ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

Unsupervised dimensionality reduction: the challenges of big data visualisation

Kerstin Bunte[1] and John Aldo Lee[2,1] *

1- Université catholique de Louvain - ICTEAM Institute - MLG
Place du Levant 3, B-1348-Louvain-la-Neuve - Belgium

2- Université catholique de Louvain - IREC Institute - MIRO
Avenue Hippocrate 55, B-1200 Bruxelles - Belgium

Abstract. Dimensionality reduction is an unsupervised task that allows high-dimensional data to be processed or visualised in lower-dimensional spaces. This tutorial reviews the basic principles of dimensionality reduction and discusses some of the approaches that were published over the past years from the perspective of their application to big data. The tutorial ends with a short review of papers about dimensionality reduction in these proceedings, as well as some perspectives for the near future.

1 Introduction

Dimensionality reduction (DR) [1, 2] aims at representing high-dimensional data in low-dimensional spaces, while preserving important structural properties, like for example (dis)similarities or neighbourhood relationships. The vast majority of DR methods work in a unsupervised way: they process data features without taking into account additional information like class labels, which are then sometimes used to assess DR quality. Dimensionality reduction can be used for different purpose, ranging from exploratory data analysis (visual inspection) to data compression before subsequent processing. In the latter case, DR can be seen as a way to defeat the so-called curse of dimensionality, which makes many complex analysis tasks like regression or classification much more difficult in high-dimensional spaces than in low-dimensional ones.

High dimensionality is not the only issue that analysts have to face in the current era of data plethora. Data collection and storage becomes easier and cheaper every day. Processing large amounts of data raises many issues, in terms of algorithmic complexity (time and memory consumption), workload distribution (vectorised, parallel, or distributed architectures), and efficient visual presentation of the results. Politics and media have coined the term “Big Data” to refer to these problems and the effort to alleviate them. In a recent interview for the INNS Big Data conference, though, Jurgen Schmidhüber said: “At any given moment, *big data* is more data than most people can conveniently store”. Thereby he pointed out nicely that big data was, is, and will always remain an open question, although it only became popular very recently.

This tutorial revisits past and recent history of DR and briefly shows how the issue of large data sets has been dealt with. Section 2 introduces the various principles of DR and weaves connections with the closely related domain

*J.A.L. and K.B. are funded by the Belgian F.R.S.-FNRS (project T.0175.13 ‘DRedVis’).

of information visualisation (InfoVis). Section 3 comes back on approaches and methods that were published in the recent and more distant past, with a special focus on how they tried to deal with big data. Section 4 briefly presents the contributions to DR in these proceedings. Finally, Section 5 sketches the expected forthcoming developments of DR.

2 Principles

In general, DR attempts to represent high dimensional data with low-dimensional counterparts while preserving as much "information" as possible. Depending on the user, data at hand, and the problem to solve, various aspects may be relevant. Therefore, DR is inherently ill-posed and a many methods have been proposed, differing in the data properties they preserve, the mathematical formulation (parametric or non-parametric, discriminative or generative), the optimisation scheme, and so on. First steps have been taken to unify a vast majority of DR techniques in a general framework [3] by summarizing a general principle. Let us assume high-dimensional data points $\Xi = \{\boldsymbol{\xi}_i \in \mathrm{I\!R}^D | i = 1 \dots N\}$ have low-dimensional counterparts $\boldsymbol{x}_i$ in the embedding space $\mathrm{I\!R}^d$, with $d \in \{2, 3\}$ for visualisation. Generally the following building blocks are used: the characteristics derived from the original data set Ξ for every data point char_Ξ, corresponding characteristics of their projection $\mathrm{char}_\mathcal{X}$, and an error measure between them. Therefore a cost function is formulated that is minimised during projection: $\mathrm{costs} := \sum_{\boldsymbol{x}_i \in X} \mathrm{error}(\mathrm{char}_\Xi, \mathrm{char}_\mathcal{X})$, possibly constrained to guarantee the uniqueness or invariance of the result.

