Contents

1 Basic kernel spectral clustering algorithms **3**
1.1 First section 4

2 From classical to least-square support vector machines **5**
2.1 First section 6

3 Kernel spectral clustering as weighted kernel PCA **7**
3.1 Classical and SVM formulation of linear PCA 8
3.1.1 Classical linear principal component analysis formulation 8
3.1.2 Support vector machine formulation of linear PCA . . 10
3.1.3 Explicit centering in the SVM formulation 13
3.1.4 Implicit centering by means of bias term 16
3.2 SVM formulation of kernel PCA and weighted kernel PCA . . 20
3.2.1 Extension of linear PCA to kernel PCA using the SVM formulation 20
3.2.2 SVM formulation of weighted kernel PCA I. 30
3.2.3 SVM formulation of weighted kernel PCA II. 38
3.2.4 Remarks on the different formulations of the weighted kernel PCA 46
3.3 The relation of weighted kerenl PCA to the basic KSC problems 50
3.3.1 The symmetric normalized graph Laplacian kernel spectral clustering algorithm 50
3.3.2 The random walk normalized graph Laplacian kernel spectral clustering algorithm 52
3.4 Weighted kernel K-means and weighted kernel PCA 53
3.5 Kernel spectral clustering through weighted kernel PCA 61
3.5.1 Multiway kernel spectral clustering as weighted kernel PCA 62

4 Sparse kernel spectral clustering algorithms **71**
4.1 First section 72

5 Kernel spectral clustering on climate data **73**
5.1 First section 74

Chapter 1

Basic kernel spectral clustering algorithms

This schapter will be on the basic kernel spectral clustering story: what is the problem, ingredients, why and how does it work?, basic algorithms,Bla...bla...blaa.....

1.1 First section

Chapter 2

From classical to least-square support vector machines

This schapter will be on the classical SVM formulation of basic problems like classification and regression and the correspond LS-SVM versions, advantages and disadvantages of LS-SVM over classical SVM formulation, some solutions of getting back sparsity in case of LS-SVM will be mentiond, sequential minima optimization algorithms for classification and regression will be given in the appendix (both the derivations of the algorithms and the algorithms because the clear details are not available), bla...bla...blaa.....

2.1 First section

Chapter 3

Kernel spectral clustering as weighted kernel PCA

This schapter will contain the detailed formalism from classical liner pincipal component analysis (PCA) to support vectors machine formulated weigted kernel PCA. Bla...bla...blaa.....

3.1 Classical and SVM formulation of linear PCA

3.1.1 Classical linear principal component analysis formulation

Given a set of N d-dimensional *zero mean* real input data $\{\bar{x}_k\}_{k=1}^N \in \mathcal{D}$, $\bar{x} \in \mathbb{R}^d$ the goal in classical principal component analysis (PCA) is to find the direction $\bar{w}$ that maximize the variance of the projected variables $\bar{w}^T\bar{x}_k$ $k = 1, 2, ..., N$. By definition

$$Var(\bar{x}) = \mathbb{E}[(\bar{x} - \mathbb{E}[\bar{x}])^2] = \mathbb{E}[(\bar{x} - \hat{\bar{\mu}}_x)^2] = \mathbb{E}[(\bar{x} - \hat{\bar{\mu}}_x)(\bar{x} - \hat{\bar{\mu}}_x])^T] \tag{3.1}$$

$$Cov(\bar{x}, \bar{y}) = \mathbb{E}[(\bar{x} - \hat{\bar{\mu}}_x)(\bar{y} - \hat{\bar{\mu}}_y)^T] \tag{3.2}$$

$$Var(\bar{x}) = \mathbb{E}[(\bar{x} - \hat{\bar{\mu}}_x)^2] = \mathbb{E}[(\bar{x} - \hat{\bar{\mu}}_x)(\bar{x} - \hat{\bar{\mu}}_x)^T] = Cov(\bar{x}, \bar{x}) \tag{3.3}$$

$$C = Cov(\bar{x}, \bar{x}) = \frac{1}{N}\sum_{k=1}^{N}(\bar{x}_k - \hat{\bar{\mu}}_x)(\bar{x}_k - \hat{\bar{\mu}}_x)^T \tag{3.4}$$

where $C \in \mathbb{R}^{dxd}$ is the covariance matrix (or variance-covariance matrix) of the $\bar{x} \in \mathbb{R}^d$ vector variable with sample mean of $\hat{\bar{\mu}}_x = \frac{1}{N}\sum_{k=1}^{N}\bar{x}_k$. The fact, that the input data have *zero mean* reduces Eq 3.4 to

$$C = Cov(\bar{x}, \bar{x}) = \frac{1}{N}\sum_{k=1}^{N}\bar{x}_k\bar{x}_k^T. \tag{3.5}$$

The classical PCA problem can be written as the following optimization problem [1]

$$\begin{aligned}
\max_{\bar{w}} Var(\bar{w}^T\bar{x}) = Cov(\bar{w}^T\bar{x}, \bar{w}^T\bar{x}) &= \frac{1}{N}\sum_{k=1}^{N}(\bar{w}^T\bar{x}_k)^2 \\
&= \frac{1}{N}\sum_{k=1}^{N}(\bar{w}^T\bar{x}_k)(\bar{w}^T\bar{x}_k)^T = \frac{1}{N}\sum_{k=1}^{N}\bar{w}^T\bar{x}_k\bar{x}_k^T\bar{w} \\
&= \bar{w}^T\left(\frac{1}{N}\sum_{k=1}^{N}\bar{x}_k\bar{x}_k^T\right)\bar{w} \\
&= \bar{w}^T C\bar{w}
\end{aligned} \tag{3.6}$$

where Eq 3.5 was used in order to obtain the last line. This objective function can be optimized (w.r.t to $\bar{w}$) under the constraint that $\bar{w}^T\bar{w} = 1$ (in order to introduce a bound) which leads to the following Lagragian of the constrained optimization problem

$$\mathcal{L}(\bar{w};\lambda) = \bar{w}^T C\bar{w} - \lambda(\bar{w}^T\bar{w} - 1) \tag{3.7}$$

with Lagrange multiplier λ. The Karush-Kuhn-Tucker (KKT) optimality conditions are given by:

$$\begin{cases} \dfrac{\partial\mathcal{L}}{\partial\bar{w}} = 0 \quad \rightarrow \quad C\bar{w} = \lambda\bar{w} \\ \\ \dfrac{\partial\mathcal{L}}{\partial\lambda} = 0 \quad \rightarrow \quad \bar{w}^T\bar{w} = 1 \end{cases} \tag{3.8}$$

The first equation in the optimilaity conditions can be recognised as an eigenvalue equation with λ eigenvalues and $\bar{w}$ corresponding eigenvectors of the covariance matrix C.

