Kick-starting GPLVM Optimization via a Connection to Metric MDS

Sebastian Bitzer
Max Planck Institute
for Human Cognitive and Brain Sciences
P.O. box 500355, 04303 Leipzig, Germany
bitzer@cbs.mpg.de

Christopher K. I. Williams
School of Informatics
University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK
c.k.i.williams@ed.ac.uk

Abstract

The Gaussian Process Latent Variable Model (GPLVM) [1] is an attractive model for dimensionality reduction, but the optimization of the GPLVM likelihood with respect to the latent point locations is difficult, and prone to local optima. Here we start from the insight that in the GPLVM, we should have that $k(\mathbf{x}_i, \mathbf{x}_j) \simeq s_{ij}$, where $k(\mathbf{x}_i, \mathbf{x}_j)$ is the kernel function evaluated at latent points $\mathbf{x}_i$ and $\mathbf{x}_j$, and s_{ij} is the corresponding estimate from the data. For an isotropic covariance function this relationship can be inverted to yield an estimate of the interpoint distances $\{d_{ij}\}$ in the latent space, and these can be fed into a multidimensional scaling (MDS) algorithm. This yields an initial estimate of the latent locations, which can be subsequently optimized in the usual GPLVM fashion. We compare two variants of this approach to the standard PCA initialization and to the ISOMAP algorithm [2], and show that our initialization converges to the best GPLVM likelihoods on all six tested motion capture data sets.

1 Introduction

The Gaussian Process Latent Variable Model (GPLVM) [1] has recently gained attention as a nonlinear dimensionality reduction method which can provide a powerful generative model of the data. This is particularly useful for the generation of new motion from examples [3, 4] or embedded in a tracking system [5]. Another advantage of the GPLVM, from a practitioner's point of view, is that prior information can be conveniently incorporated into the model in form of a prior on the latent points [e.g. 6, 7]. On the other hand, the GPLVM suffers from high computational demands and the complexity of the underlying optimization problem. In this paper we address the latter issue by relating the GPLVM to metric multidimensional scaling (MDS) which in return yields a novel initialization for the GPLVM optimization.

In particular, given a data set $\mathbf{Y} \in \mathbb{R}^{N\times D}$ of N vectors in D dimensions the GPLVM maximizes the following log-likelihood with respect to a latent configuration of points $\mathbf{X} \in \mathbb{R}^{N\times M}$ in a lower dimensional space ($M < D$):

$$L = -\frac{DN}{2}\log 2\pi - \frac{D}{2}\log|\mathbf{K}| - \frac{D}{2}\operatorname{tr}(\mathbf{K}^{-1}\mathbf{S}) \tag{1}$$

where $\mathbf{K}$ is the $N \times N$ Gram matrix with entries $k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{x}_i$ is the ith column of $\mathbf{X}^T$, $k(\cdot,\cdot)$ is a kernel function, and $\mathbf{S} = \frac{1}{D}\mathbf{Y}\mathbf{Y}^T$ is the $N \times N$ matrix with entries $s_{ij} = \sum_{d=1}^{D} y_{id}y_{jd}/D$. For a nonlinear kernel function, maximization of the log-likelihood thus implements nonlinear dimensionality reduction, but we also see that we then have to optimize with respect to NM nonlinearly related parameters $\mathbf{X}$ plus additional parameters of the kernel function. As the optimization landscape of this problem is highly complex with many local optima, the initialization of $\mathbf{X}$ is critical for successful application of the GPLVM. Part of the success of the GPLVM can, therefore, be attributed

to the insight that the result of (probabilistic) PCA often is a good, heuristic initialization. In the following we present an alternative initialization for the GPLVM which we derive directly from the model. In contrast to PCA it is nonlinear and in contrast to other potential initializations, like Isomap [2], it is particularly suited for the use with the GPLVM. Parts of this work with more detailed theoretical explanations are also published in the author's PhD thesis [8] and further experimental results will be available in a forthcoming technical report [9].

