

Semi-Supervised Kernel PCA

Christian Walder, Ricardo Henao, Morten Mørup and Lars Kai Hansen

Informatics and Mathematical Modelling
Technical University of Denmark, DK-2800
{chwa,rh,mm,lkh}@imm.dtu.dk

Abstract. We present three generalisations of Kernel Principal Components Analysis (KPCA) which incorporate knowledge of the class labels of a subset of the data points. The first, MV-KPCA, penalises within class variances similar to Fisher discriminant analysis. The second, LS-KPCA is a hybrid of least squares regression and kernel PCA. The final LR-KPCA is an iteratively reweighted version of the previous which achieves a sigmoid loss function on the labeled points. We provide a theoretical risk bound as well as illustrative experiments on real and toy data sets.

1 Introduction

In Semi-Supervised Learning (SSL) we are given a set of data points, only some of which come with class labels, and wish to infer a function which classifies new points. Alternatively we may not require the function but only its value on the unlabeled points, as in transduction. A considerable amount of work has recently been done here, see *e.g.* [1,2] for an overview and [3] for a discussion of the problem. Our approach is most closely related to the class of discriminative algorithms exemplified by the transductive support vector machine or T-SVM [4]. The classifying function of this natural semi-supervised extension of the (normal, or fully supervised) SVM can be written

$$f^* = \arg \min_{f \in \mathcal{H}} \|f\|_{\mathcal{H}}^2 + c_1 \sum_{i \in \mathcal{L}} L(f(\mathbf{x}_i), t_i) + c_2 \sum_{j=1}^m U(f(\mathbf{x}_j)),$$

where $\mathbf{x}_i \in \mathcal{X}$ ($t_i \in \pm 1$) are the data points (labels), \mathcal{L} the indices of the labeled points, and \mathcal{H} a Reproducing Kernel Hilbert Space (RKHS). The labeled loss function proposed for the T-SVM is the usual hinge loss $L(f(\mathbf{x}), t) = (1 - tf(\mathbf{x}))_+$ of the normal SVM, while the unlabeled loss $U(f(\mathbf{x})) = (1 - |f(\mathbf{x})|)_+$ is the natural unlabeled analog of L , which we depict in Figure 1. Although it appears to be as sensible as the SVM, the non-convexity of U makes the T-SVM much more difficult to handle, leading to various optimisation strategies [5].

In this paper we propose SSL algorithms which can be thought of either as generalisations of the (normally fully unsupervised) KPCA or as relaxations of the T-SVM in which U takes the simpler form $U(f(\mathbf{x})) = -f(\mathbf{x})^2$. Although

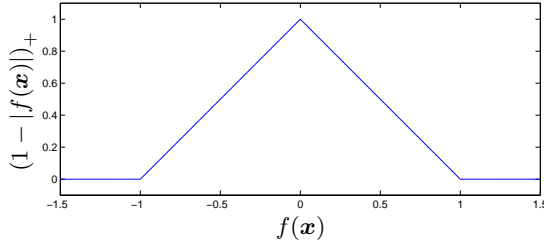


Fig. 1. The transductive SVM loss function for unlabeled points. Penalisation of this non-convex loss favours values $f(\mathbf{x})$ which are either sufficiently positive or sufficiently negative, but not too close to zero.

this term is also non-convex (it is concave), it does lead to computational advantages. In particular, choosing also a quadratic loss for L instead of the hinge, the problem is exactly solvable as we show in Section 3.3. This combination of quadratic losses with exact solution is our least squares or LS-KPCA. Building on the useful exact solvability of this (still non-convex) proxy for T-SVM, we then propose as logistic regression or LR-KPCA, an iteratively reweighted version of LS-KPCA which gets closer to the T-SVM by implementing a sigmoidal loss function L , and utilising the exact solution of LS-KPCA in an inner loop.

1.1 Overview and Organisation of the Paper

We review KPCA in Section 2 from a slightly unusual functional perspective. Our derivation relies on the representer theorem [6], which turns out to make the discussion of our SSL generalisations of KPCA rather clean and straightforward. These generalisations of KPCA make up Section 3. In Section 3.1 we introduce MV-KPCA, which differs from KPCA in that the variance should be small over some prescribed subsets of the data. This is the simplest method we propose in that it is solved by a normal (generalised) eigenvalue problem. We argue in Section 3.2 that this formulation may be problematic. Addressing these problems, in Section 3.3 we introduce LS-KPCA, the method mentioned above with purely quadratic L and U , which enjoys the risk bound we present in Section 5. LS-KPCA represents a greater departure from KPCA than MV-KPCA, but can also be solved exactly due to [7]. In Section 3.4 we move further from KPCA, with an iterative reweighting scheme which utilises this exact solution in an inner loop in order to achieve a sigmoid rather than quadratic loss function, the intuition being that this may be more appropriate for classification problems. A simple yet numerically stable optimisation procedure for LS- and LR-KPCA is outlined in Section 4. In Section 6 we compare our algorithms to previous approaches, focussing on the Spectral Graph Transducer (SGT) of [8]. We present results on standard benchmark data sets in Section 7, and finish with some conclusions in Section 8.