Models in DR can also be split in *generative* and *discriminative* [4]. These ideas can be simply summarised by *feature construction* versus *feature selection*, inspired by Bayesian or frequentist philosophy [5, 6]. Discriminative dimension reduction usually consists in a discretised representation of the input space. Supervised methods aim to extract features from the data most descriptive for a certain task (regression, classification, etc.) and will not be dealt with in this tutorial. Early unsupervised techniques include principal component analysis (PCA) [7], multidimensional scaling (MDS) [8, 9] and the self organizing map (SOM) [10]. The latter abstracts a neural network with lateral connections, modeled with a neighbourhood function to preserve the topological properties of the data. Whereas discriminative learning provide models only for the target variables conditioned on the observed data generative learning provide a full probabilistic model on all variables. Hence generative models can be used to simulate or *generate* values of any variable in its model to find a lower-dimensional parameterisation of the dataset. The model selection procedure usually involves the maximisation of the log-likelihood of the model for example by expectation maximisation (EM). Discriminative models often have an equivalent Bayesian formulation, like for example SOM has generative topographic mapping (GTM) [11], and are seen as complementary or different views of the same procedure.

Many traditional DR techniques are parametric with a functional form of the mapping that is explicit and fixed a priori: $f_W : \mathrm{I\!R}^D \rightarrow \mathrm{I\!R}^d$, $\boldsymbol{\xi} \rightarrow \boldsymbol{x} = f_W(\boldsymbol{\xi})$

and function parameters W are optimised during DR. This approach has several benefits: (1) out-of-sample extensions are immediate and require only constant time depending on the chosen form of the mapping; (2) inverse mappings may be possible (for example locally linear function can be inverted using the pseudo-inverse); (3) an implicit regularisation takes place depending on the form of the mapping function if only a few parameters need to be determined; and (4) usually only a few data points are necessary to determine the mapping parameters, which generalise to new points. Therefore, a representative subset of the full data is sufficient for training, which increases the speed and feasibility of computation for very large data sets and the investigation of the generalisation abilities with respect to new points is possible.

On the other hand, many modern techniques are non-parametric, which means that the coordinates of single point projections $\boldsymbol{\xi}_i \rightarrow \boldsymbol{x}_i$ are optimised directly. They have the advantage, that they are not confined to a restricted functional form and therefore highly non-linear and complex embeddings are possible. Spectral DR methods [12] rely on the spectrum of the neighbourhood graph as data characteristics and preserve important properties of it. In general, the mathematical objective is formulated to exhibit a unique algebraic solution and therefore these methods often base on very simple affine functions like Gaussians. Hence, they may show inferior results for boundaries, disconnected manifolds or holes. Using more complex affinities, such as geodesic distance or local neighbourhoods, can avoid those problems at the price of higher computational costs and the existence of local optima. Usually numerical optimisation is required, but due to the greater complexity their visualisation properties may be superior. Due to the mapping of a given finite set of points additional effort is necessary to include new points in the mapping. Naively out-of-sample extension can be provided if the novel points are mapped by optimizing the underlying cost function while keeping the prior data fixed. This way the new coordinates are still controlled by all data points and the computational costs depend on the size of the training set, but the generalisation ability is not clear. Therefore, a principled framework to extend non-parametric methods with explicit mapping function to combine the strength of both formulations has been shown in [3].

Interactive data visualisation has gained a lot of interest since it has great potential to engage and inform large audiences, especially since the availability of the world wide web. Therefore, considerable effort has been made to identify, express, and understand the complex concepts of human-machine communication, for example Norman's execution-evaluation cycle with its seven states of action [13]. This led to the establishment of common interaction techniques in InfoVis like, for example, *brushing*, *zooming & panning*, and *dynamic queries*. Data visualisation and user interaction have been a strong focus in InfoVis community and attempts are taken to combine it with the strong mathematical formulations and machine learning principles known from the dimensionality reduction field. Recent methods interactively select a subset of the potentially most interesting variables, employing various methods for dimensionality reduction under changes of the metric driven by the user [14, 15].

3 Approaches

Early DR methods were linear transformations of data, like principal component analysis (PCA) [7] and classical metric multidimensional scaling (MDS) [8, 9]. These two methods are dual, the former working with the eigenvalue decomposition (EVD) of the sample D-by-D covariance matrix, the latter with the EVD of the n-by-n Gram matrix of dot products. On the other hand, PCA and MDS totally differ on key aspects: PCA is parametric, MDS is not. Also, depending on whether $D \ll N$ or $D \gg N$, PCA or MDS will scale more favourably in terms of memory and time consumption.

Many nonlinear variants of MDS [8, 9, 16, 17, 18] rely on a *stress* function that quantifies distance preservation (instead of dot products in classical metric MDS). Stress is minimised with generic optimisation techniques (instead of an EVD). The stress formulation make these methods nonparametric and not suited for large N. Some publications have tried to address this issue by reducing first the data set size, for instance with vector quantisation (VQ) [19]. The drawback is naturally that not all data points are represented, unless an efficient out-of-sample extension is available too (see for instance [20, 21]). Such a strategy relying on VQ was in the same spirit as the very popular auto-organising maps [10], which elegantly combined vector quantisation and DR/visualisation.

