Supervised PCA for Interactive Data Analysis

Daniel Paurat **Dino Oglic** **Thomas Gärtner**
University of Bonn and Fraunhofer IAIS, Sankt Augustin, Germany
{daniel.paurat, dino.oglic, thomas.gaertner}@uni-bonn.de

1 Introduction

We investigate a novel approach for intuitive interaction with a data set for explorative data analysis. The key idea is that a user can directly interact with a two or three dimensional embedding of the data and actively place data points to desired locations. To achieve this, we propose a variant of semi-supervised kernel PCA which respects the placement of control points and maximizes the variance of the unlabelled data along the 'directions' of the embedding.

Knowledge Discovery is inherently an iterative process involving data understanding and modeling (Shearer, 2000). One important modelling tool facilitating better data understanding is data visualization. Yet, most algorithms for visualizing data, typically by embedding the data in the two dimensional plane, are static and non-interactive. In a previous paper (Paurat and Gärtner, 2013) we introduced the idea of interactively shaping a lower dimensional embedding of the data to emphasize specific aspects of it and gave an interpretation for the embedding. We utilized a set of selected data records within the embedding as control points to determine the overall embedding and argued that dynamic interactions with data can be achieved by employing simple, least-squares regression as a coordinate wise embedding, in this paper referred to as *least square projection* (LSP).

This approach, while facilitating dynamic interaction, has disadvantages when given only few control points on sparse data, as it is usually the case for real world applications. In this case many data records will be orthogonal to the control points, which causes them to be embedded to the origin. This is of course counter productive when interactively exploring the data. To remedy this effect, in this paper we propose to utilize a *constrained kernel PCA* (cKPCA) which not only embeds the control points to the user specified locations, but also tries to maximize the variance of the data records along the embedding directions.

In general, most of the dimensionality reduction methods used to embed data are unsupervised, i.e., they exploit only the input data and do not consider the assigned labels. A considerable amount of work has been done in this area and some of the well known methods are principal component analysis (Hastie et al., 2001.), isomap (Tenenbaum et al., 2000), locally linear embedding (Roweis and Saul, 2000), non-negative matrix factorization (Lee et al., 1999), archetypal analysis (Cutler and Breiman, 1994), and CUR decomposition (Drineas et al., 2006).

In addition to these methods, there exist several methods for the computation of supervised projections to a lower dimensional space. The most related approach to ours is the semi-supervised kernel PCA (Walder et al., 2010). The difference is that Walder et al. (2010) focus on the optimization of the least square fit subject to a constant variance. Barshan et al. (2011) rely on the empirical Hilbert-Schmidt independence criterion to compute a sequence of principal components that have maximal dependence on the response variable. This is a generalization of kernel PCA, applied to the problems of visualizations and classification on subspaces and submanifolds. Yu et al. (2006) proposed a probabilistic PCA with an efficient EM algorithm for model training for supervised and semi-supervised PCA. Also related to our work is the use of LSP in order to render static embeddings (Paiva et al., 2012; Paulovich et al., 2008), and an approach that focuses on applying an interactive hyperbolic transformation to a given static embedding (Walter and Ritter, 2002). This technique does not alter the perspective on the embedded data, but it gives a user the ability to interactively zoom into regions of interest, without loosing the context to the rest of the embedding.

2 Optimization Problem

In this section we present a variant of semi-supervised kernel PCA which respects the user defined placement of some labelled control points and maximizes the variance of the unlabelled data ‘along’ the set of unit norm functions defining the embedding. To ensure feasibility of the resulting optimization problem while retaining satisfactory visualization, we replace the usual hard orthogonality constraint by a conveniently chosen soft-orthogonality term in the objective function.

Let $X = \{x_1, \ldots, x_n\}$ be a sample from an instance space $\mathcal{X}$ with positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Without loss of generality we can assume that the first m points are labelled with $\{y_1, \ldots, y_m\}$. Furthermore, let $\mathcal{H}$ be the reproducing kernel Hilbert space of kernel k and $\mathcal{H}_X = \operatorname{span}\{k(x_i, \cdot) | x_i \in X\}$. We iteratively construct the unit $\mathcal{H}_X$-norm functions $f_1, \ldots, f_d$ by solving the following optimization problem

$$\begin{aligned} f_s = \underset{f \in \mathcal{H}}{\operatorname{argmax}} \quad & \frac{1}{n} \sum_{i=1}^{n} \left(f(x_i) - \langle f, \mu \rangle \right)^2 - \frac{\nu}{s} \left(1 + \sum_{s' < s} \langle f_{s'}, f \rangle^2 \right) \\ \text{subject to} \quad & \|f\|_{\mathcal{H}_X} = 1, \\ & 1 \leq i \leq m : \ f(x_i) = y_{is}, \end{aligned} \tag{1}$$

where $\mu = \sum_{i=1}^{n} k(x_i, \cdot)$.