$$C\bar{w} = \lambda\bar{w} \tag{3.9}$$

Thus the solution of the optimization problem is an eigenvector of the covariance matrix C. Since $C \in \mathbb{R}^{dxd}$ is a real symmetric p.s.d. matrix it has d real non-negative eigenvalues and the corresponding eigenvectors forms an orthonormal basis (normality comes from the constraint $\bar{w}^T\bar{w} = 1$). The quantity to be maximized in Eq. 3.6 can be written as

$$Var(\bar{w}^T\bar{x}) = \bar{w}^T C\bar{w} = \bar{w}^T\lambda\bar{w} = \lambda\bar{w}^T\bar{w} = \lambda \tag{3.10}$$

using the optimality conditions above. Therefore, the direction $\bar{w}$ that maximize the variance of the projected variables $\bar{w}^T\bar{x}_k \;\; k = 1, 2, ..., N$ is the eigenvector $\bar{w}^{(1)}$ of the covariance matrix C that corresponds to the largest eigenvalue λ_1. This leading eigenvector $\bar{w}^{(1)}$ is called to be the firts principal component of $\bar{x} \in \mathbb{R}^d \quad \{\bar{x}_k\}_{k=1}^N \in \mathcal{D}$ and since $Var(\bar{w}^T\bar{x}) = \lambda$ the variance of the projected data explained by this eigenvector is equal to λ_1. The second greatest variance of the projected data lies on the eigenvector $\bar{w}^{(2)}$ corresponds to the second largest eigenvalue λ_2 of the covariance matrix C and so on. As it was mentioned above, $C \in \mathbb{R}^{dxd}$ is a real symmetric p.s.d. matrix therefore it has d real non-negative eigenvalues and the corresponding d eigenvectors forms an orthonormal basis. Therefore, choosing the $t \leq d$ leading eigenvectors of the C covariance matrix leads to a new coordinate system (a new representation of the input data) in which the t projected variables are uncorreleted (because the eigenvectors are orthogonal) and as

much as possible variation (using t variables) presented in the input data set is retained [2]. Since $t \leq d$ PCA is used for dimensionality reduction that can be very efficient in case of higly correlated input variables because choosing $t \ll d$ can already retain most of the variation presented in the input data set. In order to achive this goal one can select the $t \leq d$ leading eigenvectors of the C covariance matrix and form the $W = [\bar{w}^{(1)}, ..., \bar{w}^{(t)}],\ t \leq d,\ W \in \mathbb{R}^{dxt}$ matrix. Then the t dimensional projection $Z \in \mathbb{R}^{txN}$ of the d dimensional input data set $\{\bar{x}_k\}_{k=1}^N \in \mathcal{D},\ \bar{x} \in \mathbb{R}^d$ that maximizes the retained variance can be obtained by calculating $Z = W^T X$ where X is the matrix containing the individual input data vectors as columns i.e. $X = [\bar{x}_1, ..., \bar{x}_N],\ X \in \mathbb{R}^{dxN}$.

Note that since the input data $\bar{x}$ are results of *real measurements* the d input variables can be associated with meaningfull variables while it is usually hard or more often impossible to link the t projected variables to real meaningfull quantities. Further interesting issues related to linear PCA like the interpretation of linear PCA as a reconstruction problem with information bottleneck [3] or applications are not addressed here because these are out of the scope of the present work.

3.1.2 Support vector machine formulation of linear PCA

The support vector formulation of linear PCA described in [1] [3] is adopted here. The linear PCA problem is formulated as a one class modelling problem with zero target value and the difference between the projected data points $z_k = \bar{w}^T \bar{x}_k$ and the 0 target value is considered as an error variable $e_k^2 = (0 - z_k)^2$. This error is maximized (w.r.t. to $\bar{w}$) for the given N input data which corresponds to the goal of the classical linear PCA i.e. looking for the direction $\bar{w}$ along the variance of the projected $\bar{w}^T \bar{x}$) input data is maximal. This maximization problem can be formulated as

$$\max_{\bar{w}} \sum_{k=1}^{N} [0 - \bar{w}^T \bar{x}_k]^2 \tag{3.11}$$

assuming again that the input data has *zero mean*.

This results in the following primal optimization problem

$$\boxed{\text{P:}} \quad \begin{aligned} \max_{\bar{w},e} J(\bar{w}, e) &= \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^2 - \frac{1}{2} \bar{w}^T \bar{w} \\ \textit{such that} \quad e_k &= \bar{w}^T \bar{x}_k, \quad k = 1, ..., N \end{aligned} \tag{3.12}$$

where the $-\frac{1}{2}\bar{w}^T\bar{w}$ regularization term is responsible for keeping the norm of $\bar{w}$ small during the maximization and $\gamma \in \mathbb{R}^+$ is a constant. The Lagragian of this constrained optimization problem can be written as

$$\mathcal{L}(\bar{w}, e; \alpha) = \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^2 - \frac{1}{2}\bar{w}^T\bar{w} - \sum_{k=1}^{N} \alpha_k (e_k - \bar{w}^T\bar{x}_k) \tag{3.13}$$

with Lagrange multipliers α_k $k = 1, ..., N$. The KKT conditions for optimality give

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \bar{w}} = -\bar{w} + \sum_{k=1}^{N} \alpha_k \bar{x}_k = 0 \;\; \rightarrow \;\; \bar{w} = \sum_{k=1}^{N} \alpha_k \bar{x}_k \\ \dfrac{\partial \mathcal{L}}{\partial e_k} = \gamma e_k - \alpha_k = 0 \qquad \rightarrow \;\; \alpha_k = \gamma e_k, \qquad k = 1, ..., N \\ \dfrac{\partial \mathcal{L}}{\partial \alpha_k} = -(e_k - \bar{w}^T\bar{x}_k) = 0 \;\; \rightarrow \;\; e_k - \bar{w}^T\bar{x}_k = 0, \;\; k = 1, ..., N \end{cases} \tag{3.14}$$

Eliminating the variables $\bar{w}$ and e_k by substituting the firts and second equations into the third in Eqs 3.14 yields the following set of linear equations for α_k

$$\frac{\alpha_k}{\gamma} - \sum_{l=1}^{N} \alpha_l \bar{x}_l^T \bar{x}_k = 0, \quad k = 1, ..., N \tag{3.15}$$