2 Relating the GPLVM to Metric MDS

The model defined by eq. (1) corresponds to D independent draws from a common Gaussian Process [10] with mean function $m(\mathbf{x}) = 0$ and covariance function $k(\mathbf{x}_i, \mathbf{x}_j)$. Thus a sensible preprocessing of the data is to centre each column $\mathbf{y}^c$ of $\mathbf{Y}$ to have zero mean for $c = 1, \ldots, D$, and rescale each one to have the same variance[1] (taken here to be unity). We assume below that $\mathbf{S}$ is computed using this pre-processed data.

Our initialization is based on the insight that the free-form maximization of the likelihood in eq. (1) over $\mathbf{K}$ is obtained by setting $\mathbf{K} = \mathbf{S}$. Of course, as $\mathbf{K}$ is parameterized by $\mathbf{X}$ it will not in general be possible to find locations $\mathbf{X}$ so as to make this happen. However, it does suggest that we might try setting $k(\mathbf{x}_i, \mathbf{x}_j) \simeq s_{ij}$ for all $i,\ j$. If the kernel function k is isotropic, i.e. it is a function of $d_{ij}^2 = |\mathbf{x}_i - \mathbf{x}_j|^2$ so that $k(\mathbf{x}_i, \mathbf{x}_j) = f(d_{ij}^2)$, then we have $d_{ij}^2 \simeq f^{-1}(s_{ij}) \quad \forall\ i,\ j$. For example, the squared exponential (SE) covariance function sets $k_{ij} = \exp(-d_{ij}^2/2\ell^2)$, so that $d_{ij}^2 \simeq -2\ell^2 \log(s_{ij})$. Given an $N \times N$ matrix of distances with entries d_{ij}^2 it is then straightforward to solve for the best M-dimensional Euclidean configuration using classical MDS [11].

Scaling S: It is sensible to impose the constraint that the diagonal entries in $\mathbf{S}$ are such that $d_{ii}^2 = 0$ for all $i = 1, \ldots, N$. Assuming that $f^{-1}(1) = 0$ (which holds e.g. for the SE covariance function), then this can be achieved by replacing $\mathbf{S}$ by its rescaled version $\mathbf{R}$, where

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii} s_{jj}}}, \tag{2}$$

so that $r_{ii} = 1$ for $i = 1, \ldots, N$. (This is similar to the construction of the correlation matrix from a covariance matrix, except that here the notions of sample and variable are interchanged so that $\mathbf{S}$ and $\mathbf{R}$ are $N \times N$, not $D \times D$.) This, additionally, ensures that $d_{ij}^2 \simeq -2\ell^2 \log(r_{ij})$ produces valid, i.e., positive distances. We assume below that $\mathbf{R}$ is used in place of $\mathbf{S}$.

A problem: It may happen that there is no d_{ij} corresponding to values of r_{ij} in a certain range. For example, with the SE kernel we cannot find d_{ij}'s corresponding to $r_{ij} \leq 0$, but such values may well arise in practice. Indeed, due to sampling fluctuations, negative empirical r_{ij}'s could occur even if the "true" value were positive, but they could also arise through model mis-specification. A simple approach in this case is to treat the entries with $r_{ij} \leq 0$ as missing, and apply an MDS algorithm that handles missing data as described below. However, note that small r_{ij} corresponds to large d_{ij} for the SE kernel, so there is an expectation that these missing distances in $\mathbf{x}$-space will be large.