Note that the hard constraint above can easily be relaxed using slack variables. Efficient solving remains possible in this case following along the lines outlined below.

2.1 Efficient Solution

The optimization problem (1) is defined over the reproducing kernel Hilbert space $\mathcal{H}$ with kernel $k(\cdot, \cdot)$ and the weak representer theorem (Dinuzzo and Schölkopf, 2012; Schölkopf et al., 2001) implies that $f_s = \sum_{j=1}^{n} \alpha_{sj} k(x_j, \cdot)$ for $\alpha_{s1}, \alpha_{s2}, ..., \alpha_{sn} \in \mathbb{R}$. Let K be the kernel matrix, with its first m rows denoted as $K_{[:m,:n]}$, and $H_n = \mathbf{I}_n - \frac{1}{n} \mathbf{1}_{n \times n}$. We can rewrite problem (1) as follows

$$\begin{aligned} \alpha_s = \underset{\alpha \in \mathbb{R}^n}{\operatorname{argmax}} \quad & \frac{1}{n} \alpha^T K H_n K \alpha - \frac{\nu}{s} \sum_{s' < s} \left(\alpha^T K \alpha_{s'} \right)^2 \\ \text{subject to} \quad & \alpha^T K \alpha = 1 \\ & K_{[:m,:n]} \alpha = y. \end{aligned} \tag{2}$$

Introducing a substitution $K^{\frac{1}{2}} \alpha = u$ and denoting

$$\begin{aligned} W &= K^{\frac{1}{2}} \left(\frac{1}{n} H_n - \frac{\nu}{s} \sum_{s' < s} \alpha_{s'} \alpha_{s'}^T \right) K^{\frac{1}{2}}, \\ L &= K_{[:m,:n]} K^{-\frac{1}{2}}, \end{aligned}$$

we can rewrite the last problem as

$$\begin{aligned} \underset{u \in \mathbb{R}^n}{\operatorname{argmax}} \quad & u^T W u \\ \text{subject to} \quad & u^T u = 1, \\ & L u = y. \end{aligned} \tag{3}$$

A QR factorization of the matrix L^T implies that $L = R^T Q^T$, where $Q \in \mathbb{R}^{n \times n}$ is an orthogonal matrix and $R \in \mathbb{R}^{n \times m}$ is an upper triangular matrix. Introducing a substitution

$$Q^T u = \begin{bmatrix} x \\ z \end{bmatrix}$$

the objective function of the last problem becomes
$u^T W u = u^T Q Q^T W Q Q^T u = (Q^T u)^T Q^T W Q (Q^T u)$, and with

$Q^T W Q = \begin{bmatrix} A & B^T \\ B & C \end{bmatrix}$, where $A \in \mathbb{R}^{m \times m}, B \in \mathbb{R}^{(n-m) \times m}$ and $C \in \mathbb{R}^{(n-m) \times (n-m)}$, we obtain

$u^T W u = x^T A x + 2 z^T B x + z^T C z.$

In a similar way, both constraints are transformed into

$$y = Lu = R^T(Q^T u) = \left[\begin{array}{c} P \\ \mathbf{0}_{(n-m)} \end{array} \right]^T \left[\begin{array}{c} x \\ z \end{array} \right] = P^T x \implies x = (P^T)^{-1} y \text{ and}$$
$$1 = u^T u = (Q^T u)^T (Q^T u) = x^T x + z^T z \implies z^T z = 1 - x^T x = t^2.$$

Now, we can rewrite the last problem as

$$\begin{aligned} \underset{z}{\text{argmax}} \quad & z^T C z - 2 b^T z \\ \text{subject to} \quad & z^T z = t^2, \end{aligned} \tag{4}$$

where C is a symmetric matrix and $b = -Bx$.

To compute the solution to this problem one can first form the principal Lagrangian function and show that the maximum is achieved for the largest value of the Lagrangian parameter associated with the hypersphere constraint. Then, it can be shown that finding the largest value of this parameter is equivalent to solving a quadratic eigenvalue problem. Furthermore, the quadratic eigenvalue problem can be written as a linear eigenvalue problem using block matrices.