This system of equations can be written in matrix form by itroducing $\lambda = 1/\gamma$ as

$$\left| \begin{array}{l} \boxed{\text{D:}} \;\; \textit{solve in } \alpha \\ \\ \begin{bmatrix} \bar{x}_1^T\bar{x}_1 & \bar{x}_1^T\bar{x}_2 & \cdots & \bar{x}_1^T\bar{x}_N \\ \bar{x}_2^T\bar{x}_1 & \bar{x}_2^T\bar{x}_2 & \cdots & \bar{x}_2^T\bar{x}_N \\ \vdots & \vdots & \ddots & \vdots \\ \bar{x}_N^T\bar{x}_1 & \bar{x}_N^T\bar{x}_2 & \cdots & \bar{x}_N^T\bar{x}_N \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{bmatrix} = \lambda \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{bmatrix} \end{array} \right| \tag{3.16}$$

which is the dual interpretation of Eq 3.9 that can be written in the form of

$$\Omega\bar{\alpha} = \lambda\bar{\alpha} \tag{3.17}$$

by introducing

$$\Omega_{kl} = \bar{x}_k^T\bar{x}_l, \;\; k, l = 1, ..., N. \tag{3.18}$$

It can be easily recognised from Eq. 3.18 that $\Omega \in \mathbb{R}^{NxN}$ is the Gram matrix for linear kernel $K(\bar{x}_k, \bar{x}_l) = \bar{x}_k^T \bar{x}_k$. Therfore the optimal solution of the dual problem is an eigenvector of the Gram matrix. Since Ω is real symmetric pozitive definite matrix it has N real pozitive eigenvalues and the corresponding eigenvectors forms an orthonormal basis in $\mathbb{R}^N$ (assuming that the eigenvectors are normalized i.e. $\bar{\alpha}^T \bar{\alpha} = \sum_{k=1}^N \alpha_k^2 = 1$). The quantity to be maximized in Eq. 3.12 is proportional to

$$\sum_{k=1}^{N} (\bar{w}^T \bar{x}_k)^2 = \sum_{k=1}^{N} e_k^2 = \sum_{k=1}^{N} \frac{\alpha_k^2}{\gamma^2} = \frac{1}{\gamma^2} \sum_{k=1}^{N} \alpha_k^2 = \frac{1}{\gamma^2} = \lambda^2 \tag{3.19}$$

Therfore the optimal solution of the dual problem is the eigenvector $\bar{\alpha}^{(1)}$ of the Gram matrix corresponds to the largest eigenvalue λ_1. The optimal solution in the primal space can be written as

$$\bar{w}^{(1)} = \sum_{k=1}^{N} \alpha_k^{(1)} \bar{x}_k \tag{3.20}$$

by using the first KKT condition for optimality in Eqs. 3.14. The projected data points or score variables are given by

$$z_l^{(1)} = z^{(1)}(\bar{x}_l) = \bar{w}^{(1)^T} \bar{x}_l = \left(\sum_{k=1}^{N} \alpha_k^{(1)} \bar{x}_k \right)^T \bar{x}_l = \sum_{k=1}^{N} \alpha_k^{(1)} \bar{x}_k^T \bar{x}_l \tag{3.21}$$

As it was already mentioned above Ω is real symmetric pozitive definite matrix. Therefore it has N real pozitive eigenvalues λ and the corresponding eigenvectors $\bar{\alpha}$ forms an orthonormal basis in $\mathbb{R}^N$ (assuming again that $\bar{\alpha}^T \bar{\alpha} = 1$). Similarly to the classical linear PCA formalism one can select further eigenvectors of Ω and calculate the further corresponding projections according to 3.21. The derived new variables i.e. the different projections or score variables are uncorrelated due to the fact that the $\bar{\alpha}$ eigenvectors are orhogonal. Note, that since the problem is formulated now in the dual space the available dimensions are equal to the number of datapoints N that can be higher than the dimension of the input data space d. It is different from the classical linear PCA formulation discussed in the previous section which was formulated in the primal space thus the available dimensions were d.

Since the solutions in the primal and dual spaces are

$$\begin{aligned} &\boxed{\text{P:}} \quad z(\bar{x}) = \bar{w}^T \bar{x}, && \bar{w} \in \mathbb{R}^d \\ &\boxed{\text{D:}} \quad z(\bar{x}) = \textstyle\sum_{k=1}^{N} \alpha_k \bar{x}_k^T \bar{x}, && \bar{\alpha} \in \mathbb{R}^N \end{aligned} \tag{3.22}$$

solving the problem in the primal space is advantageous in case of large number of low dimansional data ($N \gg d$) while the dual problem is suitable for handling large dimensional input spaces with lower number of input data ($N \ll d$). Also note that the existance of this freedom of choice is a consequence of the *linearity* of the currently discussed PCA. As we will see later, the presented support vector formulation offers a very elegant way to extend the current *linear* to a *non-linear* PCA by changing (the current linear) to a non-linear kernel. In doing so, the input data are implicitly mapped from the input space to a higher, more often infinite dimensional space (feature space) by means of a non-linear kernel and linear PCA is done in the feature space. Since the dimension of the feature space is very often infinite (e.g. using Gaussian kernel) only the dual problem will be solvable in this case [4].

The current linear PCA formulation can be used for dimensionality reduction very similarly to the previously discussed classical formulation. One can select the $t \leq d$ leading eigenvectors of the Ω matrix and form the $A = [\bar{\alpha}^{(1)}, ..., \bar{\alpha}^{(t)}],\ t \leq d,\ A \in \mathbb{R}^{Nxt}$ matrix. Then the t dimensional projection $Z \in \mathbb{R}^{txN}$ of the d dimensional input data set $\{\bar{x}_k\}_{k=1}^N \in \mathcal{D},\ \bar{x} \in \mathbb{R}^d$ that maximizes the retained variance can be obtained by calculating $Z = A^T X^T X$ where X is the matrix containing the individual input data vectors as columns i.e. $X = [\bar{x}_1, ..., \bar{x}_N],\ X \in \mathbb{R}^{dxN}$. The t projected variables are uncorrelated due to the fact that the $\bar{\alpha}$ eigenvectors are orthogonal. Note that there are some differencies concerning the available number of dimensions here compare to the previously discussed classical linear PCA formulation originated from the differencies between the primal-dual solutions addressed above.

Note that since the dual solutions i.e. the $\bar{\alpha}$ eigenvectors are normalized ($\bar{\alpha}^T\bar{\alpha} = \sum_{k=1}^N \alpha_k^2 = 1$) the norm of the primal solutions are $\bar{w}^T\bar{w} = \lambda$ contrary to the constraint $\bar{w}^T\bar{w} = 1$ in the case of the classical linear PCA discussed in the previous section. This can be seen easly from

$$\begin{aligned} \bar{w} &= \sum_{k=1}^{N} \alpha_k \bar{x}_k = X\bar{\alpha} \\ \bar{w}^T\bar{w} &= [X\bar{\alpha}]^T[X\bar{\alpha}] = \bar{\alpha}^T X^T X\bar{\alpha} = \bar{\alpha}^T\Omega\bar{\alpha} = \lambda \end{aligned} \tag{3.23}$$

The last equation was obtained by left multiplying Eq. 3.17 and using the normality of the eigenvectors $\bar{\alpha}^T\bar{\alpha} = 1$.

3.1.3 Explicit centering in the SVM formulation

The input data was assumed to be centered i.e. *zero mean* both in section 3.1.1 and 3.1.2. This assumption will be dropped in this section therefore an

explicit centering of the input data will appear in the maximization problem defined by Eq. 3.11. As we will see, it results in an eigenvalue problem similar to 3.17 with the difference that the Gram matrix will be centered now.