MDS with Missing Data: Iterative Minimization of Stress. When we have missing entries in the matrix of dissimilarities we cannot compute the eigendecomposition of it anymore and have to resort to other techniques. Several algorithms for MDS with missing data have been proposed, one of which is the iterative minimization of the Stress criterion [e.g. 12]: $\text{Stress}(\mathbf{X}) = \left(\frac{\sum_{i,j} w_{ij} (d_{ij} - |\mathbf{x}_i - \mathbf{x}_j|)^2}{\sum_{i,j} w_{ij} d_{ij}^2} \right)^{\frac{1}{2}}$, which minimizes the normalized squared error between the given dissimilarities and the distances between the estimated latent points. The weights w_{ij} can then be used to accommodate missing values by setting corresponding weights to 0. Stress can then be minimized with gradient descent. We employ the routine provided in the Matlab statistics toolbox (mdscale). We repeat iterative MDS with R different random initializations of $\mathbf{X}$ and select the resulting reconstruction of points which gives the highest GPLVM log-likelihood. In all experiments below we set $R = 100$.

We name this method GPLVM-Stress, as it is based on finding a latent configuration from the incomplete, approximated distance matrix via the minimization of Stress. The computational cost of

[1] This saves needing a separate signal variance parameter for each column of Y.

GPLVM-Stress is $O(RIN^2)$ where R is the number of repetitions described above, N is the number of data points and I is the number of iterations typically needed for the gradient descent on Stress to converge. When I is much larger than N, GPLVM-Stress can, therefore, be considerably slower than the original optimisation of the GPLVM which has a cost of $O(JN^3)$ where J is the number of GPLVM gradient steps.

As an alternative to Stress minimization we also used Isomap to find a latent configuration from the approximated distances. This is based on the insight that predominantly large distances will be missing, and that Isomap only uses the k nearest neighbours to approximate geodesic distances between all data points and so automatically fills in the missing distances. We name this method GPLVM-Isomap-low, because the low-dimensional distances are used. In contrast, when we apply Isomap directly on the distances between data points $\mathbf{Y}$, we name this GPLVM-Isomap-high. In both cases we select k to maximise the GPLVM (log-)likelihood.

After completing our work we became aware of [13] in which Modayil also developed the idea to use correlations between observed variables to define a low-dimensional embedding of the data. His work is in the context of discovering spatial relations between sensors of robots and he does not discuss the relation to the GPLVM. Instead of Isomap he uses regularized maximum variance unfolding [14] to fill in missing distances which gave results similar to GPLVM-Isomap-low in initial tests with our data sets, but which was roughly one order of magnitude slower than Isomap.

Variability of Covariance Estimates. Under the GPLVM model the columns of the matrix of observations $\mathbf{Y}$ are independent samples from a multivariate Gaussian distribution with mean $\mathbf{0}$ and covariance $\mathbf{K}$. Thus, the matrix $\mathbf{Y}\mathbf{Y}^T$ is Wishart distributed [15] with parameter $\mathbf{K}$ and D degrees of freedom: $\mathbf{Y}\mathbf{Y}^T \sim W_N(\mathbf{K}, D)$. The Wishart distribution has mean $E(\mathbf{Y}\mathbf{Y}^T) = D\mathbf{K}$ and the elements of $\mathbf{Y}\mathbf{Y}^T$ have variance $V(\mathbf{y}_i^T\mathbf{y}_j) = D(k_{ij}^2 + k_{ii}k_{jj})$ [see e.g. 16, ch. 3.2]. Consequently, the mean and variance for the sample covariance are $E(\mathbf{S}) = \frac{1}{D}E(\mathbf{Y}\mathbf{Y}^T) = \mathbf{K}$ and $V(s_{ij}) = \frac{1}{D^2}V(\mathbf{y}_i^T\mathbf{y}_j) = \frac{1}{D}(k_{ij}^2 + k_{ii}k_{jj})$. Therefore, $\mathbf{S}$ is an unbiased estimator of $\mathbf{K}$ and the estimate improves as the number of observation vectors D increases. In other words, as we obtain more samples of the covariances between data points from different dimensions, our estimate of the underlying covariances becomes better.