2 Kernel PCA

We treat KPCA [9] slightly differently than usual, as the problem of finding

$$f^* = \arg \max_{f \in \mathcal{H}} \sum_{i=1}^m \left(f(\mathbf{x}_i) - \frac{1}{m} \sum_j f(\mathbf{x}_j) \right)^2 \quad (1)$$

$$\text{subject to } \|f\|_{\mathcal{H}}^2 = 1, \quad (2)$$

where \mathcal{H} is the RKHS with kernel $k(\cdot, \cdot)$. The Lagrangian function associated with this problem is

$$L(f, \lambda) = \sum_{i=1}^m \left(f(\mathbf{x}_i) - \frac{1}{m} \sum_j f(\mathbf{x}_j) \right)^2 + \lambda (\|f\|_{\mathcal{H}}^2 - 1).$$

For the Lagrangian dual we maximise $L(f, \lambda)$ over $f \in \mathcal{H}$. The representer theorem [6], then implies $f^*(\mathbf{x}) = \sum_{j=1}^m \alpha_j^* k(\mathbf{x}_j, \mathbf{x})$ for some $\alpha_1, \alpha_2, \dots, \alpha_m \in \mathbb{R}$. Combined with the reproducing property $f(\mathbf{x}) = \langle f, k(\mathbf{x}, \cdot) \rangle_{\mathcal{H}}$, we obtain the simplification of (1) and (2) to

$$\boldsymbol{\alpha}^* = \arg \max_{\boldsymbol{\alpha} \in \mathbb{R}^m} \boldsymbol{\alpha}^\top (K^\top K - K^\top E_m K) \boldsymbol{\alpha} \quad (3)$$

$$\text{subject to } \boldsymbol{\alpha}^\top K \boldsymbol{\alpha} = 1, \quad (4)$$

where $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ and E_m is a square matrix of size m with entries $\frac{1}{m}$, as it shall be throughout the paper. Imposing stationarity in $\boldsymbol{\alpha}$ on the Lagrangian of (3) and (4) we find that $\boldsymbol{\alpha}^*$ is the eigenvector $\boldsymbol{\alpha}$ with largest eigenvalue λ of the generalised eigenvalue problem

$$K \boldsymbol{\alpha} = \lambda (K^\top K - K^\top E_m K) \boldsymbol{\alpha}. \quad (5)$$

It is easy to verify that this formulation of KPCA is equivalent up to a re-normalisation to the original one [9] in its *centered* form. The simpler *uncentered* version assumes that the data is centered in feature space. This leads to a slightly different eigenproblem formed by replacing the term $(K^\top K - K^\top E_m K)$ with $K^\top K$ in (5). Here we can recover the uncentered version of KPCA by replacing the objective in (1) with $\sum_{i=1}^m f(\mathbf{x}_i)^2$. That is, the variance of the function values $f(\mathbf{x}_i)$ assuming they have zero mean.

3 Semi Supervised Kernel PCA

We propose three means of incorporating label information into KPCA, with MV-KPCA (Section 3.1) incorporating a slightly different type of label information than LS- and LR-KPCA (Sections 3.3 and 3.4).

3.1 Minimum Variance Kernel PCA (MV-KPCA)

We begin with MV-KPCA, which incorporates knowledge of pairwise, or rather group-wise similarity. This is the simplest method in that it is solved by an eigenproblem very similar to that of KPCA, and also in that it involves one extra parameter rather than two. The idea of MV-KPCA is, reminiscent of the Kernel Fisher Discriminant [10], to modify the constraint (2) by adding a loss term based on the within-class variances. Given prescribed index sets $G_1, G_2, \dots, G_l \subset \{1, 2, \dots, m\}$ of similar elements from the data set $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m$, the new constraint is

$$\|f\|_{\mathcal{H}}^2 + c \sum_{j=1}^l \sum_{i \in G_j} \left(f(\mathbf{x}_i) - \frac{1}{|G_j|} \sum_{j' \in G_j} f(\mathbf{x}_{j'}) \right)^2 = 1,$$

where $c \in \mathbb{R}^+$ trades between KPCA for $c = 0$ and increasing penalisation of within class variance for larger values of c . Note that (as in our experiments in this paper) the G_j may be derived from categorical class labels by assigning points with the same label to the same group. Once again we can apply the representer theorem to the Lagrangian of this problem. Since the augmented objective function is purely quadratic (like that of the original KPCA), its optimal solution is again found by an eigenproblem, this time

$$\left(K + c \left(\sum_i K_i^\top K_i - K_i^\top E_{|G_i|} K_i \right) \right) \alpha = \lambda (K^\top K - K^\top E_m K) \alpha, \quad (6)$$

where K_i is the sub-matrix of K taking rows G_i . Note that it is not possible to obtain a convex problem by replacing the equality constraint with an inequality — although the resulting problem is equivalent but on a convex feasible region, in this case we would still be *maximising* a convex function over that region.

3.2 Difficulties Parameterising MV-KPCA

Since the objective function (1) and constraint (2) in KPCA (and MV-KPCA) both scale the same way (quadratically), the maximisation of one with the other fixed is equivalent to the maximisation of the ratio of the two. This is often referred to as the *Rayleigh quotient* form of an eigenvalue problem [11]. A consequence is that changing the constant on the right hand side of (2) only rescales f^* , which could be problematic as we now argue. The objective we are maximising is

$$\max_{f \in \mathcal{H}} \frac{\text{VAR}[f]}{\|f\|_{\mathcal{H}}^2 + c \sum_i \text{VAR}_i[f]}, \quad (7)$$

where $\text{VAR}[f]$ is the variance of the values of f over all the \mathbf{x}_i and $\text{VAR}_i[f]$ is the variance of the values of f over the points indexed by G_i . This looks like it may be interesting for SSL. After all, the objective function favours large variance on

the unlabeled points while favouring small values of two fairly standard terms for regularisation and risk, namely an RKHS norm and a type of quadratic penalty. More importantly, although this may appear to be precisely the type of semi-supervised learning objective which tends to be hard to optimise due to its non-convexity, we can solve it in $\mathcal{O}(m^3)$ time due to the convenient relationship with the eigenvalue problem (6). The unfortunate part however, is that due to the fact that changing the constraint in (2) only multiplicatively scales the solution, there is no obvious way to trade between the numerator and the denominator of (7) in the same way we can trade off within the denominator via the parameter c . This could be critical in SSL problems in which there are vastly different numbers of labeled and unlabeled points. For a second multiplicative scaling parameter in the ratio (7) to be non-trivial however, at least one of the terms would need to scale non-quadratically.

