Dimension Reduction: A Guided Tour

Dimension Reduction: A Guided Tour

Christopher J. C. Burges

Microsoft Research
One Microsoft Way
Redmond, WA 98052-6399
USA
chris.burges@microsoft.com

the essence of knowledge

Boston – Delft

Foundations and Trends® in Machine Learning

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is C. J. C. Burges, Dimension Reduction: A Guided Tour, Foundation and Trends® in Machine Learning, vol 2, no 4, pp 275–365, 2009

ISBN: 978-1-60198-378-7

Foundations and Trends® in Machine Learning
Volume 2 Issue 4, 2009

Editorial Board

Editor-in-Chief:
Michael Jordan
Department of Electrical Engineering and Computer Science
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720-1776

Editors

Editorial Scope

Foundations and Trends® in Machine Learning will publish survey and tutorial articles in the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians
Foundations and Trends® in Machine Learning, 2009, Volume 2, 4 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.

Foundations and Trends® in
Machine Learning
Vol. 2, No. 4 (2009) 275–365

DOI: 10.1561/2200000002

now
the essence of knowledge

Dimension Reduction: A Guided Tour

Christopher J. C. Burges

Microsoft Research, One Microsoft Way, Redmond, WA 98052-6399, USA, chris.burges@microsoft.com

Abstract

We give a tutorial overview of several foundational methods for dimension reduction. We divide the methods into projective methods and methods that model the manifold on which the data lies. For projective methods, we review projection pursuit, principal component analysis (PCA), kernel PCA, probabilistic PCA, canonical correlation analysis (CCA), kernel CCA, Fisher discriminant analysis, oriented PCA, and several techniques for sufficient dimension reduction. For the manifold methods, we review multidimensional scaling (MDS), landmark MDS, Isomap, locally linear embedding, Laplacian eigenmaps, and spectral clustering. Although this monograph focuses on foundations, we also provide pointers to some more modern techniques. We also describe the correlation dimension as one method for estimating the intrinsic dimension, and we point out that the notion of dimension can be a scale-dependent quantity. The Nyström method, which links several of the manifold algorithms, is also reviewed. We use a publicly available data set to illustrate some of the methods. The goal is to provide a self-contained overview of key concepts underlying many of these algorithms, and to give pointers for further reading.

Contents

1

Introduction

Dimension reduction[1] is the mapping of data to a lower dimensional space such that uninformative variance in the data is discarded, or such that a subspace in which the data lives is detected. Dimension reduction has a long history as a method for data visualization, and for extracting key low dimensional features (for example, the two-dimensional orientation of an object, from its high dimensional image representation). In some cases the desired low dimensional features depend on the task at hand. Apart from teaching us about the data, dimension reduction can lead us to better models for inference. The need for dimension reduction also arises for other pressing reasons. Stone [85] showed that, under certain regularity assumptions (including that the samples be IID), the optimal rate of convergence[2] for nonparametric regression varies

[1] We follow both the lead of the statistics community and the spirit of the paper to reduce 'dimensionality reduction' and 'dimensional reduction' to 'dimension reduction'.

[2] The definition of 'optimal rate of convergence' is technical and for completeness we reproduce Stone's definitions here [85]. A 'rate of convergence' is defined as a sequence of numbers, indexed by sample size. Let θ be the unknown regression function, Θ the collection of functions to which θ belongs, $\hat{T}_n$ an estimator of θ using n samples, and $\{b_n\}$ a sequence of positive constants. Then $\{b_n\}$ is called a lower rate of convergence if there exists $c > 0$ such that $\lim_n \inf_{\hat{T}_n} \sup_\Theta P(\|\hat{T}_n - \theta\| \geq cb_n) = 1$, and it is called an achievable rate of convergence if there is a sequence of estimators $\{\hat{T}_n\}$ and $c > 0$ such that

as $m^{-p/(2p+d)}$, where m is the sample size, the data lies in $\mathcal{R}^d$, and where the regression function is assumed to be p times differentiable. We can get a very rough idea of the impact of sample size on the rate of convergence as follows. Consider a particular point in the sequence of values corresponding to the optimal rate of convergence: $m = 10,000$ samples, for $p = 2$ and $d = 10$. Suppose that d is increased to 20; what number of samples in the new sequence gives the same value? The answer is approximately 10 million. If our data lies (approximately) on a low dimensional manifold $\mathcal{L}$ that happens to be embedded in a high dimensional manifold $\mathcal{H}$, then modeling the data directly in $\mathcal{L}$ rather than in $\mathcal{H}$ may turn an infeasible problem into a feasible one.

