Discussion of "Spectral Dimensionality Reduction via Maximum Entropy"

Laurens van der Maaten
Delft University of Technology
lvdmaaten@gmail.com

Since the introduction of LLE (Roweis and Saul, 2000) and Isomap (Tenenbaum et al., 2000), a large number of non-linear dimensionality reduction techniques (manifold learners) have been proposed. Many of these non-linear techniques can be viewed as instantiations of Kernel PCA; they employ a cleverly designed kernel matrix[1] that preserves local data structure in the "feature space" (Bengio et al., 2004). The kernel matrices of the first manifold learners were handcrafted: for instance, LLE uses an inverse squared graph Laplacian of the reconstruction weight matrix as kernel matrix, Isomap uses a centered geodesic distance matrix, and Laplacian Eigenmaps uses an inverse neighborhood graph Laplacian (Belkin and Niyogi, 2002).

More recently, several techniques have been proposed that, instead of designing the kernel matrix by hand, try to learn a good kernel matrix from the data (Weinberger and Saul, 2009; Shaw and Jebara, 2009). In particular, these techniques impose linear constraints on the kernel matrix $\mathbf{K}$ that are designed to preserve local data structure in the feature space; the techniques optimize an objective function that approximately minimizes the rank of the kernel matrix subject to these constraints. The (approximate) rank minimization is required because we wish to obtain a compact data representation. The main differences between "kernel-learning" dimensionality reduction techniques are in the way the approximate rank constraint is implemented: for instance, Maximum Variance Unfolding (MVU) maximizes the variance of the embedding (Weinberger and Saul, 2009), whereas Structure Preserving Embedding maximizes the similarity between the eigenvectors of the kernel matrix and those of the data adjacency matrix (Shaw and Jebara, 2009).

Maximum Entropy Unfolding (MEU; Lawrence (2011)) fits in the group of kernel-learning dimensionality reduction techniques. In particular, MEU uses exactly the same linear constraints as MVU to preserve small pairwise distances in the feature space. The main novelty in MEU is that the approximate rank constraint is implemented by developing a probabilistic interpretation of dimensionality reduction: Lawrence (2011) uses the principle of maximum entropy (Jaynes, 1986) to develop a distribution $p(\mathbf{Y})$ over the data set $\mathbf{Y} \in \mathbb{R}^{N \times D}$, and obtains the low-dimensional data representation through an SVD of the inverse covariance (i.e., kernel) matrix of $p(\mathbf{Y})$. The entropy maximization[2] is performed subject to a set of distance preservation constraints (the same constraints are also used in MVU). The distribution $p(\mathbf{Y})$ then takes the form of a Gaussian Random Field (GRF) in which each data point corresponds to a node:

$$p(\mathbf{Y}) = \frac{|\mathbf{L} + \gamma\mathbf{I}|^{\frac{1}{2}}}{(2\pi)^{\frac{ND}{2}}} \exp\left\{-\frac{1}{2}\,\mathrm{trace}\left((\mathbf{L} + \gamma\mathbf{I})\mathbf{Y}\mathbf{Y}^T\right)\right\}.$$

Herein, $\mathbf{L}$ is the Laplacian of the matrix of Lagrange multipliers $\mathbf{\Lambda}$ that correspond to the distance preservation constraints; γ is a hyperparameter. The embedding $\mathbf{X} \in \mathbb{R}^{N \times d}$ is formed by the d principal eigenvectors of the kernel matrix $\mathbf{K} = (\mathbf{L} + \gamma\mathbf{I})^{-1}$.

Below, we discuss the potential impact of MEU on our understanding of: (1) the connections between manifold learning and generative modeling, (2) the curses and blessings of dimensionality, and (3) the way in which manifold learners obtain low-rank solutions.

Manifold learning versus generative modeling. The paper by Lawrence (2011) mainly focuses on the connections between MEU and existing manifold learners such as Isomap and LLE; it presents new connections between manifold learners that go beyond those discussed by Bengio et al. (2004). A connection that remains underexposed is that between MEU and the non-linear probabilistic PCA model known as GPLVM (Lawrence, 2005). Both MEU and the

[1] The kernel function $\kappa(\mathbf{y}_i, \mathbf{y}_j)$ typically uses other data points $\mathbf{y}_k$ (with $k \neq i, j$) as parameters, which leads to kernel matrices that are not Mercer but still Gramian.

[2] The entropy maximization is performed relative to a Gaussian base distribution to prevent $p(\mathbf{Y})$ from blowing up in unconstrained directions.

Table 1: Four dimension reduction approaches.

–	*Non-generative model*	*Generative model*
Global struct.	PCA, Autoenc.	pPCA, GPLVM
Local struct.	Isomap, LLE, MVU	**MEU**

GPLVM model the distribution $p(\mathbf{Y})$ as a GRF, but the covariance $\mathbf{K}$ of the two models is learned differently: in the GPLVM, the embedding $\mathbf{X}$ is learned directly and $\mathbf{K} = \mathbf{X}\mathbf{X}^T$; whereas in MEU, the Lagrange multipliers $\mathbf{\Lambda}$ are learned and $\mathbf{K} = (\mathbf{L} + \gamma\mathbf{I})^{-1}$.

The connection between MEU and the GPLVM is particularly interesting as it shows that MEU provides a unifying framework for two seemingly very different approaches: (1) manifold learning using techniques like Isomap and LLE and (2) generative modeling using models like probabilistic PCA and the GPLVM. Manifold learners learn a smooth mapping *from the data space to the embedding space*. A smooth mapping in this direction preserves mainly local data structure: if similar data points would be modeled far apart in the embedding, a non-smooth mapping from the data space to the embedding space would be required. By contrast, generative models aim to learn a smooth mapping *from the embedding space to the data space*. A smooth mapping in this direction preserves mainly global data structure: if dissimilar points are close together in the embedding, the mapping from the embedding to the data space needs to be non-smooth.