3.3 Least Squares Kernel PCA (LS-KPCA)

Continuing the previous argument, although many non purely quadratic surrogates for any of the three terms in (7) are possible, few of the interesting ones will lead to computational problems as straightforward as solving the eigenvalue problem (6). Fortunately however, the classic squared loss does turn out to be fairly convenient. Hence, as LS-KPCA we now propose the following modification of KPCA. First, instead of maximising the first term (the objective (1)) with the second term (the constraint (2)) fixed, we minimise the second with the first fixed. Actually, this is still KPCA as these problems are of course equivalent. Next, we add onto the new objective function a squared loss term, to get

$$f^* = \arg \min_{f \in \mathcal{H}} \|f\|_{\mathcal{H}}^2 + c \sum_{i \in \mathcal{L}} (f(\mathbf{x}_i) - y_i)^2 \quad (8)$$

$$\text{subject to } \text{VAR}[f] = s^2. \quad (9)$$

An example solution of the above problem is depicted in Figure 2. Unlike the original KPCA constraint and MV-KPCA, the above constraint does break the scale invariance of the ratio of the objective function and the constraint function and the problem cannot be written as a ratio similar to (7). In other words, the part of (8) which is linear in f makes the relationship between s and the corresponding optimal f^* non-trivial. Furthermore, although the parameterisation is unusual, we are now able to control the relative importance of the three terms, $\text{VAR}[f]$, $\|f\|_{\mathcal{H}}^2$ and the squared error part of (8), via the parameters c and s^2 . Applying the representer theorem as before yields

$$\boldsymbol{\alpha}^* = \arg \min_{\boldsymbol{\alpha} \in \mathbb{R}^m} \boldsymbol{\alpha}^\top K \boldsymbol{\alpha} + c \|K_{\mathcal{L}} \boldsymbol{\alpha} - \mathbf{t}\|^2 \quad (10)$$

$$\text{subject to } \boldsymbol{\alpha}^\top (K^\top K - K^\top E_m K) \boldsymbol{\alpha} = s^2, \quad (11)$$

where $\mathbf{t} \in \mathbb{R}^{|\mathcal{L}|}$ is the sub-vector of \mathbf{y} taking indices \mathcal{L} , and $K_{\mathcal{L}}$ is the submatrix of K taking rows \mathcal{L} . To solve (10) and (11) we can make use of the ideas in [7],

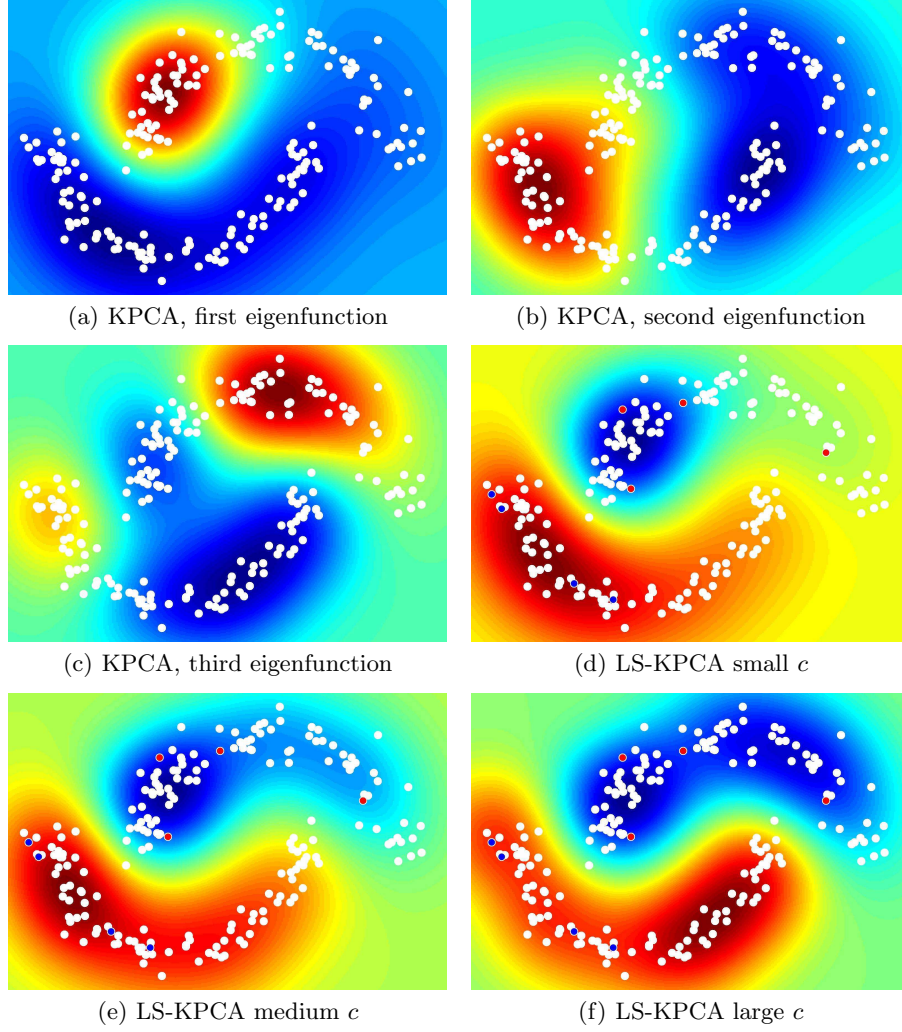


Fig. 2. An example in \mathbb{R}^2 of unlabeled (white) and labeled (red/blue, four per class) points using a spherical Gaussian kernel on the *two moons* toy dataset. The value of f^* (from equation (1) for (a)-(c) and (8) for (d)-(f)) is rendered in colour ranging from around +3 (dark red) to -3 (dark blue). LS-KPCA approaches KPCA for small c , hence there is a smooth transition (a)-(d)-(e)-(f), ignoring the sign change in (a) which is arbitrary for KPCA.