The purpose of this monograph is to describe the mathematics and key ideas underlying the methods, and to provide some links to the literature for those interested in pursuing a topic further.[3] The subject of dimension reduction is vast, so we use the following criterion to limit the discussion: we restrict our attention to the case where the inferred feature values are continuous. The observables, on the other hand, may be continuous or discrete. Thus this review does not address clustering methods, or, for example, feature selection for discrete data, such as text. This still leaves a very wide field, and so we further limit the scope by choosing not to cover probabilistic topic models (in particular, latent Dirichlet allocation, nonnegative matrix factorization, probabilistic latent semantic analysis, and Gaussian process latent variable models). Furthermore, implementation details, and important theoretical details such as consistency and rates of convergence of sample quantities to their population values, although important, are not discussed. For an alternative, excellent overview of dimension reduction methods, see Lee and Verleysen [62]. This monograph differs from that work in several ways. In particular, while it is common in the literature to see methods applied to artificial, low dimensional data sets such as the famous Swiss Roll, in this monograph we prefer to use higher dimensional data: while low dimensional toy data can be valuable to

$\lim_n \sup_\Theta P(\|\hat{T}_n - \theta\| \geq c b_n) = 0$; $\{b_n\}$ is called an optimal rate of convergence if it is both a lower rate of convergence and an achievable rate of convergence. Here the $\inf_{\hat{T}_n}$ is over all possible estimators $\hat{T}_n$.

[3] This monograph is a revised and extended version of Burges [17].

express ideas and to illustrate strengths and weaknesses of a method, high dimensional data has qualitatively different behavior from two- or three-dimensional data. Here, we use the publicly available KDD Cup [61] training data. This is anonymized breast cancer screening data for 1,712 patients, 118 of whom had a malignant cancer; each feature vector has 117 features, and a total of 102,294 such samples are available. The goal of the Cup was to identify those patients with a malignant tumor from the corresponding feature vectors in a test set. We use the data here because it is relevant to an important real-world problem, it is publicly available, and because the training data has labels (some of the techniques we describe below are for supervised problems).

Regarding notation: we denote the sample space (the high dimensional space in which the data resides) as $\mathcal{H}$, the low dimensional space (to which many of the methods discussed below map the data) as $\mathcal{L}$, and we reserve $\mathcal{F}$ to denote a *feature space* (often a high or infinite-dimensional Hilbert space, to which the kernel versions of the methods below map the data as an intermediate step). Vectors are denoted by boldface, whereas components are denoted by x_a, or by $(\mathbf{x}_i)_a$ for the a-th component of the i-th vector. Random variables are denoted by upper case; we use $E[X|y]$ as shorthand for the function $E[X|Y = y]$, in contrast to the random variable $E[X|Y]$. Following Horn and Johnson [54], the set of p by q matrices is denoted M_{pq}, the set of (square) p by p matrices by M_p, the set of symmetric p by p matrices by S_p, and the set of (symmetric) positive semidefinite matrices by S_p^+ (all matrices considered are real). $\mathbf{e}$ with no subscript is used to denote the vector of all ones; on the other hand $\mathbf{e}_a$ denotes the a-th eigenvector. We denote sample size by m, and dimension usually by d or d', with typically $d' \ll d$. δ_{ij} is the Kronecker delta (the ij-th component of the unit matrix).

We place dimension reduction techniques into two broad categories: methods that rely on projections (Section 3) and methods that attempt to model the manifold on which the data lies (Section 4). Section 3 gives a detailed description of principal component analysis; apart from its intrinsic usefulness, PCA is interesting because it serves as a starting point for many modern algorithms, some of which (kernel PCA,

probabilistic PCA, and oriented PCA) are also described here. However, it has clear limitations: it is easy to find even low dimensional examples where the PCA directions are far from optimal for feature extraction [33], and PCA ignores correlations in the data that are higher than second order. We end Section 3 with a brief look at projective methods for dimension reduction of *labeled* data: sliced inverse regression, and kernel dimension reduction. Section 4 starts with an overview of the Nyström method, which can be used to extend, and link, several of the algorithms described in this monograph. We then examine some methods for dimension reduction which assume that the data lies on a low dimensional manifold embedded in a high dimensional space, namely locally linear embedding, multidimensional scaling, Isomap, Laplacian eigenmaps, and spectral clustering.

Before we begin our exploration of these methods, however, let's investigate a question that is more fundamental than, and that can be explored independently of, any particular dimension reduction technique: if our data lives on a manifold $\mathcal{M}$ that is embedded in some Euclidean space, how can we estimate the dimension of $\mathcal{M}$?