A related approach, at least in its biological inspiration, is DR with auto-encoders (AEs), namely, feed-forward artificial neural networks working in auto-association mode (outputs must match inputs) [22, 23, 24, 25]. AEs are deep networks and consist of at least three hidden layers, with the middle one including d neurons. Hence, DR occurs in the first half of the network, while the second reconstruct data. Being parametric, AEs also have a controllable model complexity, scale rather well for both D and N, and have a straightforward out-of-sample extension. The principle of AEs is much more elegant than that of the hybrid approaches in [20, 21]. However, early AEs could only rely on back-propagation and suffered thus from inefficient learning, until the development of specific training techniques [25].

Weighted distance preservation [16, 17, 18] in a stress function is not to only way to generalise classical metric MDS to nonlinear DR. A preliminary nonlinear transformation can send data to a feature space, where classical MDS can be carried out with the same guarantee of identifying a global optimum thanks to the EVD. Kernel PCA [26] implements this principle with Mercer kernels, which transpose dot products from the data space to some (unknown) feature space [27, 28]. Many other DR methods have followed this pioneering idea and are now known as spectral DR [12]. For instance, Isomap [29] and maximum variance unfolding (MVU) [30] rely on fixed or adaptive geodesic distances (instead of Euclidean ones) to induce the feature space. Laplacian eigenmaps (LE) [31] and locally linear embedding (LLE) [32] harness graph theory and indirectly involve random walks and commute time distance like in spectral clustering [33, 34]. Isomap and MVU entail Gram matrices and do not scale very well with N. In contrast, LE and LLE involve only adjacency or affinity matrices that are advantageously sparse. Like classical MDS it derives from, spectral DR is non-parametric and requires out-of-sample extensions to process large data sets [35].

The latest and most promising DR methods are variants of stochastic neighbour embedding (SNE) [36], like t-SNE [37], NeRV [38], and Jensen-Shannon embedding [39]. In spirit, these methods are close to stress-based MDS, except that they replace distances with specific pairwise similarities/affinities that are quite robust against distance concentration [40, 41]. If these methods are very successful in terms of DR quality [37, 38, 39], especially for very high-dimensional data, they do not scale well for large N, since they involves pairwise similarities. An early but rather inefficient workaround [37] has relied on $L \ll N$ landmarks, which are randomly picked in the data sets. Instead of all possible pairs of data, the method considers only data-landmarks pairs. In contrast, the latest developments borrow ideas from astronomy and mechanics, domains where large N-body problems are solved approximately. Data structures like quadtrees [42] or fast multipole methods [43] allow reducing the time complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(N \log(N))$ or even $\mathcal{O}(N)$. These accelerated methods can process tens of thousands of data in minutes, but the quality of their results still need further quantitative assessment.

4 Contributions in these ESANN proceedings

The special session about unsupervised DR includes six contributions.

Payen et al. and Delion et al. propose two applicative papers involving recent nonlinear DR methods. The former investigates efficient clustering for spatial bird population analysis along the Loire river. They analyse the spatio-temporal distribution of bird communities to study river zonation by detecting ecological discontinuities due to geomorphology of landscapes using quantitative evaluation based on neighbourhood ranking. The latter contribution compares nonlinear DR techniques for high-dimensional Near InfraRed Spectroscopy (NIRS) data for vineyard soil characterisation. They analyse double variability, namely the inter-specific due to similar sites and the intra-specific with respect to the sample, using fractional metrics.

Gianniotis et al. present their approach to visualisation of time series data. They employ an echo state network (ESN) with fixed reservoir to capture the long-term latent dynamics and convert the time series into vector representation by training a linear readout vector. Visualisation is then constructed using the bottleneck activations of an AE. The core of their contribution is the definition of an objective function that quantifies the reconstruction error in a meaningful way, namely, how well the reconstructed readout vector can reproduce the time series when plugged into the same fixed ESN reservoir.

Alaíz et al. propose a measure of neighbourhood preservation to fix parameters for diffusion maps without requiring problem-specific knowledge. It assumes that a certain diffusion metric can approximate the metric of the low-dimensional Riemannian manifold. They show next that the list of varying parameters can be chosen depending on a neighbourhood preservations measure and that a linear relation binds the fitness criteria and the model accuracy.