The aforementioned explicit centering leads to projected data or score variables $z_k = \bar{w}^T(\bar{x}_k - \hat{\bar{\mu}}_x)$ and the maximization problem given by Eq. 3.11 is changed to

$$\max_{\bar{w}} \sum_{k=1}^{N} \left[0 - \bar{w}^T \left(\bar{x}_k - \hat{\bar{\mu}}_x\right)\right]^2 \tag{3.24}$$

The corresponding primal optimization problem can be formulated as

$$\left| \begin{array}{l} \boxed{\text{P:}} \\ \qquad \max_{\bar{w},e} J(\bar{w}, e) \quad = \quad \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^2 - \frac{1}{2} \bar{w}^T \bar{w} \\ \qquad \textit{such that} \quad e_k \quad = \quad \bar{w}^T(\bar{x}_k - \hat{\bar{\mu}}_x), \quad k = 1, ..., N \end{array} \right| \tag{3.25}$$

where the $-\frac{1}{2}\bar{w}^T\bar{w}$ regularization term is responsible for keeping the norm of $\bar{w}$ small during the maximization and $\gamma \in \mathbb{R}^+$ is a constant. The Lagragian of this constrained optimization problem can be written as

$$\mathcal{L}(\bar{w}, e; \alpha) = \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^2 - \frac{1}{2} \bar{w}^T \bar{w} - \sum_{k=1}^{N} \alpha_k [e_k - \bar{w}^T(\bar{x}_k - \hat{\bar{\mu}}_x)] \tag{3.26}$$

with Lagrange multipliers α_k $k = 1, ..., N$. The KKT conditions for optimality give

$$\left\{ \begin{array}{lll} \frac{\partial \mathcal{L}}{\partial \bar{w}} = -\bar{w} + \sum_{k=1}^{N} \alpha_k (\bar{x}_k - \hat{\bar{\mu}}_x) = 0 & \rightarrow & \bar{w} = \sum_{k=1}^{N} \alpha_k (\bar{x}_k - \hat{\bar{\mu}}_x) \\ \frac{\partial \mathcal{L}}{\partial e_k} = \gamma e_k - \alpha_k = 0 & \rightarrow & \alpha_k = \gamma e_k, \qquad k = 1, ..., N \\ \frac{\partial \mathcal{L}}{\partial \alpha_k} = -[e_k - \bar{w}^T(\bar{x}_k - \hat{\bar{\mu}}_x)] = 0 & \rightarrow & e_k - \bar{w}^T(\bar{x}_k - \hat{\bar{\mu}}_x) = 0, \quad k = 1, ..., N \end{array} \right. \tag{3.27}$$

Eliminating the variables $\bar{w}$ and e_k by substituting the firts and second equations into the third in Eqs 3.27 yields the following set of linear equations

for α_k

$$\frac{\alpha_k}{\gamma} - \sum_{l=1}^{N} \alpha_l [\bar{x}_l - \hat{\bar{\mu}}_x]^T [\bar{x}_k - \hat{\bar{\mu}}_x] = 0, \quad k = 1, ..., N \tag{3.28}$$

This system of equations can be written in matrix form by itroducing $\lambda = 1/\gamma$ as

D: *solve in* α

$$\begin{bmatrix} (\bar{x}_1 - \hat{\bar{\mu}}_x)^T(\bar{x}_1 - \hat{\bar{\mu}}_x) & \cdots & (\bar{x}_1 - \hat{\bar{\mu}}_x)^T(\bar{x}_N - \hat{\bar{\mu}}_x) \\ (\bar{x}_2 - \hat{\bar{\mu}}_x)^T(\bar{x}_1 - \hat{\bar{\mu}}_x) & \cdots & (\bar{x}_2 - \hat{\bar{\mu}}_x)^T(\bar{x}_N - \hat{\bar{\mu}}_x) \\ \vdots & \ddots & \vdots \\ (\bar{x}_N - \hat{\bar{\mu}}_x)^T(\bar{x}_1 - \hat{\bar{\mu}}_x) & \cdots & (\bar{x}_N - \hat{\bar{\mu}}_x)^T(\bar{x}_N - \hat{\bar{\mu}}_x) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{bmatrix} = \lambda \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{bmatrix} \tag{3.29}$$

This can be written as

$$\Omega_c \bar{\alpha} = \lambda \bar{\alpha} \tag{3.30}$$

where

$$\Omega_{c_{kl}} = (\bar{x}_k - \hat{\bar{\mu}}_x)^T(\bar{x}_l - \hat{\bar{\mu}}_x), \; k, l = 1, ..., N. \tag{3.31}$$

As it can be seen $\Omega_c = M_c \Omega M_c$ is the centered Gram matrix for a linear kernel where $M_c \in \mathbb{R}^{NxN}$ is the centering matrix

$$M_c = I_N - \frac{1}{N} \bar{1}_N \bar{1}_N^T \tag{3.32}$$

$I_N \in \mathbb{R}^{NxN}$ is the identity matrix and $\bar{1}_N$ is a column vector of N ones. Comparing 3.17 and 3.30 one can see that the explicit centering of the input data in the maximization problem given by Eq. 3.24 results in an eigenvalue problem similar to 3.17 with the difference that the Gram matrix will now be centered. The quantity to be maximized in Eq. 3.25 is proportional to

$$\sum_{k=1}^{N} [\bar{w}^T(\bar{x}_k - \hat{\bar{\mu}}_x)]^2 = \sum_{k=1}^{N} e_k^2 = \sum_{k=1}^{N} \frac{\alpha_k^2}{\gamma^2} = \frac{1}{\gamma^2} \sum_{k=1}^{N} \alpha_k^2 = \frac{1}{\gamma^2} = \lambda^2 \tag{3.33}$$

Therfore the optimal solution of the dual problem is the eigenvector $\bar{\alpha}^{(1)}$ of the centered Gram matrix Ω_c corresponds to the largest eigenvalue λ_1. One can select further eigenvectors $\bar{\alpha}$ of the centered Gram matrix Ω_c and calculate the further score variables in a similar way described in the previous section with similar properties. The solutions in the primal space $\bar{w}$ can be written by means of the selected dual solutions $\bar{\alpha}$ as

$$\bar{w} = \sum_{k=1}^{N} \alpha_k (\bar{x}_k - \hat{\bar{\mu}}_x) \tag{3.34}$$

by using the current fisrt KKT conditions for optimality in Eqs. 3.27. The (uncorrelated) projected data points or score variables can be given similarly to Eq. 3.21 using Eq. 3.34

$$\begin{aligned} z_l = z(\bar{x}_l) = \bar{w}^T(\bar{x}_l - \hat{\bar{\mu}}_x) &= \left(\sum_{k=1}^{N} \alpha_k(\bar{x}_k - \hat{\bar{\mu}}_x)\right)^T (\bar{x}_l - \hat{\bar{\mu}}_x) \\ &= \sum_{k=1}^{N} \alpha_k(\bar{x}_k - \hat{\bar{\mu}}_x)^T(\bar{x}_l - \hat{\bar{\mu}}_x) \end{aligned} \tag{3.35}$$

The solutions in the primal and dual spaces can be given as

$$\begin{aligned} &\boxed{\text{P:}} \quad z(\bar{x}) = \bar{w}^T(\bar{x} - \hat{\bar{\mu}}_x), && \bar{w} \in \mathbb{R}^d \\ &\boxed{\text{D:}} \quad z(\bar{x}) = \sum_{k=1}^{N} \alpha_k(\bar{x}_k - \hat{\bar{\mu}}_x)^T\bar{x}, && \bar{\alpha} \in \mathbb{R}^N \end{aligned} \tag{3.36}$$

Note that the sample mean explicitly appears both in the primal and the dual solutions.