which studies the problem

$$\mathbf{z}^* = \arg \min_{\mathbf{z} \in \mathbb{R}^m} \mathbf{z}^\top G \mathbf{z} - 2\mathbf{b}^\top \mathbf{z} \quad (12)$$

$$\text{subject to } \mathbf{z}^\top \mathbf{z} = s^2. \quad (13)$$

It is shown that the solution to this non-convex problem is $\mathbf{z}^* = (G - \lambda^* I)^{-1} \mathbf{b}$, where λ^* is the smallest eigenvalue of the problem

$$\begin{pmatrix} G & -I \\ -\frac{1}{s^2} \mathbf{b} \mathbf{b}^\top & G \end{pmatrix} \begin{pmatrix} \gamma \\ \eta \end{pmatrix} = \lambda \begin{pmatrix} \gamma \\ \eta \end{pmatrix}.$$

Note that this result was used in a related context [8], as we discuss in Section 6. Making the change of variables $\mathbf{z} = P^{\frac{1}{2}} \boldsymbol{\alpha}$ where

$$P = K^\top K - K^\top E_m K,$$

we can use this result to derive that

$$\boldsymbol{\alpha}^* = (C - \zeta^* P)^{-1} \mathbf{b}, \quad (14)$$

where $C = K + cK_\mathcal{L}^\top K_\mathcal{L}$, $\mathbf{b} = cK_\mathcal{L} \mathbf{t}$ and ζ^* is the smallest eigenvalue of the generalised eigenvalue problem

$$\begin{pmatrix} C & -P \\ -\frac{1}{s^2} \mathbf{b} \mathbf{b}^\top & C \end{pmatrix} \begin{pmatrix} \gamma \\ \eta \end{pmatrix} = \zeta \begin{pmatrix} P & \mathbf{0} \\ \mathbf{0} & P \end{pmatrix} \begin{pmatrix} \gamma \\ \eta \end{pmatrix}. \quad (15)$$

The change of variables is unnecessary however, as we can repeat the arguments in [7] with the constraint in (13) replaced by $\mathbf{z}^\top P \mathbf{z} = s^2$.

3.4 Logistically Loss via Reweighting (LR-KPCA)

Generalising LS-KPCA to arbitrary L and U for the labeled and unlabeled loss functions, we get

$$f^* = \arg \min_{f \in \mathcal{H}} \|f\|_{\mathcal{H}}^2 + c \sum_{i \in \mathcal{L}} L(f(\mathbf{x}_i), y_i) \quad (16)$$

$$\text{subject to } \sum_i U(f(\mathbf{x}_i)) = s^2. \quad (17)$$

Note that the U we intend here and for the remainder of the paper differs from that of the T-SVM formulation in Section 1 by a sign change. The purely quadratic losses of LS-KPA may not be appropriate for classification. Leaving L and U unspecified but abusing the notation by extending them element-wise to vectors, the representer theorem still applies, so we can write the problem in $\boldsymbol{\alpha}$ as

$$\boldsymbol{\alpha}^* = \arg \min_{\boldsymbol{\alpha} \in \mathbb{R}^m} \boldsymbol{\alpha}^\top K \boldsymbol{\alpha} + c \mathbf{e}^\top L(K \boldsymbol{\alpha}, \mathbf{y}) \quad (18)$$

$$\text{subject to } \mathbf{e}^\top U(K \boldsymbol{\alpha}) = s^2, \quad (19)$$

where \mathbf{e} is a vector of ones. We would like to use more sophisticated losses U and L in the above formulation. For this, it is natural to try to leverage the powerful result that we can solve the least squares formulation exactly, by employing the *iteratively reweighted least squares* idea [12]. This is essentially a Newton-Raphson method, but with the interpretation that each step solves a least squares problem with modified weights.

A Newton-Raphson step solves a local second order approximation of the problem. To be able to apply the exact solution of the previous section, we are forced to choose a U which is purely second order (*i.e.* with no linear term). Then the local second order approximation of the constraint (19) is still purely second order (and exact), and the form of the optimisation problem remains that of LS-KPCA. Hence we maintain our initial choice $U(f(\mathbf{x})) = f(\mathbf{x})^2$. We can try to improve on L however, as doing so does not change the form of the objective function on taking a local second order approximation, since the LS-KPCA objective (10) already has a linear part. Hence, motivated by logistic regression, as LR-KPCA we propose the sigmoid

$$L(f(\mathbf{x}), y) = 1/(1 + \exp(-yf(\mathbf{x})))$$

as the loss term for labeled points. A Huber-like differentiable approximation of the support vector machine hinge loss could just as easily be used, however.

By arguments similar to those of logistic regression [12], we can solve (18)-(19) by iteratively reweighted LS-KPCA. We can derive as usual that the following steps constitute a Newton-Raphson update. Given the current solution $\boldsymbol{\alpha}_n$, we compute $\mathbf{g}, \mathbf{z}, \mathbf{s} \in \mathbb{R}^{|\mathcal{L}|}$ as $\mathbf{g} = K_{\mathcal{L}}\boldsymbol{\alpha}_n$, and

$$\begin{aligned} z_i &= 1/(1 + \exp(-t_i g_i)), \\ r_i &= z_i(1 - z_i), \\ s_i &= g_i - (z_i - t_i)(1 - z_i)/z_i, \end{aligned} \tag{20}$$

for $i = 1, 2, \dots, |\mathcal{L}|$. The next iterate $\boldsymbol{\alpha}_{n+1}$ is defined like $\boldsymbol{\alpha}^*$ of (14) and (15), but with a different C and \mathbf{b} , which now depend on \mathbf{r} and \mathbf{s} according to

$$C = K + cK_{\mathcal{L}}^{\top} R K_{\mathcal{L}}, \quad \mathbf{b} = cK_{\mathcal{L}}^{\top} R \mathbf{s},$$

where R is diagonal with $R_{ii} = r_i$.