References

[1] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, "A learning algorithm for Boltzmann machines," *Cognitive Science*, vol. 9, 1985.

[2] M. A. Aizerman, E. M. Braverman, and L. I. Rozoner, "Theoretical foundations of the potential function method in pattern recognition learning," *Automation and Remote Control*, vol. 25, pp. 821–837, 1964.

[3] S. Akaho, "A kernel method for canonical correlation analysis," in *Proceedings of the International Meeting of the Psychometric Society (IMPS2001)*, 2001.

[4] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics, 2003.

[5] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *Journal of Machine Learning Research*, vol. 3, pp. 1–48, 2002.

[6] P. F. Baldi and K. Hornik, "Learning in linear neural networks: A survey," *IEEE Transactions on Neural Networks*, vol. 6, pp. 837–858, July 1995.

[7] A. Basilevsky, *Statistical Factor Analysis and Related Methods*. Wiley, New York, 1994.

[8] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, pp. 1373–1396, 2003.

[9] Y. Bengio, J. Paiement, and P. Vincent, "Out-of-sample extensions for LLE, Isomap, MDS, Eigenmaps and spectral clustering," in *Advances in Neural Information Processing Systems 16*, MIT Press, 2004.

[10] C. Berg, J. P. R. Christensen, and P. Ressel, *Harmonic Analysis on Semigroups*. Springer-Verlag, 1984.

[11] C. M. Bishop, "Bayesian PCA," in *Advances in Neural Information Processing Systems 11*, MIT Press, 1999.

[12] I. Borg and P. Groenen, "Modern Multidimensional Scaling: Theory and Applications," Springer, 1997.
[13] B. E. Boser, I. M. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *Fifth Annual Workshop on Computational Learning Theory*, pp. 144–152, ACM, Pittsburgh, 1992.
[14] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
[15] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.
[16] C. J. C. Burges, "Some notes on applied mathematics for machine learning," in *Advanced Lectures on Machine Learning*, (O. Bousquet, U. von Luxburg, and G. Rätsch, eds.), pp. 21–40, Springer Lecture Notes in Artificial Intelligence, 2004.
[17] C. J. C. Burges, "Geometric methods for feature selection and dimensional reduction," in *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, (L. Rokach and O. Maimon, eds.), Kluwer Academic, 2005.
[18] C. J. C. Burges, "Simplified support vector decision rules," in *Proceedings of the Thirteenth International Conference on Machine Learning*, pp. 71–77, 1996.
[19] C. J. C. Burges, J. C. Platt, and S. Jana, "Extracting noise-robust features from audio," in *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, pp. 1021–1024, IEEE Signal Processing Society, 2002.
[20] C. J. C. Burges, J. C. Platt, and S. Jana, "Distortion discriminant analysis for audio fingerprinting," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 165–174, 2003.
[21] F. R. K. Chung, *Spectral Graph Theory*. American Mathematical Society, 1997.
[22] R. D. Cook, *Regression Graphics*. Wiley, 1998.
[23] R. D. Cook, "Model based sufficient dimension reduction for regression," Isaac Newton Institute Lectures on Contemporary Frontiers in High-Dimensional Statistical Data Analysis, http://www.newton.ac.uk/webseminars/pg+ws/2008/sch/schw01/0108/cook/, 2008.
[24] R. D. Cook and L. Forzani, "Likelihood-based sufficient dimension reduction," *Journal of the American Statistical Association*, vol. 104, pp. 197–208, 2009.
[25] R. D. Cook and H. Lee, "Dimension reduction in binary response regression," *Journal of the American Statistical Association*, vol. 94, pp. 1187–1200, 1999.
[26] R. D. Cook and S. Weisberg, "Sliced inverse regression for dimension reduction: Comment," *Journal of the American Statistical Association*, vol. 86, pp. 328–332, 1991.
[27] T. F. Cox and M. A. A. Cox, *Multidimensional Scaling*. Chapman and Hall, 2001.
[28] R. B. Darlington, "Factor analysis," Technical report, Cornell University, http://comp9.psych.cornell.edu/Darlington/factor.htm, 1997.