3.1.4 Implicit centering by means of bias term

The support vector machine formulation of the PCA problem given in 3.1.2 offers a straightforward way to introduce a bias term in the function to be maximized. As we will see in this section, it implicitly leads to the centering of the Gram matrix instead of the previously discussed explicit centering of the input data given by Eq. 3.24.

Introducing a bias term b into the projection leads to projected data $z_k = \bar{w}^T\bar{x}_k + b$ and the maximization problem given by Eq. 3.11 is changed to

$$\max_{\bar{w}} \sum_{k=1}^{N} \left[0 - (\bar{w}^T\bar{x}_k + b)\right]^2 \tag{3.37}$$

The corresponding primal optimization problem can be formulated as

$$\left|\begin{array}{l} \boxed{\text{P:}} \\ \qquad \max_{\bar{w},e} J(\bar{w}, e) \;=\; \gamma\frac{1}{2}\sum_{k=1}^{N} e_k^2 - \frac{1}{2}\bar{w}^T\bar{w} \\ \qquad \textit{such that} \quad e_k \;=\; \bar{w}^T\bar{x}_k + b, \quad k = 1, ..., N \\ \end{array}\right| \tag{3.38}$$

where the $-\frac{1}{2}\bar{w}^T\bar{w}$ regularization term is responsible for keeping the norm of $\bar{w}$ small during the maximization and $\gamma \in \mathbb{R}^+$ is a constant. The Lagragian of this constrained optimization problem can be written as

$$\mathcal{L}(\bar{w}, e; \alpha) = \gamma\frac{1}{2}\sum_{k=1}^{N} e_k^2 - \frac{1}{2}\bar{w}^T\bar{w} - \sum_{k=1}^{N}\alpha_k(e_k - \bar{w}^T\bar{x}_k - b) \tag{3.39}$$

with Lagrange multipliers α_k $k = 1, ..., N$. The KKT conditions for optimality give

$$\begin{cases} \dfrac{\partial\mathcal{L}}{\partial\bar{w}} = -\bar{w} + \sum_{k=1}^{N}\alpha_k\bar{x}_k = 0 & \rightarrow \quad \bar{w} = \sum_{k=1}^{N}\alpha_k\bar{x}_k \\ \dfrac{\partial\mathcal{L}}{\partial e_k} = \gamma e_k - \alpha_k = 0 & \rightarrow \quad \alpha_k = \gamma e_k, \qquad k = 1, ..., N \\ \dfrac{\partial\mathcal{L}}{\partial b} = \sum_{k}^{N}\alpha_k = 0 & \rightarrow \quad \sum_{k=1}^{N}\alpha_k = 0 \\ \dfrac{\partial\mathcal{L}}{\partial\alpha_k} = -[e_k - \bar{w}^T\bar{x}_k - b] = 0 & \rightarrow \quad e_k - \bar{w}^T\bar{x}_k - b = 0, \quad k = 1, ..., N \end{cases} \tag{3.40}$$

Eliminating the variables $\bar{w}$ and e_k by substituting the firts and second equations into the fourth in Eqs 3.40 yields

$$\frac{\alpha_k}{\gamma} - \sum_{l=1}^{N}\alpha_l\bar{x}_l^T\bar{x}_k - b = 0, \quad k = 1, ..., N \tag{3.41}$$

Then applying the third KKT condition from Eqs. 3.40 (i.e. $\sum_{k=1}^{N}\alpha_k = 0$) to Eq. 3.41 gives the expression for the bias term b

$$\begin{aligned} &\sum_{k=1}^{N}\frac{\alpha_k}{\gamma} - \sum_{k=1}^{N}\sum_{l=1}^{N}\alpha_l\bar{x}_l^T\bar{x}_k - \sum_{k=1}^{N}b = 0 \\ &\frac{1}{\gamma}\sum_{k=1}^{N}\alpha_k - \sum_{k=1}^{N}\sum_{l=1}^{N}\alpha_l\bar{x}_l^T\bar{x}_k - Nb = 0 \\ &b = -\frac{1}{N}\sum_{k=1}^{N}\sum_{l=1}^{N}\alpha_l\bar{x}_l^T\bar{x}_k \end{aligned} \tag{3.42}$$

Substituting this back into Eq. 3.41 leads to the following set of expressions for $\bar{\alpha}$

$$\frac{\alpha_k}{\gamma} - \sum_{l=1}^{N} \alpha_l \bar{x}_l^T \bar{x}_k + \frac{1}{N} \sum_{k=1}^{N} \sum_{l=1}^{N} \alpha_l \bar{x}_l^T \bar{x}_k = 0, \quad k = 1, ..., N \tag{3.43}$$

The second and third terms on the left hand side can be written as

$$\begin{aligned} \sum_{l=1}^{N} \alpha_l \bar{x}_l^T \bar{x}_k - \frac{1}{N} \sum_{k=1}^{N} \sum_{l=1}^{N} \alpha_l \bar{x}_l^T \bar{x}_k &= \sum_{l=1}^{N} \alpha_l \bar{x}_l^T \bar{x}_k - \sum_{l=1}^{N} \alpha_l \bar{x}_l^T \frac{1}{N} \sum_{k=1}^{N} \bar{x}_k \\ &= \sum_{l=1}^{N} \alpha_l \bar{x}_l^T \bar{x}_k - \sum_{l=1}^{N} \alpha_l \bar{x}_l^T \hat{\bar{\mu}}_x = \sum_{l=1}^{N} \alpha_l \bar{x}_l^T [\bar{x}_k - \hat{\bar{\mu}}_x] \\ &= \sum_{l=1}^{N} \alpha_l [\bar{x}_l - \hat{\bar{\mu}}_x]^T [\bar{x}_k - \hat{\bar{\mu}}_x], \quad k = 1, ..., N \end{aligned} \tag{3.44}$$