The solution to the problem (4) is (Gander et al., 1989):

$$z^* = (C - \lambda_{max} \mathbf{I}_{(n-m)})^{-1} b,$$

where λ_{max} is the largest real eigenvalue of

$$\left[\begin{array}{cc} C & -\mathbf{I}_{(n-m)} \\ -\frac{1}{t^2} b b^T & C \end{array} \right] \left[\begin{array}{c} \gamma \\ \eta \end{array} \right] = \lambda \left[\begin{array}{c} \gamma \\ \eta \end{array} \right].$$

Hence, the solution to problem (2) is given by

$$K^{-\frac{1}{2}} Q \left[\begin{array}{c} \left(P^T\right)^{-1} y \\ \left(C - \lambda_{max} \mathbf{I}_{(n-m)}\right)^{-1} b \end{array} \right].$$

3 Experiments

Figure 1 (left) illustrates the problem with using LSP as an embedding technique on the ICDM 2001 abstracts dataset (Kontonasios and Bie, 2010) and shows how cKPCA is able to overcome it. The LSP embedding collapses towards the origin mainly because the dataset has sparse entries. The five control points (highlighted in red) have few to no attributes in common with the other embedded data records, which leaves LSP unable to embed them anywhere but the origin. As cKPCA also maximizes the variance, the resulting embedding has more spread. This gives the user more insights about the underlying structure of the data, as well as the possibility to better select new control points and interact with the embedding. Note that the small number of control points reflects the actual use case of an interactive embedding. In general, a user would not want to interact with too many control points, but rather with a few known or highly expressive ones.

The middle picture of Figure 1 shows how the average pairwise distance of the embedded data develops depending on the amount of control points. In this experimental setting we choose a number of random control points and place them according to their third and fourth principal components coordinates. One can see how cKPCA starts as a regular PCA and develops with more and more control points selected towards the new embedding, while keeping the high spread among the embedded data records from the beginning on. LSP on the other hand initially places all points at the origin and only slowly develops the desired spread. The right part of Figure 1 shows how cKPCA and LSP develop in terms of root mean squared error (rmse) between the third and forth principal component and the resulting embedding over the number of control points set. With an increasing number of control points, cKPCA develops away from the regular PCA embedding towards the new embedding.

We also evaluated the stray of cKPCA and LSP on a total of eight datasets. Five of them are from the well known UCI repository (Frank and Asuncion, 2010) and three are real world datasets which

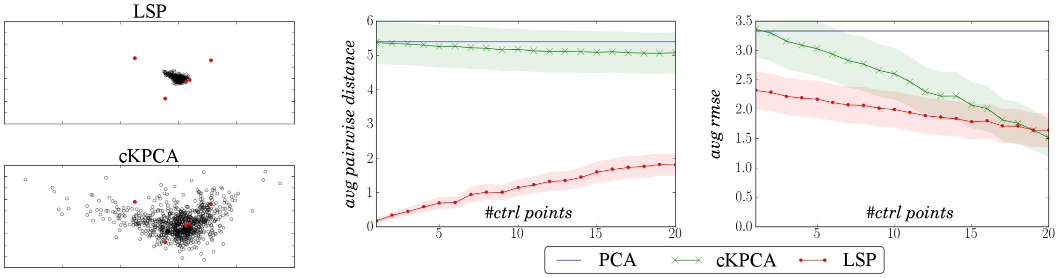

Figure 1: *Left, LSP and cKPCA embedding of the ICDM 2001 abstracts dataset. Middle and right, control points are placed according to their third and fourth principal components coordinates. The middle one shows the development of the averaged pairwise distance of the embedded data over the number of control points selected. The right one shows the development of the root mean squared error between the third and forth principal component and the actual embedding.*

we consider sparse: The already mentioned ICDM 2001 abstracts and two datasets which describe ingredients of food and cocktails crawled from the websites `food.com` and `webtender.com`.[1] For a fixed number of five control points we performed the above explained experimental setup, with ν set to 1.0 and utilizing a linear kernel. Table 1 shows the averaged results over 50 runs. One can see, how the resulting embeddings, especially for the sparser datasets (marked with *), tend to have more average pairwise distance among the embedded data, while having a rmse comparable to the one LSP makes. This can be interpreted as a sign of more stray among the embedded data.