3 Experiments

Synthetic Data. In extensive evaluations on synthetic data we have confirmed this somewhat surprising result which means that results of dimensionality reduction improve (better reflecting the true underlying configuration of points) with increasing dimensionality of the data, if the data follows the assumptions of the model. We also showed that a predominant part of the error in representing the true latent configuration is contributed by badly-approximated small covariances which correspond to large distances. Furthermore, we demonstrated that the resulting GPLVM-Stress and, especially, GPLVM-Isomap-low configurations were indeed better initializations of the GPLVM than the standard PCA initialisation in terms of log likelihood (before and after GPLVM optimization). In particular, we investigated data sets with varying dimensionality ($D \in \{3, 5, 8, 14, 23, 39, 65, 108, 180, 300\}$) and we found that for data sets with marked nonlinearities and $D \geq 14$ initialization of the GPLVM with GPLVM-Isomap-low lead to larger likelihoods after GPLVM optimization in 99, 75 and 87 per cent of the tested data sets (336 in total) when compared to PCA, GPLVM-Stress and GPLVM-Isomap-high, respectively. For further details of these experiments we refer the interested reader to the forthcoming technical report [9].

Human Motion Capture Data. We tested the four initialization methods on 6 different motion capture data sets. The first two were our own and represent punches of a single person; the first consisted of the full recorded movements of three punches, while in the second we cut out the retraction at the end of the punches. The remaining data sets have been used in other publications to demonstrate the working of the GPLVM and its variants. In particular, data set 3 (running) has been used in Lawrence and Quinonero-Candela [17] and data sets 4 (walking of a single person), 5 (walking of 4 different people) and 6 (4 golf swings of a single person) in Wang et al. [6]. We normalized these data sets (see [9] for details) and chose latent dimensionality M according to our prior beliefs about the intrinsic dimensionality of the data (3, 2, 3, 3, 3 and 3, respectively, for the 6 data sets). The data sets contained 264, 97, 217, 130, 288 and 255 data points in 57, 57, 102, 53, 53 and 52 dimensions, respectively.

(a) before optimization

	DS1	DS2	DS3	DS4	DS5	DS6
PCA	−458	−365	6	−257	−733	−141
ISO-high	45	−54	185	−160	−597	23
Stress	66	−5	173	−128	**−426**	43
ISO-low	**114**	**35**	**193**	**−120**	−499	**71**

(b) after optimization

	DS1	DS2	DS3	DS4	DS5	DS6
PCA	635	200	587	−5	106	276
ISO-high	658	204	597	−3	**109**	276
Stress	689	208	589	−3	107	281
ISO-low	**702**	**209**	**601**	**−2**	**109**	**289**

Table 1: Normalized GPLVM log-likelihoods (L/D in eq. 1) for the 6 mocap data sets (1-uncut punches, 2-cut punches, 3-run, 4-walk, 5-walks of 4 people, 6-golf swings)

After normalization we applied PCA, GPLVM-Isomap-high, GPLVM-Isomap-low and GPLVM-Stress to the motion capture data sets, and initialized a GPLVM with the resulting latent points and covariance function parameters $\ell = 1, \phi = 1$ and $\sigma^2 = 0.01$. We computed GPLVM log-likelihoods using eq. (1) as before. Subsequently, we optimized the GPLVM using scaled conjugate gradients for 500 steps and recorded the log-likelihood of the result. We note that more than 50% of the entries in the matrix $\mathbf{S}$ were negative in all data sets (52, 56, 58, 56, 55 and 52% for the 6 data sets). From our experience with the synthetic data we, therefore, did not expect GPLVM-Stress to perform well on these data sets. Nevertheless, it consistently achieved better GPLVM likelihoods than PCA before and after optimization, although likelihoods after optimization were quite similar on three data sets (see Table 1). However, the advantage of our approach is even clearer for GPLVM-Isomap-low which outperformed all other methods before and after optimization except on data set 5. As an example, we show in Fig. 1 the two-dimensional latent configurations resulting from data set 2.

