Rank Priors for Continuous Non-Linear Dimensionality Reduction

Raquel Urtasun[1], Andreas Geiger[2], Trevor Darrell[1]
UC Berkeley EECS & ICSI, Berkeley, U.S.A.[1]
Karlsruhe Institute of Technology, Germany[2]
rutasun@csail.mit.edu, geiger@mrt.uka.de, trevor@eecs.berkeley.edu

Many problems involve high dimensional datasets that are computationally challenging to analyze. In such cases it is desirable to reduce the dimensionality of the data while preserving the original information in the data distribution, allowing for more efficient learning and inference. Linear dimensionality reduction techniques (e.g., PCA) have been very popular in the past, due to their simplicity and efficiency. However in practice they can result in poor approximations when dealing with complex datasets.

Graph-based methods [4] exploit local neighborhood distances to approximate the geodesic distance in the manifold. They have been shown to be very effective when dealing with large datasets that are homogeniously sampled. However, they suffer in the presence of noisy and sparse data. Unfortunately, a large set of real world datasets are sparse. Human motion datasets are comprised of small number of examples of motions from different subjects performing different activities. While these databases are typically densely sampled in time, they are sparse in the motion style and activity type. Object recognition databases also suffer from sparsity: only a few objects are labeled for categories with large variation in appearence.

Non-linear probabilistic models, such as the GPLVM [3], can recover complex manifolds, and have received considerable attention in recent years [6, 1]. However, they have only been applied to small databases typically composed of very few examples of a single activity [6]. Moreover, the latent dimensionality was either chosen by the user or optimized by cross-validation, which is computationally expensive. While their representation power is desirable, such methods suffer from local minima, since they rely on computationally expensive optimization of complex non-linear functions that are generally non-convex. Even with the right dimensionality, if initialized far from the optimimum, they can result in poor representations [5]. Factors which contribute to this include the distortion introduced by the initialization and the non-convexity of the optimization. This is aggravated when optimizing very low-dimensional latent spaces, which is typically the case in applications such as tracking [6].

In this paper we present a new learning paradigm that mitigates the problem of local minima by performing *continuous* dimensionality reduction. In contrast to previous GPLVM-based approaches, no distortion is introduced by an initialization step since the latent coordinates are initialized to be the original observations. By introducing a prior over the dimensionality of the latent space that encourages sparsity of the singular values, our method is able to simultaneously estimate the latent space and its dimensionality.

Let $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_N]^T$ be the set of observations $\mathbf{y}_i \in \Re^D$, and let $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]^T$ be the set of latent variables $\mathbf{x}_i \in \Re^Q$, with $Q \ll D$. Probabilistic LVMs relate the latent variables to a set of observed variables via a probabilistic mapping, $y^{(d)} = f(\mathbf{x}) + \eta$, with $y^{(d)}$ the d-th coordinate of $\mathbf{y}$, and $\eta \sim \mathcal{N}(0, \theta_3)$ iid Gaussian noise. The Gaussian Process Latent Variable Model (GPLVM) places a Gaussian process prior over the space of mapping functions f. Marginalizing over f and assuming conditional independence of the output dimensions given the latent variables results in

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{d=1}^{D} \mathcal{N}(\mathbf{Y}^{(d)}|\mathbf{0}, \mathbf{K})$$

where $\mathbf{Y}^{(d)}$ is the d-th column in $\mathbf{Y}$, and $\mathbf{K}$ is the covariance matrix, typically defined in terms of a kernel function. Learning is performed by maximizing the posterior $p(\mathbf{X}|\mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})$ with respect to the latent variables $\mathbf{X}$ and the kernel hyperparameters Θ. $p(\mathbf{X})$ encodes prior knowledge about the latent space.