where the last equation was obtained using $\sum_{k=1}^{N} \alpha_k = 0$ since

$$\begin{aligned} \sum_{l=1}^{N} \alpha_l [\bar{x}_l - \hat{\bar{\mu}}_x]^T [\bar{x}_k - \hat{\bar{\mu}}_x] &= \sum_{l=1}^{N} \alpha_l [\bar{x}_l^T \bar{x}_k - \bar{x}_l^T \hat{\bar{\mu}}_x - \hat{\bar{\mu}}_x^T \bar{x}_k + \hat{\bar{\mu}}_x^T \hat{\bar{\mu}}_x] \\ &= \sum_{l=1}^{N} \alpha_l \bar{x}_l^T \bar{x}_k - \sum_{l=1}^{N} \alpha_l \bar{x}_l^T \hat{\bar{\mu}}_x - \hat{\bar{\mu}}_x^T \bar{x}_k \underbrace{\sum_{l=1}^{N} \alpha_l}_{0} + \hat{\bar{\mu}}_x^T \hat{\bar{\mu}}_x \underbrace{\sum_{l=1}^{N} \alpha_l}_{0} \\ &= \sum_{l=1}^{N} \alpha_l \bar{x}_l^T \bar{x}_k - \sum_{l=1}^{N} \alpha_l \bar{x}_l^T \hat{\bar{\mu}}_x = \sum_{l=1}^{N} \alpha_l \bar{x}_l^T [\bar{x}_k - \hat{\bar{\mu}}_x] \end{aligned} \tag{3.45}$$

Therefore, Eq. 3.43 can be written as

$$\frac{\alpha_k}{\gamma} - \sum_{l=1}^{N} \alpha_l [\bar{x}_l - \hat{\bar{\mu}}_x]^T [\bar{x}_k - \hat{\bar{\mu}}_x] = 0, \quad k = 1, ..., N \tag{3.46}$$

using Eq. 3.44. This system of equations can be written in matrix form by itroducing $\lambda = 1/\gamma$ as

D: *solve in* α

$$\begin{bmatrix} (\bar{x}_1 - \hat{\bar{\mu}}_x)^T (\bar{x}_1 - \hat{\bar{\mu}}_x) & \cdots & (\bar{x}_1 - \hat{\bar{\mu}}_x)^T (\bar{x}_N - \hat{\bar{\mu}}_x) \\ (\bar{x}_2 - \hat{\bar{\mu}}_x)^T (\bar{x}_1 - \hat{\bar{\mu}}_x) & \cdots & (\bar{x}_2 - \hat{\bar{\mu}}_x)^T (\bar{x}_N - \hat{\bar{\mu}}_x) \\ \vdots & \ddots & \vdots \\ (\bar{x}_N - \hat{\bar{\mu}}_x)^T (\bar{x}_1 - \hat{\bar{\mu}}_x) & \cdots & (\bar{x}_N - \hat{\bar{\mu}}_x)^T (\bar{x}_N - \hat{\bar{\mu}}_x) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{bmatrix} = \lambda \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{bmatrix} \tag{3.47}$$

This can be written as

$$\Omega_c \bar{\alpha} = \lambda \bar{\alpha} \tag{3.48}$$

where

$$\Omega_{c_{kl}} = (\bar{x}_k - \hat{\bar{\mu}}_x)^T (\bar{x}_l - \hat{\bar{\mu}}_x),\ k, l = 1, ..., N. \tag{3.49}$$

As it can be seen $\Omega_c = M_c \Omega M_c$ is the centered Gram matrix for a linear kernel where $M_c \in \mathbb{R}^{NxN}$ is the centering matrix

$$M_c = I_N - \frac{1}{N} \bar{\mathbb{1}}_N \bar{\mathbb{1}}_N^T \tag{3.50}$$

$I_N \in \mathbb{R}^{NxN}$ is the identity matrix and $\bar{\mathbb{1}}_N$ is a column vector of N ones.

Note that Eqs 3.46- 3.49 here are identical to Eqs. 3.28- 3.31. obtained in the previous section with explicit centering of the input data in the problem formulation (see Eq. 3.24). Therefore, considering a bias term in the problem formulation (Eq. 3.37) automatically leads to the centering of the Garm matrix.

The quantity to be maximized in Eq. 3.38 is proportional to

$$\sum_{k=1}^{N} [\bar{w}^T \bar{x}_k + \hat{\bar{\mu}}_x]^2 = \sum_{k=1}^{N} e_k^2 = \sum_{k=1}^{N} \frac{\alpha_k^2}{\gamma^2} = \frac{1}{\gamma^2} \sum_{k=1}^{N} \alpha_k^2 = \frac{1}{\gamma^2} = \lambda^2 \tag{3.51}$$

Therfore the optimal solution of the dual problem is again the eigenvector $\bar{\alpha}^{(1)}$ of the centered Gram matrix Ω_c corresponds to the largest eigenvalue λ_1. The solution in the primal space can be written as

$$\bar{w} = \sum_{k=1}^{N} \alpha_k \bar{x}_k \tag{3.52}$$

using the current fisrt KKT conditions for optimality in Eq. 3.40. The (uncorrelated) projected data points or score variables are given by

$$\begin{aligned} z_l = z(\bar{x}_l) = \bar{w}^T \bar{x}_l + b &= \left(\sum_{k=1}^{N} \alpha_k \bar{x}_k) \right)^T \bar{x}_l + b \\ &= \sum_{k=1}^{N} \alpha_k \bar{x}_k^T \bar{x}_l + b \end{aligned} \tag{3.53}$$

Therefore the solutions in the primal and dual spaces are

$$\begin{aligned} &\boxed{\text{P:}} \qquad z(\bar{x}) = \bar{w}^T \bar{x} + b, \qquad\qquad \bar{w} \in \mathbb{R}^d \\ &\boxed{\text{D:}} \qquad z(\bar{x}) = \sum_{k=1}^{N} \alpha_k \bar{x}_k^T \bar{x} + b, \qquad \bar{\alpha} \in \mathbb{R}^N \\ &\textit{where} \quad b = -\frac{1}{N} \sum_{k=1}^{N} \sum_{l=1}^{N} \alpha_l \bar{x}_l^T \bar{x}_k \end{aligned} \tag{3.54}$$

Note that contrary to Eqs. 3.36, the sample mean does not appear here.