4 Discussion

In this paper we have derived a relationship between metric MDS and the GPLVM which allows us to kickstart the optimization of the GPLVM likelihood using MDS procedures which can cope with missing distances. The resulting method is a particular instance of metric MDS based on the inverse of an isotropic covariance function. Our experiments on synthetic data have shown that GPLVM-Stress and GPLVM-Isomap-low clearly outperform PCA as initialization for the GPLVM, and outperform GPLVM-Isomap-high in the majority of cases.

We noted that using GPLVM-Stress may lead to a substantial increase in computational costs. GPLVM-Isomap-low, on the other hand, is as fast as GPLVM-Isomap-high and performs best on all real-world data sets. We note that although GPLVM-Stress and GPLVM-Isomap-low are based on the same approximated distances in latent space the resulting configuration of latent points may still differ considerably (e.g. Fig. 1).

Acknowledgements: This work is supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007- 216886. This publication only reflects the authors' views.

References

[1] N. Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816, 2005.

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

[3] K. Grochow, S. L. Martin, A. Hertzmann, and Z. Popovic. Style-based inverse kinematics. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, 2004.

[4] S. Bitzer and S. Vijayakumar. Latent spaces for dynamic movement primitives. In *Proc. 9th IEEE RAS International Conference on humanoid Robots (Humanoids 2009)*, 2009.

[5] R. Urtasun, D. J. Fleet, and P. Fua. 3d people tracking with gaussian process dynamical models. In *Conference on Computer Vision and Pattern Recognition, CVPR*, 2006.

[6] J. M. Wang, D. J. Fleet, and A. Hertzmann. Gaussian process dynamical models for human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):283–298, Feb 2008.

[7] R. Urtasun, D. J. Fleet, A. Geiger, J. Popovic, T. J. Darrell, and N. D. Lawrence. Topologically-constrained latent variable models. In A. McCallum and S. Roweis, editors, *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, pages 1080–1087. Omnipress, 2008.

[8] S. Bitzer. *Nonlinear Dimensionality Reduction for Motion Synthesis and Control*. PhD thesis, School of Informatics, University of Edinburgh, submitted July 2010.

[9] S. Bitzer and C. K. I. Williams. Kick-starting GPLVM optimization via a connection to metric MDS. Technical Report, in preparation.

[10] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[11] T. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Chapman & Hall/CRC, 2nd edition, 2000.

[12] A. Buja, D. F. Swayne, M. L. Littman, N. Dean, and H. Hofmann. Xgvis: Interactive data visualization with multidimensional scaling. Technical report, AT&T Labs, 2001.

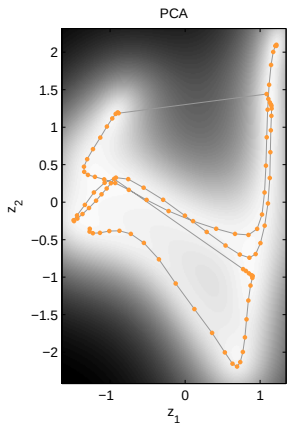
[13] J. Modayil. Discovering sensor space: Constructing spatial embeddings that explain sensor correlations. In *Proc. of International Conference on Development and Learning (ICDL 2010)*, pages 120–125, 2010.

[14] K. Weinberger, F. Sha, Q. Zhu, and L. Saul. Graph regularization for maximum variance unfolding with an application to sensor localization. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems, NIPS, 19*. MIT Press, Cambridge, MA, 2007.

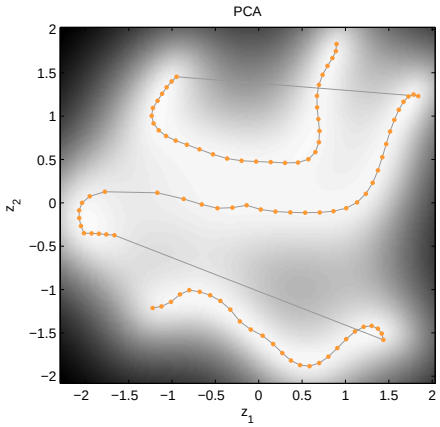
[15] J. Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 20A(1-2):32–52, 1928.