[29] V. De Silva and J. B. Tenenbaum, "Global versus local methods in nonlinear dimensionality reduction," *Advances in Neural Information Processing Systems*, vol. 15, pp. 705–712, MIT Press, 2002.
[30] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society B*, vol. 39, pp. 1–22, 1977.
[31] P. Diaconis and D. Freedman, "Asymptotics of graphical projection pursuit," *Annals of Statistics*, vol. 12, pp. 793–815, 1984.
[32] K. I. Diamantaras and S. Y. Kung, *Principal Component Neural Networks*. Wiley, 1996.
[33] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. Wiley, 1973.
[34] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the Nyström method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, 2004.
[35] J. H. Friedman and W. Stuetzle, "Projection pursuit regression," *Journal of the American Statistical Association*, vol. 76, pp. 817–823, 1981.
[36] J. H. Friedman, W. Stuetzle, and A. Schroeder, "Projection Pursuit density estimation," *Journal of the American Statistical Association*, vol. 79, pp. 599–608, 1984.
[37] J. H. Friedman and J. W. Tukey, "A projection pursuit algorithm for exploratory data analysis," *IEEE Transactions on Computers*, vol. 23, pp. 881–890, 1974.
[38] K. Fukumizu, F. R. Bach, and M. I. Jordan, "Kernel dimension reduction in regression," *Annals of Statistics*, vol. 37, pp. 1871–1905, 2009.
[39] A. Globerson and N. Tishby, "Sufficient dimensionality reduction," *Journal of Machine Learning Research*, vol. 3, 2003.
[40] J. Goldberger, S. T. Roweis, G. E. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," *Advances in Neural Information Processing Systems*, vol. 17, 2005.
[41] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Johns Hopkins, 3rd ed., 1996.
[42] M. Gondran and M. Minoux, *Graphs and Algorithms*. Wiley, 1984.
[43] P. Grassberger and I. Procaccia, "Measuring the strangeness of strange attractors," *Physica*, vol. 9D, pp. 189–208, 1983.
[44] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, "Measuring statistical dependence with Hilbert-Schmidt norms," in *Algorithmic Learning Theory, Springer Lecture Notes in Computer Science*, vol. 3734, pp. 63–77, 2005.
[45] G. Grimmet and D. Stirzaker, *Probability and Random Processes*. Oxford University Press, 3rd ed., 2001.
[46] J. Ham, D. D. Lee, S. Mika, and B. Schölkopf, "A kernel view of dimensionality reduction of manifolds," in *Proceedings of the International Conference on Machine Learning*, 2004.
[47] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 12, pp. 2639–2664, 2004.

[48] T. J. Hastie and W. Stuetzle, "Principal curves," *Journal of the American Statistical Association*, vol. 84, pp. 502–516, 1989.
[49] N. J. Higham, "Computing the nearest symmetric positive semidefinite matrix," *Linear Algebra and its Applications*, vol. 103, pp. 103–118, 1988.
[50] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, pp. 1771–1800, 2002.
[51] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 7, pp. 1527–1554, 2006.
[52] G. E. Hinton and S. E. Roweis, "Stochastic neighbor embedding," *Advances in Neural Information Processing Systems*, vol. 14, pp. 833–840, 2002.
[53] G. E. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504–507, 2007.
[54] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.
[55] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1936.
[56] T. Hsing and H. Ren, "An RKHS formulation of the inverse regression dimension-reduction problem," *Annals of Statistics*, vol. 37, pp. 726–755, 2009.
[57] P. J. Huber, "Projection pursuit," *Annals of Statistics*, vol. 13, pp. 435–475, 1985.
[58] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley, 2001.
[59] T. L. Kelley, *Crossroads in the Mind of Man: A study of Differentiable Mental Abilities*. Stanford University Press, 1928.
[60] G. S. Kimeldorf and G. Wahba, "Some results on Tchebycheffian spline functions," *Journal of Mathematical Analysis and Applications*, vol. 33, pp. 82–95, 1971.
[61] Knowledge Discovery and Data Mining Cup, http://www.kddcup2008.com/index.html, 2008.
[62] J. Lee and M. Verleysen, *Nonlinear Dimensionality Reduction*. Springer, 2007.
[63] B. Li, H. Zha, and F. Chiaromonte, "Contour regression: A general approach to dimension reduction," *The Annals of Statistics*, vol. 33, pp. 1580–1616, 2005.
[64] C.-K. Li, "Sliced Inverse Regression for dimension reduction," *Journal of the American Statistical Association*, vol. 86, pp. 316–327, 1991.
[65] C.-K. Li, "On Principal Hessian Directions for data visualization and dimension reduction: Another application of Stein's lemma," *Journal of the American Statistical Association*, vol. 87, pp. 1025–1039, 1992.
[66] M. Meila and J. Shi, "Learning segmentation by random walks," *Advances in Neural Information Processing Systems*, vol. 12, pp. 873–879, 2000.
[67] S. Mika, B. Schölkopf, A. J. Smola, K.-R. Müller, M. Scholz, and G. Rätsch, "Kernel PCA and de-noising in feature spaces," *Advances in Neural Information Processing Systems*, vol. 11, MIT Press, 1999.
[68] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in Neural Information Processing Systems*, vol. 14, MIT Press, 2002.