PCA and graph-based techniques are commonly used to initialize the latent space in GPLVM-based dimensionality reduction; both offer closed-form solutions. However, PCA cannot capture non-linear dependencies, LLE gives a good initialization *only* if the data points are uniformly sampled along the manifold, and Isomap has difficulty with non-convex datasets [2]. Generally, when initialized far from the global minimum,

the GPLVM optimization can get stuck in local minima [3, 5]. To avoid this problem different priors over the latent space have been developed. In [7] a prior was introduced in the form of a Gaussian process over the dynamics in the latent space. This results in smoother manifolds but performs poorly when learning stylistic variations of a motion or multiple motions [5]. In [5] a prior over the latent space was proposed, inspired by the LLE cost function, that encourages smoothness and allows the introduction of prior knowledge, e.g., topological information about the manifold. However, such prior knowledge is not commonly available, reducing considerably the applicability of their technique.

Here we introduce a continuous dimensionality reduction technique that initializes the latent space to the observation space to avoid initial distorsions, and learns the latent space and its dimensionality by introducing a prior that penalizes latent spaces with high dimensionality. The dimensionality of the latent space can be described by the rank of the Gram matrix of the latent coordinates, which can be computed as the number of non-zero eigenvalues. However, it is difficult to enforce directly a prior on the rank since it is a discrete quantity. Instead, we propose a relaxation that results in a penalty function which gradients are continuous and can be easily computed. In particular, we introduce a prior of the form

$$p(\mathbf{X}) = \frac{1}{Z} \exp\left(-\alpha \sum_{i=1}^{D} \phi(s_i)\right) \tag{1}$$

where s_i are the normalized singular values of the mean-subtracted matrix of latent coordinates, with D the dimensionality of the latent space, and Z a normalization constant.

Different penalty functions ϕ can be considered. Common choices for sparsity are the power family and the (generalized) elastic net. In the power family

$$\phi(s_i, p) = |s_i|^p \tag{2}$$

sparsity is achieved for $p \leq 1$. The L_2 norm (i.e., $p = 2$) is a well studied penalty, but does not encourage sparsity. It is equivalent to a Gaussian prior over the singular values in (1). The most commonly used penalty that encourage sparsity is the L_1 norm (i.e., $p = 1$), that results in a Laplace prior over the singular values in Eq. (1). This case is in general attractive since the penalty function is linear, and when the objective function is also linear the optimization can be effectively solved with a Linear Program, even with large number of variables. However here we are interested in learning non-linear latent spaces; our objective function is non-linear even when ϕ is linear. In particular, we minimize the negative log posterior

$$\mathcal{L} = \frac{D}{2} \ln |\mathbf{K}| + \frac{D}{2} tr(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^T) + \alpha \sum_{i=1}^{D} \phi(s_i) \ , \tag{3}$$

where α controls the influence of the penalty in the optimization. Of particular interest to us are functions ϕ that drive small singular values faster towards 0 than larger ones. Examples of such functions are the power family with $p < 1$, logarithmic and sigmoid functions.

We demonstrate the effectiveness of our approach to discover the latent structure and its dimensionality in a variety of artificial datasets. We then illustrate the application of our method to the problem of tracking and classifying 3D articulated motion. Our approach proofs superior to tracking in the original space and tracking using standard GPLVM in a variety of synthetic and real databases.

References

[1] K. Grochow, S. Martin, A. Hertzmann, and Z. Popovic. Style-based inverse kinematics. In *SIGGRAPH*, 2004.

[2] Stefan Harmeling. Exploring model selection techniques for nonlinear dimensionality reduction. Technical report, Edinburgh University, 2007.

[3] Neil Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *JMLR*, 6:1783–1816, 2005.

[4] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 2000.

[5] R. Urtasun, D. J. Fleet, A. Geiger, J. Popovic, T. Darrell, and Lawrence N.D. Topologically-constrained latent variable models. In *ICML*, 2008.

[6] R. Urtasun, D. J. Fleet, A. Hertzman, and P. Fua. Priors for people tracking from small training sets. In *ICCV*, volume 1, pages 403–410, Beijing, China, 2005.

[7] Jack M. Wang, David J. Fleet, and Aaron Hertzmann. Gaussian process dynamical models for human motion. *PAMI*, 30(2):283–298, 2008.