[16] R. J. Muirhead. *Aspects of multivariate statistical theory*. Wiley, 2nd edition, 2005.

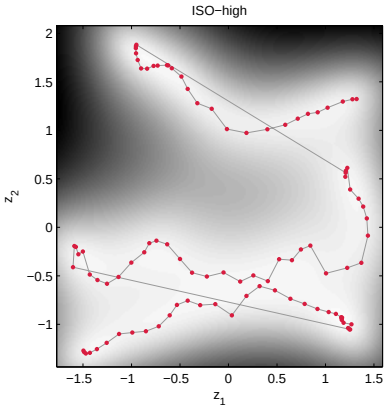
[17] N. D. Lawrence and J. Quinonero-Candela. Local distance preservation in the GP-LVM through back constraints. In *Proceedings of the International Conference in Machine Learning (ICML)*, 2006.



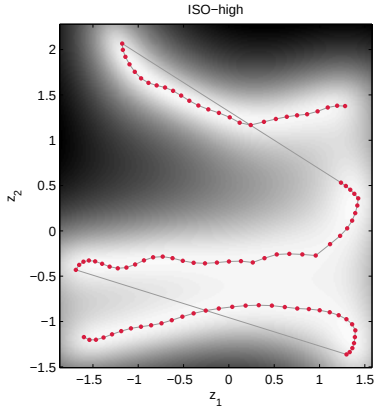
(a) PCA, before optimisation (-365)



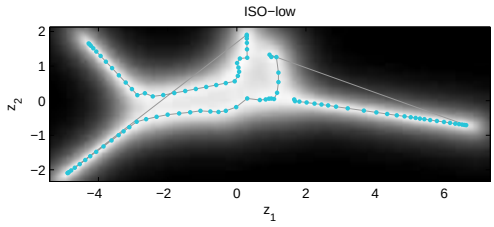
(b) PCA, after optimisation (200)



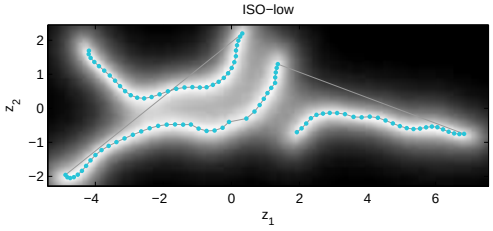
(c) ISO-high, before optimisation (-54)



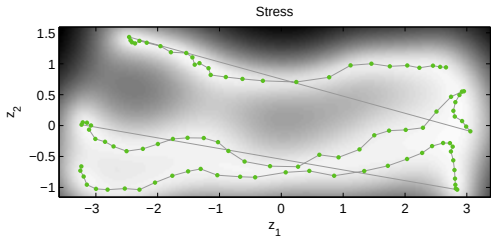
(d) ISO-high, after optimisation (204)



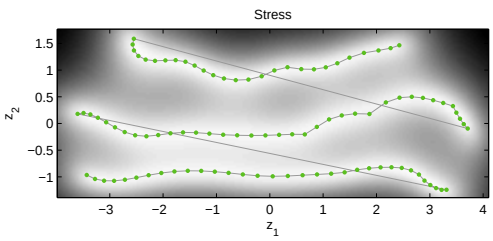
(e) ISO-low, before optimisation (35)



(f) ISO-low, after optimisation (209)



(g) Stress, before optimisation (-5)



(h) Stress, after optimisation (208)

Figure 1: Latent points and log GPLVM predictive confidences (shading) before and after GPLVM optimisation for data set 2 (3 punches without retraction). Numbers in parantheses are normalised log-likelihoods repeated from Table 1. Note that on this data set GPLVM-ISO-low had the largest GPLVM log-likelihood after optimisation. Gray lines indicate temporal order of data points.