[69] J. Nilsson, F. Sha, and M. I. Jordan, "Regression on manifolds using kernel dimension reduction," in *Proceedings of the 24th International Conference on Machine Learning*, 2007.

[70] J. C. Platt, "Fastmap, MetricMap, and landmark MDS are all Nyström algorithms," in *Proceedings of the 10th International Conference on Artificial Intelligence and Statistics*, 2005.

[71] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vettering, *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 2nd ed., 1992.

[72] A. Rahimi and B. Recht, "Clustering with normalized cuts is clustering with a hyperplane," Workshop on Statistical Learning in Computer Vision, Prague, 2004.

[73] S. M. Ross, *Introduction to Probability Models*. Academic Press, 10th ed., 2010.

[74] S. M. Ross and E. A. Peköz, *A Second Course in Probability*. www.Probability Bookstore.com, Boston, MA, 2007.

[75] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

[76] L. K. Saul and S. T. Roweis, "Think globally, fit locally: Unsupervised learning of low dimensional manifolds," *Journal of Machine Learning Research*, vol. 4, pp. 119–155.

[77] L. K. Saul, K. Q. Weinberger, J. H. Ham, F. Sha, and D. D. Lee, "Spectral methods for dimensionality reduction," in *Semisupervised Learning*, (O. Chapelle, B. Schölkopf, and A. Zien, eds.), pp. 293–308, MIT Press, 2006.

[78] I. J. Schoenberg, "Remarks to Maurice Frechet's article *Sur la définition axiomatique d'une classe d'espace distanciés vectoriellement applicable sur l'espace de hilbert*," *Annals of Mathematics*, vol. 36, pp. 724–732, 1935.

[79] B. Schölkopf, "The kernel trick for distances," *Advances in Neural Information Processing Systems*, vol. 13, pp. 301–307, MIT Press, 2001.

[80] B. Schölkopf and A. Smola, *Learning with Kernels*. MIT Press, 2002.

[81] B. Schölkopf, A. Smola, and K.-R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, pp. 1299–1319, 1998.

[82] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 888–905, 2000.

[83] C. E. Spearman, "'General intelligence' objectively determined and measured," *American Journal of Psychology*, vol. 5, pp. 201–293, 1904.

[84] I. Steinwart, "On the influence of the kernel on the consistency of support vector machines," *Journal of Machine Learning Research*, vol. 37, pp. 726–755, 2001.

[85] C. J. Stone, "Optimal global rates of convergence for nonparametric regression," *Annals of Statistics*, vol. 10, pp. 1040–1053, 1982.

[86] J. Sun, S. Boyd, L. Xiao, and P. Diaconis, "The fastest mixing Markov process on a graph and a connection to a maximum variance unfolding problem,"

Society for Industrial and Applied Mathematics (SIAM) Review, vol. 48, pp. 681–699, 2006.

[87] J. B. Tenenbaum, "Mapping a manifold of perceptual observations," *Advances in Neural Information Processing Systems*, vol. 10, MIT Press, 1998.

[88] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Computation*, vol. 11, pp. 443–482, 1999.

[89] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society*, vol. 61, p. 611, 1999.

[90] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377, 1999.

[91] L. van der Maaten, "t-Distributed Stochastic Neighbor Embedding," http://homepage.tudelft.nl/19j49/t-SNE.html.

[92] L. van der Maaten, "Learning a parametric embedding by preserving local structure," in *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2009.

[93] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[94] L. van der Maaten, E. Postma, and J. van den Herik, "Dimensionality reduction: a comparative review," Tilburg University Technical Report TiCC-TR 2009–005, 2009.

[95] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[96] K. Q. Weinberger and L. K. Saul, "An introduction to nonlinear dimensionality reduction by maximum variance unfolding," in *Proceedings of the Twenty First National Conference on Artificial Intelligence (AAAI-06)*, pp. 1683–1686, 2006.

[97] S. Wilks, *Mathematical Statistics*. Wiley, 1962.

[98] C. K. I. Williams, "On a connection between kernel PCA and metric multidimensional scaling," *Advances in Neural Information Processing Systems*, vol. 13, pp. 675–681, MIT Press, 2001.

[99] C. K. I. Williams and M. Seeger, "Using the Nyström method to speed up kernel machines," *Advances in Neural Information Processing Systems*, vol. 13, pp. 682–688, MIT Press, 2001.

[100] Z. Zhang and M. I. Jordan, "Multiway spectral clustering: A margin-based perspective," *Statistical Science*, vol. 23, pp. 383–403, 2008.