3.2 SVM formulation of kernel PCA and weighted kernel PCA

3.2.1 Extension of linear PCA to kernel PCA using the SVM formulation

Support vector mechine formulation of the *linear* pricipal component analysis was discussed so far that resulted in an eigenproblem of the (centered) Gram matrix in the dual space. As it was already mentioned in 3.1.2, the presented support vector machine formulation offers a very elegant and straigthforward way to extend the *linear* PCA to its kernel version that gives the possibility of *non-linear* PCA. In the kernel PCA, as in case of other kernel methods, the input data $\{\bar{x}_k\}_{k=1}^N \in \mathcal{D}$, $\bar{x} \in \mathbb{R}^d$ are mapped $\varphi(\cdot) : \mathbb{R}^d \to \mathbb{R}^{n_h}$ from the d dimensional *input space* $\mathbb{R}^d$ to a high (n_h) dimensional space $\mathbb{R}^{n_h}$ called *feature space* and linear PCA is done in this feature space insted of the input space. As we will see, only dot products of the mapped input data vectors $\varphi(\bar{x})$ appear in the problem when it is formulated in the dual space. Therefore, the explicit calculation of the mapping to the high or infinite dimensional feature space is not necessary since the *kernel trick* can be applied in case of using Mercer kernels. Thus, the *kernel trick* makes possible to work in the high or infinite dimensional *feature space* without doing explicit computations in this space. The kernel version of the PCA was first given by *Schölkopf et al* [5] and the support vector machine formulation of the kernel PCA discussed here was given by *Suykens* [1]. Support vector machine formulation of kernel PCA both using the explicit and implicit centering through a bias term will be given.

Formulation with explicit centerin

The problem can be formulated similarly to Eq. 3.24 with the differnce that the input data are mapped into the feature space by using $\varphi(\cdot) : \mathbb{R}^d \to \mathbb{R}^{n_h}$ thus the score variables are

$$z_k = z(\bar{x}_k) = \bar{w}^T(\varphi(\bar{x}_k) - \hat{\mu}_{\varphi(\bar{x})}) \tag{3.55}$$

where $\hat{\mu}_{\varphi(\bar{x})} = \frac{1}{N}\sum_{k=1}^{N}\varphi(\bar{x}_k)$. Note, that $\varphi(\bar{x}) \in \mathbb{R}^{n_h}$ where n_h is high and can be even infinite. Thus the maximization problem can now be written as

$$\max_{\bar{w}} \sum_{k=1}^{N}[0 - \bar{w}^T(\varphi(\bar{x}_k) - \hat{\mu}_{\varphi(\bar{x})})]^2 \tag{3.56}$$

The corresponding primal optimization problem can be formulated as

$$\left|\begin{array}{l} \boxed{\text{P:}} \\ \quad \max_{\bar{w},e} J(\bar{w}, e) \;=\; \gamma\frac{1}{2}\sum_{k=1}^{N} e_k^2 - \frac{1}{2}\bar{w}^T\bar{w} \\ \quad \textit{such that} \quad e_k \;=\; \bar{w}^T(\varphi(\bar{x}_k) - \hat{\mu}_{\varphi(\bar{x})}), \quad k = 1, ..., N \end{array}\right| \tag{3.57}$$

The Lagragian of this constrained optimization problem is

$$\mathcal{L}(\bar{w}, e; \alpha) = \gamma\frac{1}{2}\sum_{k=1}^{N} e_k^2 - \frac{1}{2}\bar{w}^T\bar{w} - \sum_{k=1}^{N}\alpha_k[e_k - \bar{w}^T(\varphi(\bar{x}_k) - \hat{\mu}_{\varphi(\bar{x})})] \tag{3.58}$$

with Lagrange multipliers α_k $k = 1, ..., N$. The KKT conditions for optimality give

$$\left\{\begin{array}{lll} \dfrac{\partial\mathcal{L}}{\partial\bar{w}} = -\bar{w} + \sum_{k=1}^{N}\alpha_k(\varphi(\bar{x}_k) - \hat{\mu}_{\varphi(\bar{x})}) = 0 & \to & \bar{w} = \sum_{k=1}^{N}\alpha_k(\varphi(\bar{x}_k) - \hat{\mu}_{\varphi(\bar{x})}) \\ \dfrac{\partial\mathcal{L}}{\partial e_k} = \gamma e_k - \alpha_k = 0 & \to & \alpha_k = \gamma e_k,\ k = 1, ..., N \\ \dfrac{\partial\mathcal{L}}{\partial\alpha_k} = -[e_k - \bar{w}^T(\varphi(\bar{x}_k) - \hat{\mu}_{\varphi(\bar{x})})] = 0 & \to & e_k - \bar{w}^T(\varphi(\bar{x}_k) - \hat{\mu}_{\varphi(\bar{x})}) = 0, \\ & & k = 1, ..., N \end{array}\right. \tag{3.59}$$

Eliminating the variables $\bar{w}$ and e_k by substituting the firts and second equations into the third in Eqs 3.59 yields the following set of equations for α_k

$$\frac{\alpha_k}{\gamma} - \sum_{l=1}^{N} \alpha_l [\varphi(\bar{x}_l) - \hat{\mu}_{\varphi(\bar{x})}]^T [\varphi(\bar{x}_k) - \hat{\mu}_{\varphi(\bar{x})}] = 0, \quad k = 1, ..., N \tag{3.60}$$

This system of equations can be written in matrix form by itroducing $\lambda = 1/\gamma$ as

$$\boxed{\text{D:}}\ solve\ in\ \alpha \quad \Omega_c \bar{\alpha} = \lambda \bar{\alpha} \tag{3.61}$$

where

$$\Omega_{c_{kl}} = (\varphi(\bar{x}_k) - \hat{\mu}_{\varphi(\bar{x})})^T (\varphi(\bar{x}_l) - \hat{\mu}_{\varphi(\bar{x})}), \quad k, l = 1, ..., N.$$

$$\Omega_c = \begin{bmatrix} (\varphi(\bar{x}_1) - \hat{\mu}_{\varphi(\bar{x})})^T (\varphi(\bar{x}_1) - \hat{\mu}_{\varphi(\bar{x})}) & \cdots & (\varphi(\bar{x}_1) - \hat{\mu}_{\varphi(\bar{x})})^T (\varphi(\bar{x}_N) - \hat{\mu}_{\varphi(\bar{x})}) \\ (\varphi(\bar{x}_2) - \hat{\mu}_{\varphi(\bar{x})})^T (\varphi(\bar{x}_1) - \hat{\mu}_{\varphi(\bar{x})}) & \cdots & (\varphi(\bar{x}_2) - \hat{\mu}_{\varphi(\bar{x})})^T (\varphi(\bar{x}_N) - \hat{\mu}_{\varphi(\bar{x})}) \\ \vdots & \ddots & \vdots \\ (\varphi(\bar{x}_N) - \hat{\mu}_{\varphi(\bar{x})})^T (\varphi(\bar{x}_1) - \hat{\mu}_{\varphi(\bar{x})}) & \cdots & (\varphi(\bar{x}_N) - \hat{\mu}_{\varphi(\bar{x})})^T (\varphi(\bar{x}_N) - \hat{\mu}_{\varphi(\bar{x})}) \end{bmatrix} \tag{3.62}$$