Due to the form of the logistic function (20) we have that $r_i > 0$ and so the resulting Hessian C is always positive definite. As is also the case for normal logistic regression however, we have no guarantee that it will improve the objective function, making some form of back-tracking line search necessary. Due to the constraint (17), this is not as simple as moving back on the line $\lambda\boldsymbol{\alpha}_n + (1 - \lambda)\boldsymbol{\alpha}_{n-1}$, $0 \leq \lambda \leq 1$. Instead, in order to guarantee convergence we check the objective function, and as long as it is not better than the previous iterate, we solve a modified problem with an additional regularisation term $\lambda\|\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_{n-1}\|^2$, where λ is a parameter we increase until we see an improvement in the (unmodified) objective function. It is important to note that similar line search heuristics are also required in the iteratively reweighted maximum likelihood solver of the standard logistic regression model.

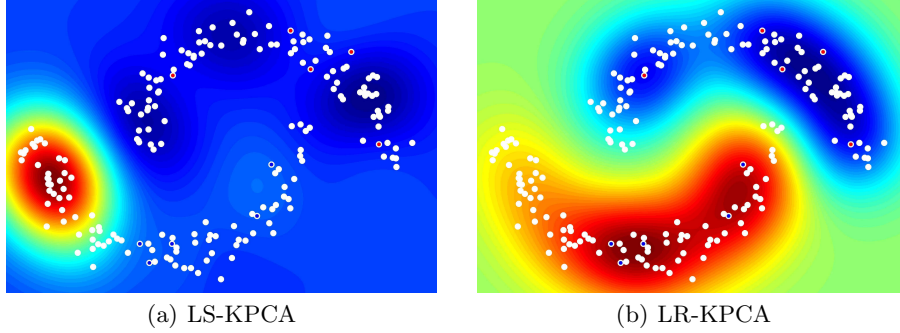


Fig. 3. LS-KPCA (*left*) suffers here due to the value of s^2 in (11) being large. This essentially enforces a large squared value of the function at certain points, conflicting with the squared loss which favours values near ± 1 , so that the energy of the function gets concentrated away from the labeled points. The sigmoid loss of LR-KPCA (*right*) can discount labeled points which are well classified, leaving the energy unhampered. The values range from around -50 (dark red) to +10 (dark blue) for LS-KPCA, and from around -3 to +3 for LR-KPCA.

4 Efficient Solution

We need to compute the ζ^* of (14), both for LS-KPCA and the inner loop of LR-KPCA. As argued in [7], doing so via the eigenvalue equation (15) directly can be highly unstable since the matrix on the left hand side is not symmetric, is twice the size of the normal KPCA eigenvalue problem, and is typically badly conditioned. From (11), the solution α^* must satisfy $\alpha^\top P \alpha = s^2$. As it was shown in [7], for $\zeta < \delta$ where δ is the smallest eigenvalue of C , (12) and (13) have a unique solution if and only if the characteristic polynomial (or secular equation) $f(\zeta) = \alpha^\top P \alpha - s^2 = 0$ is satisfied, provided that $f(\zeta)$ is strictly increasing for $\zeta \in (-\infty, \delta)$. As a result, to obtain α^* we need to find the unique root of $f(\zeta)$ to the left of $\delta - |\mathbf{u}^\top \mathbf{b}|/s$, where \mathbf{u} is the eigenvector with associated eigenvalue δ . Here we use Brent's method [13], allowing ζ^* to be calculated to high precision. Note that the uncentered $\text{VAR}[f]$ allows us to make a straight-forward change of variables $\mathbf{f} = K\alpha$ and solve the problem in \mathbf{f} . The resulting eigenvalue problem is of the form (12)-(13), with an isotropic constraint, and can be solved more efficiently via (to use their terminology) the simpler *explicit* characteristic polynomial of [7], as opposed to the *implicit* one used here. Transforming to an isotropically constrained variable in this way would be possible for the centered formulation as well, however this transformation leads to numerical problems due to the required matrix inverse. Moreover, the inverse in (14) means that it is not possible to reduce the computational time complexity from cubic.

5 Risk Bound

It is straightforward to obtain a risk bound for LS-KPCA using [14], with the analysis being similar to that of the SGT in that work. Assume the uncentered version of (9), so that $P = K^\top K$. From (10) and (11) we have that the soft decision function values $\mathbf{z}^* = K\boldsymbol{\alpha}^*$, an Unlabeled-Labeled Decomposition (ULD). Letting $m = l + n$, where l (resp. n) is the number of labeled (unlabeled) points, we can obtain an error bound for LS-KPCA

Theorem 1 ([14]).

Let $\mathbf{z}^ = K\boldsymbol{\alpha}^*$ be the ULD of a transductive algorithm. Let $\|\boldsymbol{\alpha}\|_2 \leq \mu$, $c = \sqrt{32 \ln(4e)/3} < 5.05$ and $r = 1/l + 1/n$. With probability of at least $1 - \delta$ over the choice of the training set \mathcal{L} from X , for all \mathbf{z} in the set of all possible hypothesis that can be generated by the algorithm for a given X , all possible partitions and all possible labelings of the training set, the risk $\mathcal{R}_n(\mathbf{z})$ is bounded from above by*

$$\mathcal{R}_l(\mathbf{z}) + \sqrt{\frac{2\mu^2}{ln}} \|K\|_{\text{Fro}}^2 + cr\sqrt{\min(l, n)} + \sqrt{2r \ln \frac{1}{\delta}}.$$

To apply this to LS-KPCA, we replace $\mathbf{z}^* = K\boldsymbol{\alpha}^*$ in (10) and (11), eigen-decompose $G = K^{-1}CK^{-1} = QDQ^\top$ (so that $Q^\top Q = I$) and put $\boldsymbol{\alpha} = Q^\top \mathbf{z}$. We see $\mathbf{z}^* = Q\boldsymbol{\alpha}^*$ with $\boldsymbol{\alpha}^*$ as in (16). Since Q depends only on the data \mathbf{X} , the latter is a ULD of LS-KPCA. From (12) $\boldsymbol{\alpha}^\top \boldsymbol{\alpha} = s^2$ and $\|Q\|_{\text{Fro}}^2 = q$, where q is the rank of G . Hence $\mathcal{R}_n(\mathbf{z})$ is bounded from above by

$$\mathcal{R}_l(\mathbf{z}) + \sqrt{\frac{2qs^2}{ln}} + cr\sqrt{\min(l, n)} + \sqrt{2r \ln \frac{1}{\delta}}.$$

The second term is an upper bound on the Rademacher complexity of ULD algorithms, for the SGT it is $\sqrt{2\tilde{q}r}$ [14], where \tilde{q} is the number of non-zero eigenvalues of the Laplacian. Since both algorithms use a squared loss for \mathcal{R}_l , their bounds differ only by the Rademacher complexity, which may even be made equal by choosing s appropriately. This is to be expected, since as we explain in the next section LS-KPCA differs from the SGT only by its regulariser.