dataset	P-dist$_{cKPCA}$	P-dist$_{LSP}$	rmse$_{cKPCA}$	rmse$_{LSP}$
auto93	2.26±0.20	1.14±0.10	1.09±0.12	0.62±0.07
autoPrice	2.03±0.17	1.35±0.11	0.66±0.08	0.48±0.05
bodyfat	1.68±0.17	1.05±0.10	0.67±0.08	0.45±0.05
pollution	2.36±0.23	1.55±0.13	0.81±0.12	0.56±0.06
servo	2.09±0.16	1.47±0.12	0.69±0.06	0.62±0.06
food.com*	3.56±0.50	0.55±0.13	2.17±0.33	1.37±0.23
ICDM 2001 abstracts*	5.29±0.62	0.63±0.16	3.03±0.47	2.16±0.32
webtender.com*	2.26±0.40	0.31±0.10	1.27±0.30	0.83±0.21

Table 1: *Root mean squared error (rmse) and average pairwise distance (P-dist) for a fixed amount of five control points on eight datasets. The obtained results are averaged over 50 runs.*

4 Conclusion

We have proposed a variant of semi-supervised kernel PCA which respects the placement of control points and maximizes the variance of the unlabelled data along the directions of the embedding. Furthermore, we have given a closed form solution to the formulated non-convex optimization problem. The proposed method can be used in explorative data analysis when utilizing interactive embeddings. Our experiments indicate that the resulting embeddings, especially for sparse datasets, are able to retain most of the variance and that cKPCA, in contrast to LSP, does not suffer from the initialization problem. In the future, on the one hand, we would like to integrate cKPCA into our interactive embedding tool (Paurat and Gärtner, 2013) with additional knowledge based constraints which would enable a more diverse incorporation of domain knowledge. On the other hand, we would like to improve the scalability of the method using an approximation instead of a full span expansion of the optimizer.

Acknowledgement
Part of this work was supported by the German Science Foundation (DFG) under the reference number 'GA 1615/1-1'.

[1] The datasets and the interactive LSP embedding tool can be downloaded here: `http://www-kd.iai.uni-bonn.de/index.php?page=software_details&id=31`

References

E. Barshan, A. Ghodsi, Z. Azimifar, and M.Z. Jahromi. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition*, 44(7):1357–1371, 2011.

A. Cutler and L. Breiman. Archetypal analysis. *Technometrics*, 36 (4):338–347, 1994.

F. Dinuzzo and B. Schölkopf. The representer theorem for Hilbert spaces: a necessary and sufficient condition. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Lon Bottou, Weinberger, and Q. Kilian, editors, *NIPS*, pages 189–196, 2012.

P. Drineas, R. Kannan, and M. W. Mahoney. Fast Monte-Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition. *SIAM Journal on Computing*, 36 (1), 2006.

A. Frank and A. Asuncion. UCI machine learning repository, 2010. URL http://archive.ics.uci.edu/ml.

W. Gander, G. Golub, and U. von Matt. A constrained eigenvalue problem. *Linear Algebra and Its Applications*, 114-115:815–839, 1989.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics, 2001.

K. Kontonasios and T. De Bie. An information-theoretic approach to finding informative noisy tiles in binary databases. In *Proceedings of the 2010 SIAM International Conference on Data Mining (SDM)*, 2010.

D. D. Lee, H. S. Seung, and et al. Learning the parts of objects by non-negative matrix. *Nature*, 401 (6755):788–791, 1999.

J. G. S. Paiva, W. R. Schwartz, H. Pedrini, and R. Minghim. Semi-supervised dimensionality reduction based on partial least squares for visual analysis of high dimensional data. *Computer Graphics Forum*, 31(3pt4):1345–1354, 2012.

F. V. Paulovich, L. G. Nonato, R. Minghim, and H. Levkowitz. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. *IEEE Transactions on Visualization and Computer Graphics*, 14(3):564–575, 2008.

D. Paurat and T. Gärtner. Invis: A tool for interactive visual data analysis. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2013.

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

B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *Proceedings of the 14th Conference on Computational Learning Theory*. Springer, 2001.

C. Shearer. The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing*, 5(4):13–22, 2000.

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

C. Walder, R. Henao, M. Mørup, and L. K. Hansen. Semi-Supervised Kernel PCA. *Computing Research Repository (CoRR)*, abs/1008.1398, 2010.

J. A. Walter and H. Ritter. On interactive visualization of high-dimensional data using the hyperbolic plane. In *Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD 2002, pages 123–132, 2002.

S. Yu, K. Yu, V. Tresp, H. P. Kriegel, and M. Wu. Supervised probabilistic principal component analysis. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD 2006, pages 464–473, 2006.