Blöbaum's et al. contribution investigates unsupervised DR for transfer learning, which establish a link of source (training) and target (test) domain by representing data in a common latent space. Transfer learning is currently a hot topic in the context of big data, distributed systems and life-long learning, where source and target data might follow a different underlying distribution or is contained in a different spaces. A shared distribution of source and target data in the latent space can be enforced by EM optimisation of the log-likelihood and employ modern non-linear DR methods as t-SNE and kernel embedding. Transfer learning quality is evaluated by training a linear SVM on the projected source data, classify the projected target data, and compare their labels.

Finally, Peluffo-Ordóñez et al. propose an interactive data visualisation based on a geometrical homotopy, in order to transpose concepts like interaction and controllability from InfoVis to DR. They define a bi-parametric mixture of kernel matrices representing different spectral DR methods, such that non-experts can select or combine methods by simply picking points in a polygonal surface.

5 Perspectives for the near future

After more than a century of research, DR is reaching maturity. Starting from early linear methods like PCA and MDS, the field has known several (r)evolution. First, the extension from linear projection to nonlinear mapping, even with classical spectral optimisation techniques. Next, biological influences (SOMs, AEs) have initiated the moves towards data visualisation (InfoVis) and scalability. Neighbourhood preservation [36, 37, 38, 39] has then led to methods able to deal with very high dimensions, as well as new quality criteria [38, 44]. Tightening the connection with InfoVis and improving scalability without degrading quality are now the necessary steps to face the big data era and make DR applicable in real situations.

References

[1] M.A. Carreira-Perpiñán. A review of dimension reduction techniques. Technical report, University of Sheffield, Sheffield, January 1997.

[2] J.A. Lee and M. Verleysen. *Nonlinear dimensionality reduction.* Springer, 2007.

[3] K. Bunte, M. Biehl, and B. Hammer. A general framework for dimensionality reducing data visualization using explicit mapping functions. *Neural Computation*, 24(3):771–804, 2012.

[4] G. Bouchard. Bias-variance tradeoff in hybrid generative-discriminative models. In *Proceedings of the Sixth International Conference on Machine Learning and Applications*, ICMLA '07, pages 124–129, Washington, DC, USA, 2007. IEEE Computer Society.

[5] C. Pal, M. Kelm, X. Wang, G. Druck, and A. McCallum. On discriminative and semi-supervised dimensionality reduction. In *Advances in Neural Information Processing Systems, Workshop on Novel Applications of Dimensionality Reduction, (NIPS Workshop)*, 2006.

[6] N. Batmanghelich, B. Taskar, and C. Davatzikos. Generative-discriminative basis learning for medical imaging. *IEEE Trans. Med. Imaging*, 31(1):51–69, 2012.

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

[7] I.T. Jolliffe. *Principal Component Analysis.* Springer-Verlag, New York, NY, 1986.

[8] T.F. Cox and M.A.A. Cox. *Multidimensional Scaling.* Chapman & Hall, London, 1995.

[9] I. Borg and P.J.F. Groenen. *Modern Multidimensional Scaling: Theory and Applications.* Springer-Verlag, New York, 1997.

[10] T. Kohonen. *Self-Organizing Maps.* Springer, Heidelberg, 2nd edition, 1995.

[11] C.M. Bishop, M. Svensén, and K.I. Williams. GTM: A principled alternative to the self-organizing map. *Neural Computation*, 10(1):215–234, 1998.

[12] L.K. Saul, K.Q. Weinberger, J.H. Ham, F. Sha, and D.D. Lee. Spectral methods for dimensionality reduction. In O. Chapelle, B. Schoelkopf, and A. Zien, editors, *Semisupervised Learning.* MIT Press, 2006.

[13] D. A. Norman. *The Design of Everyday Things.* Basic Books, Inc., New York, NY, USA, 2002.

[14] S. J. Fernstad, J. Shaw, and J. Johansson. Quality-based guidance for exploratory dimensionality reduction. *Information Visualization*, 12(1):44–64, 2013.

[15] I. D. Blanco, A. A. Cuadrado Vega, D. Pérez-López, F. J. García-Fernández, and M. Verleysen. Interactive dimensionality reduction for visual analytics. In *22th European Symposium on Artificial Neural Networks, ESANN 2014, Bruges, Belgium, April 23-25, 2014*, 2014.

[16] J.W. Sammon. A nonlinear mapping algorithm for data structure analysis. *IEEE Transactions on Computers*, CC-18(5):401–409, 1969.

[17] P. Demartines and J. Hérault. Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *IEEE Transactions on Neural Networks*, 8(1):148–154, January 1997.