6 Relationship to Other Methods

Firstly, MV-KPCA is similar to the Kernel Fisher Discriminant, penalising a different set of variances but also leading to an eigenvalue problem [10]. Next, LS-KPCA is related to [8] and [15], the former relationship being the clearest. In particular, following [16] we can interpret Joachims' SGT [8] as a special case of LS-KPCA. To do this, in (8) and (9) we define the RKHS of functions as the set of real valued functions defined on the vertices of the graph and satisfying a particular linear constraint (normalised cut balancing constraint, related to our centered variance for (9)) so that $\mathcal{H} = \{\mathbf{f} \in \mathbb{R}^m : \mathbf{f}^\top \mathbf{e} = 0\}$ where \mathbf{e} is a vector

of ones. We define the graph Laplacian matrix L as in [8], and let the kernel matrix K be given by L^+ , the pseudo-inverse of L . If we further restrict to the simpler uncentered version of (9), and use the fact that the first eigenvector of L is a scaled \mathbf{e} , then simplifications lead to equations (19) and (20) in [8]. Hence the SGT is LS-KPCA with an RKHS defined by a graph based regulariser. Such regularisers have proven highly effective in SSL. A similar combination was proposed in [17] but with a non-convex gradient descent and more sophisticated loss functions.

The RKHS derivation of SGT has various advantages. First, our experiments show that in some cases it can be more effective to use a normal Gaussian kernel RKHS regulariser rather than a graph based one. As we see in Section 7, this happens particularly when the data density is adversarial in the sense of defying the so-called manifold assumption (that the data lie near a low dimensional sub-manifold of the input space [17]), in which case the graph based regulariser may be inappropriate. It is also straightforward to smoothly transition between the two regularisers as in [18]. This transitioning can equivalently be obtained by simply including the graph Laplacian regulariser as an additional term in (16) of the form $\sum_{i,j} w_{ij} (f(\mathbf{x}_i) - f(\mathbf{x}_j))^2$, since the overall problem can be converted just as easily as before to an optimisation problem in α via the Lagrangian/representer theorem. Various other interesting options are straightforward due to this flexibility. For example, problem specific invariances may be incorporated as in [19] by penalising loss terms which are functions of the gradient of f . This simply leads to an expansion for the optimal f^* which contains gradients of the kernel function [6], and this leads immediately to finite dimensional optimisations similar to those in α in our formulations. Finally, compared with the graph cut derivation of the SGT our RKHS derivation permits a natural out of sample extension.

To complete our comparison with other methods let us finally mention LR-KPCA. This algorithm is a greater departure from previous work. It is related to logistic regression and of course LS-KPCA, but seems to be the first iterative algorithm to take advantage of the exact solution of LS-KPCA provided by [7] as part of an inner loop.

7 Experiments and Discussion

We tested on the six binary benchmark data sets of [2] as follows.

7.1 Gaussian Kernel

Each error in Table 1 corresponds to a mean (standard deviation) over the twelve test splits supplied with the data sets, for each of the two supplied cases: 10 (top half of table) and 100 (bottom half) labeled points. We used the Gaussian kernel $k(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2)$, so for each split we had to choose the parameters c (for LS- and LR-KPCA), s (for LS-, LR- and MV-KPCA) and γ (for all three and also KPCA). To choose these parameters for each split, we performed 10

	<i>g241c</i>	<i>g241d</i>	<i>Digit1</i>	<i>USPS</i>	<i>BCI</i>	<i>Text</i>
LR-KPCA	15.09 (4.57)	49.71 (3.96)	<i>15.52</i> (4.01)	<i>21.41</i> (3.06)	<i>47.46</i> (1.75)	<i>32.5</i> (3.57)
LS-KPCA	14.70 (1.83)	48.74 (3.91)	<i>13.86</i> (2.99)	<i>23.75</i> (3.20)	<i>48.44</i> (2.57)	<i>32.71</i> (3.00)
MV-KPCA	14.12 (2.28)	50.24 (4.03)	<i>12.32</i> (3.69)	62.02 (12.70)	<i>50.04</i> (0.89)	<i>32.47</i> (3.35)
MV-KPCA-10	36.34 (7.54)	<i>32.96</i> (8.56)	21.37 (4.69)	55.59 (3.92)	<i>48.46</i> (2.15)	<i>34.74</i> (7.88)
KPCA-10	28.75 (4.26)	<i>32.98</i> (8.58)	21.36 (4.67)	35.27 (13.59)	<i>48.16</i> (2.91)	36.00 (9.76)
LR-KPCA	12.64 (0.46)	23.8 (3.00)	<i>4.60</i> (1.07)	11.35 (1.87)	<i>32.25</i> (2.04)	27.66 (2.88)
LS-KPCA	13.12 (0.45)	22.93 (3.19)	<i>4.08</i> (1.34)	<i>7.51</i> (1.01)	29.03 (2.20)	<i>24.96</i> (2.61)
MV-KPCA	12.82 (0.42)	49.95 (1.64)	<i>3.89</i> (1.08)	20.11 (5.61)	49.33 (1.59)	30.80 (2.03)
MV-KPCA-10	17.78 (1.63)	<i>17.04</i> (2.70)	9.40 (1.59)	10.08 (1.82)	37.17 (3.76)	27.50 (1.57)
KPCA-10	18.11 (1.85)	21.14 (3.06)	7.42 (1.41)	14.46 (1.53)	48.50 (1.71)	33.79 (3.59)