MEU is the first technique that combines generative modeling with the preservation of local data structure (see Table 1). As a result, MEU can be used to obtain new insights in the ongoing discussion on whether it is better to preserve local data structure or to preserve global data structure. For instance, future work may compare the log-likelihood of test data[3] under both a MEU and a GPLVM model in an attempt to investigate whether it is better to preserve local or global data structure (or a combination of the two).

An oddity of the MEU generative model is that it (unlike the GPLVM) does not treat the embedding $\mathbf{X}$ as latent variables. Instead, the embedding is constructed by performing an arbitrary MDS algorithm on the inverse covariance of the GRF. So far, it is unclear whether this peculiarity can be somehow eliminated.

Blessing of dimensionality. An interesting characteristic of MEU is that the quality of the parameter estimates tends to improve with the dimensionality of the data: that is, MEU benefits from a *blessing of dimensionality* (Lawrence, 2011). In essence, MEU benefits from additional dimensions because they may provide additional information that can help to reduce the variance of the parameter estimates. The blessing of dimensionality appears to contradict the famous *curse of dimensionality* (Bellman, 1961); but this is mainly because the curse of dimensionality is often misinterpreted. The curse of dimensionality refers to problematic phenomenons such as concentration of distances in high-dimensional spaces, which do occur when the features are independent, but are largely absent when the features are correlated (Szekely et al., 2011). The extent to which such phenomena surface mainly depends on the number of underlying parameters of the data, and often, it is thus better to speak of the *curse of intrinsic dimensionality*.

Like any other learner, MEU does suffer from the curse of intrinsic dimensionality: identifying 300 underlying parameters is undoubtedly harder than identifying 3 underlying parameters (it leads to higher-rank kernels). Redundancies in the dimensions, however, can increase the quality of the parameter estimates, e.g., because one dimension may correct for noise in another dimension. It is important to note that such blessings of dimensionality are not specific to MEU: all learners in which the number of parameters does not increase with the number of dimensions may benefit from it (Donoho, 2000). MEU nicely highlights the blessing of dimensionality of these learners.

Rank minimization. A notable difference between MVU and MEU is that maximizing entropy appears to be a looser rank minimization technique than maximizing variance. Maximizing variance (like in MVU) is identical to minimizing the sum of the eigenvalues, whereas maximizing entropy (like in MEU) is similar to maximizing the sum of the log-eigenvalues: maximizing entropy thus heavily penalizes solutions that give rise to infinitesimal eigenvalues. It is yet unclear which of the two approaches is better, but results from other studies suggest that different ways of dealing with dissimilar points can produce very different results. For instance, a study by Carreira-Perpiñán (2010) on the similarities between Stochastic Neighbor Embedding (SNE; Hinton and Roweis (2003)) and Laplacian Eigenmaps suggests that how dissimilar points are pushed apart is much more important than how similar points are pulled together. In particular, Laplacian Eigenmaps tends to collapse points and SNE does not, whilst the only difference between the two is in how they deal with dissimilar points. So far, good results have been obtained by pushing dissimilar data far away (van der Maaten and Hinton, 2008; Weinberger and Saul, 2009), but perhaps MEU can provide new ideas on this.

[3] A notable problem here is that both MEU and the GPLVM do not provide a way to compute the exact log-likelihoods of test data.

Acknowledgements

The author has benefitted from discussions with Geoffrey Hinton, Lawrence Saul, Marco Loog, and Neil Lawrence. The author is supported by NWO award number 680.50.0908, and by EU-FP7 SSPNet.

References

M. Belkin and P. Niyogi. Laplacian Eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems*, volume 14, pages 585–591, 2002.

R. Bellman. *Adaptive control processes: A guided tour*. Princeton University Press, 1961.

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 *Advances in Neural Information Processing Systems*, volume 16, Cambridge, MA, 2004. The MIT Press.

M.Á. Carreira-Perpiñán. The elastic embedding algorithm for dimensionality reduction. In *Proceedings of the* 27^{th} *International Conference on Machine Learning*, pages 167–174, 2010.

D.L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. Technical report, Stanford University, 2000.

G.E. Hinton and S.T. Roweis. Stochastic Neighbor Embedding. In *Advances in Neural Information Processing Systems*, volume 15, pages 833–840, 2003.

E.T. Jaynes. Bayesian methods: General background. In J.H. Justice, editor, *Maximum Entropy and Bayesian Methods in Applied Statistics*, pages 1–25, 1986.

N.D. Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, 6(Nov):1783–1816, 2005.

N.D. Lawrence. Spectral dimensionality reduction via maximum entropy. In *Proceedings of the* 14^{th} *International Conference on Artificial Intelligence and Statistics*, 2011.

N.D. Lawrence and J. Quiñonero Candela. Local distance preservation in the GP-LVM through back constraints. In *Proceedings of the International Conference on Machine Learning*, pages 513–520, 2006.

S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326, 2000.

B. Shaw and T. Jebara. Structure preserving embedding. In *Proceedings of the International Conference on Machine Learning*, pages 937–944, 2009.

E. Szekely, E. Bruno, and S. Marchand-Maillet. The duality of the notions of distance and nearest neighbours in a high-dimensional context. In *Proceedings of KDD (in press)*, 2011.

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

L.J.P. van der Maaten and G.E. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2431–2456, 2008.

K.Q. Weinberger and L.K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10 (Feb):207–244, 2009.