[18] J. Venna and S. Kaski. Local multidimensional scaling. *Neural Networks*, 19:889–899, 2006.

[19] R.M. Gray. Vector quantization. *IEEE Acoustics, Speech and Signal Processing Magazine*, 1:4–29, April 1984.

[20] J. Mao and A.K. Jain. Artificial neural networks for feature extraction and multivariate data projection. *IEEE Transactions on Neural Networks*, 6(2):296–317, March 1995.

[21] D. de Ridder and R.P.W. Duin. Sammon's mapping using neural networks: A comparison. *Pattern Recognition Letters*, 18(11–13):1307–1316, 1997.

[22] M. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243, 1991.

[23] E. Oja. Data compression, feature extraction, and autoassociation in feedforward neural networks. In T. Kohonen, K. Mäkisara, O. Simula, and J. Kangas, editors, *Artificial Neural Networks*, volume 1, pages 737–745. Elsevier Science Publishers, B.V., North-Holland, 1991.

[24] D. DeMers and G.W. Cottrell. Nonlinear dimensionality reduction. In D. Hanson, J. Cowan, and L. Giles, editors, *Advances in Neural Information Processing Systems (NIPS 1992)*, volume 5, pages 580–587. Morgan Kaufmann, San Mateo, CA, 1993.

[25] G.E. Hinton and R.R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, July 2006.

[26] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

[27] C.K.I. Williams. On a connection between Kernel PCA and metric multidimensional scaling. In T.K. Leen, T.G. Diettrich, and V. Tresp, editors, *Advances in Neural Information Processing Systems (NIPS 2000)*, volume 13, pages 675–681. MIT Press, Cambridge, MA, 2001.

[28] J. Ham, D.D. Lee, S. Mika, and B. Schölkopf. A kernel view of the dimensionality reduction of manifolds. In *21th International Conference on Machine Learning (ICML-04)*, pages 369–376, 2004.

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

[30] K.Q. Weinberger and L.K. Saul. Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*, 70(1):77–90, 2006.

[31] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems (NIPS 2001)*, volume 14. MIT Press, 2002.

[32] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[33] M. Saerens, F. Fouss, L. Yen, and P. Dupont. The principal components analysis of a graph, and its relationships to spectral clustering. In *Proceddings of the 15th European Conference on Machine Learning (ECML 2004)*, pages 371–383, 2004.

[34] L. Yen, D. Vanvyve, F. Wouters, F. Fouss, M. Verleysen, and M. Saerens. Clustering using a random-walk based distance measure. In M. Verleysen, editor, *Proceedings of ESANN 2005, 13th European Symposium on Artificial Neural Networks*, pages 317–324, Bruges, Belgium, April 2005. d-side.

[35] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouimet. Out-of-sample extensions for LLE, isomap, MDS, eigenmaps, and spectral clustering. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems (NIPS 2003)*, volume 16. MIT Press, Cambridge, MA, 2004.

[36] G. Hinton and S.T. Roweis. Stochastic neighbor embedding. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems (NIPS 2002)*, volume 15, pages 833–840. MIT Press, 2003.

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

[38] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11:451–490, 2010.

[39] J.A. Lee, E. Renard, G. Bernard, P. Dupont, and M. Verleysen. Type 1 and 2 mixtures of Kullback-Leibler divergences as cost functions in dimensionality reduction based on similarity preservation. *Neurocomputing*, 112:92–108, 2013.

[40] J.A. Lee and M. Verleysen. Shift-invariant similarities circumvent distance concentration in stochastic neighbor embedding and variants. In *Proc. International Conference on Computational Science (ICCS 2011)*, pages 538–547, Singapore, 2011.

[41] M. Vladymyrov and M.Á. Carreira-Perpiñán. Entropic affinities: Properties and efficient numerical computation. In *Proc. 30th International Conference on Machine Learning (ICML)*, volume 28 of *JMLR: W&CP*. Atlanta, Georgia, 2013.

[42] Z. Yang, J. Peltonen, and S. Kaski. Scalable optimization of neighbor embedding for visualization. In *Proc. 30th International Conference on Machine Learning (ICML)*, volume 28 of *JMLR: W&CP*, pages 786–794, Atlanta, Georgia, 2013.

[43] M. Vladymyrov and M. Á Carreira-Perpiñán. Linear-time training of nonlinear low-dimensional embeddings. In *Proc. 17th International Conference on Artificial Intelligence and Statistics (AISTATS 2014)*, pages 968–977, 2014.

[44] J.A. Lee and M. Verleysen. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*, 72(7–9):1431–1443, 2009.