Table 1. *mean (std-dev)* errors over 12 splits for 10 (top half) and 100 (bottom) labeled points out of 1500, on the benchmark sets of [2].

fold cross validation over the labeled points of that split. This model selection procedure can be problematic, especially for the 10 labels case in which the cross validation estimate is especially unreliable, but to be fair we always followed this procedure. Italics in the table indicate significantly best results amongst our algorithms, in the sense of having a mean error less than the mean error plus standard deviation of the method with lowest mean. Bold indicates best mean result over all published results in the study [2].

The algorithms listed in Table 1 are the following. LS- and MV-KPCA correspond to using the first eigenfunction from those problems as the classifying function. For MV-KPCA-10 and KPCA-10 however we took the top ten eigenfunctions, and used the values of these functions to train a hard margin, linear SVM (which sees these ten values on the labeled points only). Our motivation for testing MV-KPCA-10 was that by penalising within class variances rather than a signed class label loss term, this algorithm may be flexible enough to extract multiple relevant features. We include KPCA-10 as a baseline for comparison. The SVM was also applied to the first eigenfunction of LR-, LS- and MV-KPCA, although there the optimisation is trivial and merely sets a threshold.

Directly comparable error rates for eleven other algorithms are included in [2], and the chapter with these numbers is freely available online. It turns out that we obtain best mean performance compared with all methods on *g241c* (for 10 and 100 labels) and for *BCI* (for 100 labels only), and generally competitive performance overall. Comparing the different methods it turns out that one of the strongest competitors is the SGT [8], which appears overall to be significantly better than our methods on these data sets. However, as explained in Section 6 the SGT is in fact a special case of our LS-KPCA. The main difference lies in the graph based regularisation of the SGT rather than our plain Gaussian kernel RKHS norm regularisation. Other more subtle differences include Joachims’ choice of graph connectivity and weights, use of a non-trivial spectral renormalisation of the graph Laplacian, and rebalancing based on relative class frequencies [8]. Since these options are also possible for the more general LS-KPCA, we can argue that LS-KPCA should be attributed with the best performance of the published SGT results and our results here. Most importantly, it is clear that the most meaningful features of our results are hence captured

in the relative performances of our different methods. However, one point we can make with regard to absolute performance measures, is that our best performance on *g241c* (a sythetic data set composed as samples from two highly overlapping Gaussians, one per class) seems to indicate that LS- and MV-KPCA are able to handle non manifold like data effectively, presumably due to the non graph based regularisation.

Our algorithms do utilise the unlabeled examples, as we significantly outperform the purely supervised baseline methods [2]. This is further evidenced by the fact that KPCA-10 is never significantly better than the other variants, and often significantly worse. The mediocre performance of KPCA-10 on these datasets is surprising, as this algorithm seems reasonable for SSL, and was previously proposed for exactly that [20]. We also found that LR-KPCA is rather similar to LS-KPCA. This does not seem to be due to computational problems since the iterative reweighting scheme always converged to high precision within 20 iterations. It may be due to the coarse grid we were forced to use for the cross validation parameter search of LR-KPCA, due to its being rather expensive to solve. It is expensive since each re-weighting step requires the LS-KPCA type solution, which itself requires of the order of 10 matrix inverses of size m during the zero finding phase described in Section 4. Although expensive, a more refined search would be possible with sufficient computational resources, but would presumably only lead to modest improvement. Rather, it seems that the squared loss of LS-KPCA is reasonable for these problems, which agrees with the fact that a significant amount of work has been in precisely the opposite direction to our LS-KPCA \rightarrow LR-KPCA. By this we mean the least squares SVM [21] where the hinge loss is actually replaced by a squared loss for classification (although the use of a squared classification loss is relatively uncommon overall in the literature). Moreover, in SSL the labeled loss term plays a diminished role in comparison to normal supervised learning as in the LS-SVM. Nonetheless, LR-KPCA performs strongly and intuitively on the two moons toy dataset as depicted in Figure 3.

7.2 Combined Graph Diffusion and Gaussian Kernel

The main difference between our LS-KPCA and the SGT [8] is the extra degree of freedom afforded the fact that LS-KPCA may utilise an arbitrary kernel function, rather than being restricted to a graph based representation. To demonstrate the value of this degree of freedom we now experiment with LS-KPCA using a convex combination of a graph diffusion and a normal Gaussian kernel, namely

$$K = wK_\gamma + (1 - w) \exp(-\tau L),$$

where K_γ is the kernel matrix associated with the Gaussian kernel as in the previous sub-section. L is the normalised graph Laplacian defined by $L = \text{diag}(S\mathbf{e}) - S$, where $S = \text{diag}(W\mathbf{e})^{-\frac{1}{2}} W \text{diag}(W\mathbf{e})^{-\frac{1}{2}}$. Here W is the edge weight matrix for the graph. To construct W we employed a standard nearest neighbour connectivity, and assigned Gaussian edge weights with respect to the pairwise

	Digit1	USPS	BCI	g241c	COIL	g241d	Text
<i>G-LS</i>	18.15 (7.79)	25.37 (13.47)	48.93 (2.33)	42.29 (7.91)	64.42 (5.21)	46.81 (4.00)	41.85 (7.08)
<i>M-LS</i>	15.95 (7.11)	25.70 (11.38)	48.93 (1.74)	36.20 (13.70)	73.58 (10.87)	47.66 (3.80)	39.92 (7.60)
<i>G-LS</i>	2.70 (0.95)	5.75 (1.28)	48.06 (2.64)	29.36 (6.19)	25.33 (10.96)	32.50 (6.06)	26.67 (2.35)
<i>M-LS</i>	3.88 (3.00)	5.70 (2.43)	37.33 (7.93)	16.63 (3.54)	31.98 (23.72)	24.02 (4.34)	25.71 (2.04)

Table 2. Mean and standard deviation percentage errors for 10 (top half of table) and 100 (bottom half) labeled points out of 1500. *G-LS* is LS-KPCA with a graph diffusion kernel, while *M-LS* is LS-KPCA with a combined Gaussian and graph diffusion kernel.

