iteratively in the process of the analysis.
- If the datasets are small, as they are likely to be at the start of a design of experiments effort, sophisticated techniques from data mining might not be the best first choice for analysis. This is also true if the datasets are unbalanced because they were generated for some other analysis and are now being re-analyzed to gain new insights. In such cases, we should use simpler analysis techniques to see if we can gain qualitative insights into the data. In addition, we can experiment with data mining techniques to determine if they can shed some light on the data, for example, by indicating regions of feature space where there is a greater likelihood of finding materials with the desired properties.

5 Project team

The project team consisted of Ya Ju Fan, a post-doctoral fellow in the Center for Applied Scientific Computing who was hired for this effort, and Chandrika Kamath, both at LLNL.

6 Publications and presentations from this work

The following publications and presentations resulted from this work:

1. Poster Presentation: Ya Ju Fan and Chandrika Kamath. "A Comparison of Non-linear Techniques for Dimension Reduction". LLNL Postdoc Poster Symposium, June 1, 2011. LLNL-POST-484645.

2. Conference Presentation: Ya Ju Fan and Chandrika Kamath. "Intrinsic Dimensionality Using Non-linear Dimension Reduction Techniques". Institute for Operations Research and Management Sciences Annual Meeting, Charlotte, NC, November 13-16, 2011.

3. Conference Presentation: Ya Ju Fan and Chandrika Kamath. "Comparison of Dimensionality Reduction Techniques in Scientific Applications". SIAM Conference on Uncertainty Quantification, Raleigh, NC, April 2-5, 2012.

4. Technical Report: Ya Ju Fan and Chandrika Kamath. "On the Selection of Dimension Reduction Techniques for Scientific Applications". Accepted with minor revisions in *Annals of Information Systems*. February 2012. LLNL-TR-531131.

5. Technical Report: Chandrika Kamath, "Analysis of the formation enthalpy dataset," LLNL Technical Report LLNL-TR-582974, September 2012.

6. Technical Report: Chandrika Kamath, "Analysis of the band-gap type dataset," LLNL Technical Report LLNL-TR-577712, August 2012.

7. Book chapter: Chandrika Kamath and Ya Ju Fan, " Data Mining for Materials Science and Engineering," in preparation for the forthcoming book, *Informatics for Materials Science and Engineering,* edited by Prof. Krishna Rajan, to be published by Elsevier.

Acknowledgment

This work was done in collaboration with Mayeul d'Avezac and Alberto Franceschetti from NREL, who provided the data and the domain expertise for the formation enthalpy and band-gap type problems, respectively. Others from NREL, including Haowei Peng and Liping Yu, also provided responses to various questions.

I would like to thank William Tumas and Alex Zunger, the directors of the EFRC, as well as Lori Diachin from LLNL, for their support of this work. The work on analysis of formation enthalpy data for Co2ZNO4 and Rh2ZnO4, and the analysis of band-gap type 1 data, was funded by the SciDAC-e program in the Office of Science, US Department of Energy. The analysis of the Mn2CrO4 spinel material and the band-gap types 2, 3, and 4, were unfunded work.

LLNL-TR-583076: This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

References

[1] Christopher G. Atkeson, Andrew W. Moore Y, and Stefan Schaal Z. Locally weighted learning. *Artificial Intelligence Review*, 11:11–73, 1997.

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

[3] Matthew Brand. Charting a manifold. In *Advances in Neural Information Processing Systems 15*, pages 961–968. MIT Press, 2003.

[4] L. Breiman, J.H. Friedman, R. A. Olshen, and C.J. Stone. *Classification and Regression Trees*. CRC Press, Boca Raton, Florida, 1984.

[5] Center for Inverse Design web page, 2012. `http://www.centerforinversedesign.org/`.

[6] William S. Cleveland and Susan J. Devlin. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610, September 1988.

[7] K.-T. Fang, R. Li, and A. Sudjianto. *Design and Modeling for Computer Experiments*. Chapman and Hall/CRC Press, Boca Raton, FL, 2005.

[8] M. A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In *Proc. 17th International Conf. on Machine Learning*, pages 359–366. Morgan Kaufmann, San Francisco, CA, 2000.

[9] Balazs Kegl. Intrinsic dimension estimation using packing numbers. In *Advances in Neural Information Processing Systems: NIPS*, pages 681–688. MIT Press, December 2002.

[10] MINDES: Data Mining for Inverse Design Project web page, 2012. `https://computation.llnl.gov/casc/StarSapphire/MINDES.html`.

[11] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

[12] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[13] Marko Robnik-Šikonja and Igor Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Mach. Learn.*, 53:23–69, October 2003.

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

[15] Lawrence K. Saul, Sam T. Roweis, and Yoram Singer. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4:119–155, 2003.

[16] Leonard A. Smith. Intrinsic limits on dimension calculations. *Physics Letters A*, 133(6):283 – 288, 1988.

[17] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[18] J. B. Tenenbaum, V. de Silva, and J. C. Langford. The isomap algorithm and topological stability – response. *Science*, 295(5552):7, 2002.

[19] Gerard V. Trunk. Statistical estimation of the intrinsic dimensionality of a noisy signal collection. *Computers, IEEE Transactions on*, C-25(2):165 –171, February 1976.

[20] Laurens van der Maaten, Eirc Postma, and Jaap van den Herik. Dimensionality reduction: A comparative review. Technical Report TiCC TR 2009–005, Tilburg University, October 2009.

[21] Zhenyue Zhang and Hongyuan Zha. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *SIAM Journal of Scientific Computing*, 26:313–338, 2002.