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

Full text available at: http://dx.doi.org/10.1561/2200000002
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.
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 1

2 Estimating the Dimension 5
   2.1 A Cautionary Note 6
   2.2 Empirical Investigation 8

3 Projective Methods 13
   3.1 Independent Component Analysis 15
   3.2 Principal Component Analysis (PCA) 17
   3.3 Probabilistic PCA (PPCA) 24
   3.4 The Kernel Trick 27
   3.5 Kernel PCA 29
   3.6 Canonical Correlation Analysis 33
   3.7 Linear Discriminant Analysis 40
   3.8 Oriented PCA and Distortion Discriminant Analysis 42
   3.9 Sufficient Dimension Reduction 45

4 Manifold Modeling 57
   4.1 The Nyström Method 57
   4.2 Multidimensional Scaling 61
   4.3 Isomap 68
   4.4 Locally Linear Embedding 69
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 \cite{stone} 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 \cite{stone}. A ‘rate of convergence’ is defined as a sequence of numbers, indexed by sample size. Let \(\theta\) be the unknown regression function, \(\Theta\) the collection of functions to which \(\theta\) belongs, \(\hat{T}_n\) an estimator of \(\theta\) 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 \(\lim_n \inf_{\hat{T}_n} \sup_{\theta \in \Theta} P(\|\hat{T}_n - \theta\| \geq cb_n) = 1\), 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. 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

\[ \lim_{n \to \infty} \sup_{\Theta} P(\|\hat{T}_n - \theta\| \geq cb_n) = 0; \]  
\{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 $\hat{T}_n$.

\footnote{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 $(x_i)_a$ 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). $e$ with no subscript is used to denote the vector of all ones; on the other hand $e_a$ denotes the $a$-th eigenvector. We denote sample size by $m$, and dimension usually by $d$ or $d'$, with typically $d' \ll d$. $\delta_{ij}$ 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


References


References

References


References