Euclidean distance of connected vertices. We set the bandwidth of this edge weight Gaussian to be equal to the mean squared pairwise distance between connected points, and removed self connections so that W has zero entries on the main diagonal. Starting with $w = 0$ (pure graph kernel), we chose the number of nearest neighbours, τ , and the parameters c and s of (8)-(9) using a leave one out procedure which we accelerated by exploiting Cholesky up- and down-dates. This pure graph kernel based method is listed as *G-LS* in Table 2. Given those optimal parameters, we then fixed τ and the number of nearest neighbours, and then selected w , γ , c and s again using leave one out, in order to assess the relative benefit of using a non graph based kernel in this experimental setting. The results for the case of mixing the Gaussian and diffusion kernels is listed in Table 2 as *M-LS*. We see that incorporating the Gaussian kernel in this manner never degrades the performance of the pure graph based algorithm, while for data sets *g241c* and *BCI* it significantly improves the performance.

8 Conclusions

We have proposed three variants of KPCA for semi-supervised learning. All three are able to benefit from the unlabeled data, and lead to competitive overall results on benchmark sets. LS-KPCA generalises the powerful SGT algorithm [8], thereby admitting various alternative algorithms due to the flexibility of the RKHS setting. Moreover, our RKHS based derivation of the more general case is conceptually cleaner than that of the SGT, which was originally derived from a relaxed spectral graph cut perspective. Both LS-KPCA and the SGT utilise [7] to obtain the globally optimal solution to their non-convex optimisation problems. We interpret this as a useful tool for problems related to that of the T-SVM, by considering the variance term in LS-KPCA as analogous to the T-SVM unlabeled loss function. We also proposed the more sophisticated LR-KPCA, which implements a classification oriented sigmoid loss function via a reweighting scheme. This reweighting scheme also utilises the globally optimal solution of LS-KPCA in an inner loop.

Generally speaking, we believe that the formulations in [7] are powerful, and perhaps under-utilised in machine learning. We hope to uncover a family of interesting algorithms (particularly for semi supervised learning) by studying re-weighted versions of these formulations. Here we presented an iterative re-weighting of the loss in (16) (as in LR-KPCA). Also interesting is the possibility

of reweighting the summand in (17) in order to obtain more sophisticated unlabeled loss terms. This is more complex, and we plan to investigate this direction in the future.

References

1. Zhu, X.: Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison (2007)
2. Chapelle, O., Schölkopf, B., Zien, A., eds.: Semi-Supervised Learning. MIT Press, Cambridge (2006)
3. Seeger, M.: Learning with labeled and unlabeled data. Technical report, Univ. Edinburgh (Dec. 2002)
4. Vapnik, V.: Statistical Learning Theory. John Wiley and Sons, inc., New York (1998)
5. Chapelle, O., Sindhwani, V., Keerthi, S.S.: Optimization techniques for semi-supervised support vector machines. *JMLR* **9** (2008) 203–233
6. Kimeldorf, G., Wahba, G.: Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Appl.* **33** (1971) 82–95
7. Gander, W., Golub, G., von Matt, U.: A constrained eigenvalue problem. *Linear Algebra and its Appl.* **114**(115) (1989) 815–839
8. Joachims, T.: Transductive learning via spectral graph partitioning. In: *ICML*. (2003) 290–297
9. Schölkopf, B., Smola, A., Müller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation* **10** (1998) 1299–1319
10. Mika, S., Rätsch, G., Weston, J., Schölkopf, B., Smola, A., Müller, K.R.: Constructing descriptive and discriminative non-linear features: Rayleigh coefficients in feature spaces. *IEEE PAMI* **25**(5) (2003) 623–628
11. Horn, R.A., Johnson, C.R.: *Matrix Analysis*. Cambridge University Press (1990)
12. McCullagh, P., Nelder, J.: *Generalized Linear Models*, Second Edition. Chapman & Hall (1989)
13. Brent, R.P.: *Algorithms for Minimization without Derivatives*. Prentice-Hall, Englewood Cliffs, N.J. (1973)
14. El-Yaniv, R., Pechyony, D.: Transductive Rademacher complexity and its applications. In: *COLT*. Volume 4539 of *Lecture Notes in Computer Science*., Springer (2007) 157–171
15. El-Yaniv, R., Pechyony, D., Vapnik, V.: Large margin vs. large volume in transductive learning. *Mach. Learn.* **72**(3) (2008)
16. Herbster, M., Pontil, M., Wainer, L.: Online learning over graphs. In: *ICML '05: Proceedings of the 22nd international conference on Machine learning*, New York, NY, USA, ACM (2005) 305–312
17. Chapelle, O., Zien, A.: Semi-supervised classification by low density separation. In: *AISTATS 2005*, Society for Artificial Intelligence and Statistics (2005) 57–64
18. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR* **7** (November 2006) 2399–2434
19. Chapelle, O., Schölkopf, B.: Incorporating invariances in nonlinear support vector machines. In: *Advances in Neural Information Processing Systems 14*. MIT Press, Cambridge, MA (2001) 609–616

20. Su, W., Carpuat, M., Wu, D.: Semi-supervised training of a kernel pca-based model for word sense disambiguation. In: Proc. of the 20th intl. conf. on Computational Linguistics. (2004)
21. Suykens, J.A.K., Vandewalle, J.: Least squares support vector machine classifiers. Neural Proc. Lett. **9**(3) (1999)