$\Omega \in \mathbb{R}^{NxN}$, $\Omega_{kl} = \varphi(\bar{x}_k)^T \varphi(\bar{x}_l) = K(\bar{x}_k, \bar{x}_l)$ (using the kernel trick) is the kernel matrix and $\Omega_c = M_c \Omega M_c$ is the centered kernel matrix where $M_c \in \mathbb{R}^{NxN}$ is the centering matrix

$$M_c = I_N - \frac{1}{N} \bar{1}_N \bar{1}_N^T \tag{3.63}$$

$I_N \in \mathbb{R}^{NxN}$ is the identity matrix and $\bar{1}_N$ is a column vector of N ones. The elements of the centered kernel matrix can be given

$$\begin{aligned}
\Omega_{c_{kl}} =&[\varphi(\bar{x}_k) - \hat{\mu}_{\varphi(\bar{x})}]^T[\varphi(\bar{x}_l) - \hat{\mu}_{\varphi(\bar{x})}] \\
=&[\varphi(\bar{x}_k) - \frac{1}{N}\sum_{i=1}^{N}\varphi(\bar{x}_i)]^T[\varphi(\bar{x}_l) - \frac{1}{N}\sum_{i=1}^{N}\varphi(\bar{x}_i)] \\
=&\varphi(\bar{x}_k)^T\varphi(\bar{x}_l) - \varphi(\bar{x}_k)^T\frac{1}{N}\sum_{i=1}^{N}\varphi(\bar{x}_i) \\
&- \varphi(\bar{x}_l)^T\frac{1}{N}\sum_{i=1}^{N}\varphi(\bar{x}_i) + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\varphi(\bar{x}_i)^T\varphi(\bar{x}_j) \\
=&K(\bar{x}_k,\bar{x}_l) - \frac{1}{N}\sum_{i=1}^{N}K(\bar{x}_k,\bar{x}_i) - \frac{1}{N}\sum_{i=1}^{N}K(\bar{x}_l,\bar{x}_i) \\
&+ \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}K(\bar{x}_i,\bar{x}_j)
\end{aligned} \tag{3.64}$$

by using the kernel trick $K(\bar{x}_k,\bar{x}_l) = \varphi(\bar{x}_k)^T\varphi(\bar{x}_l)$ and assuming that the kernel satisfy the Mercer conditions. This result is given in this form by *Suykens* in [1] and this solution is equivalent with kernel PCA as proposed by *Schölkopf et al.* in [5]. According to Eq. 3.61 the pool of candidate solutions is formed by the $\bar{\alpha}$ eigenvectors of the centered kernel matrix Ω_c. Since $\Omega_c \in \mathbb{R}^{NxN}$ is a real symmetric p.s.d. matrix it has N non-negative eigenvalues and the corresponding eigenvectors are orthogonal. If the normality of the eigenvectors is further assumed i.e. $\bar{\alpha}^T\bar{\alpha} = \sum_{k=1}^{N}\alpha_k^2 = 1$ the quantity to be maximized in Eq 3.56 is

$$\sum_{k=1}^{N}[0-\bar{w}^T(\varphi(\bar{x}_k)-\hat{\mu}_{\varphi(\bar{x})})]^2 = \sum_{k=1}^{N}e_k^2 = \sum_{k=1}^{N}\frac{\alpha_k^2}{\gamma^2} = \frac{1}{\gamma^2}\sum_{k=1}^{N}\alpha_k^2 = \frac{1}{\gamma^2} = \lambda^2 \tag{3.65}$$

Therefore, the optimal solution of the dual problem is the $\bar{\alpha}^{(1)}$ eigenvector of the centered kernel matrix Ω_c corresponds to the largest eigenvalue λ_1 as

earlier. The score variables can be given by

$$\begin{aligned} z(\bar{x}) &= \bar{w}^T(\varphi(\bar{x}) - \hat{\mu}_{\varphi(\bar{x})}) = \\ &= \left(\sum_{k=1}^{N} \alpha_k(\varphi(\bar{x}_k) - \hat{\mu}_{\varphi(\bar{x})})\right)^T (\varphi(\bar{x}) - \hat{\mu}_{\varphi(\bar{x})}) \\ &= \sum_{k=1}^{N} \alpha_k[\varphi(\bar{x}_k) - \hat{\mu}_{\varphi(\bar{x})}]^T[\varphi(\bar{x}) - \hat{\mu}_{\varphi(\bar{x})}] \\ &= \sum_{k=1}^{N} \alpha_k \left\{ K(\bar{x}_k, \bar{x}) - \frac{1}{N}\sum_{i=1}^{N} K(\bar{x}_k, \bar{x}_i) - \frac{1}{N}\sum_{i=1}^{N} K(\bar{x}, \bar{x}_i) \right. \\ &\qquad \left. + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} K(\bar{x}_i, \bar{x}_j) \right\} \end{aligned} \tag{3.66}$$

using Eq 3.55, first equation from Eqs 3.59 and 3.64. This result is given in this form in [1].

However, the kernel PCA eigenvalue problem given by Eq 3.61 as well as the expression for the score variables given above by Eq. 3.66 can be further simplified after realising that the optimal solutions of the dual problem i.e. $\bar{\alpha}$ vectors are going to be centered i.e. $\sum_{k=1}^{N} \alpha_k = 0$. This condition is explicitly not among the KKT conditions for optimality given by Eqs 3.59 but it is there implicitly since

$$\begin{aligned} \sum_{k=1}^{N} \alpha_k &= \sum_{k=1}^{N} \gamma e_k = \gamma \sum_{k=1}^{N} \bar{w}^T(\varphi(\bar{x}_k) - \hat{\mu}_{\varphi(\bar{x})}) \\ &= \gamma \bar{w}^T \sum_{k=1}^{N} (\varphi(\bar{x}_k) - \hat{\mu}_{\varphi(\bar{x})}) = \gamma \bar{w}^T \underbrace{\left[\sum_{k=1}^{N} \varphi(\bar{x}_k) - \sum_{i=1}^{N} \varphi(\bar{x}_i)\right]}_{\bar{0}} \end{aligned} \tag{3.67}$$

Then Eqs. 3.60 can be written as

$$\begin{aligned}
\frac{\alpha_k}{\gamma} &= \sum_{l=1}^{N} \alpha_l [\varphi(\bar{x}_l) - \hat{\mu}_{\varphi(\bar{x})}]^T [\varphi(\bar{x}_k) - \hat{\mu}_{\varphi(\bar{x})}] = \\
&= \sum_{l=1}^{N} \alpha_l [\varphi(\bar{x}_l)^T \varphi(\bar{x}_k) - \varphi(\bar{x}_l)^T \hat{\mu}_{\varphi(\bar{x})} - \hat{\mu}_{\varphi(\bar{x})}^T \varphi(\bar{x}_k) + \hat{\mu}_{\varphi(\bar{x})}^T \hat{\mu}_{\varphi(\bar{x})}] \\
&= \sum_{l=1}^{N} \alpha_l \varphi(\bar{x}_l)^T \varphi(\bar{x}_k) - \sum_{l=1}^{N} \alpha_l \varphi(\bar{x}_l)^T \hat{\mu}_{\varphi(\bar{x})} \\
&\quad - \hat{\mu}_{\varphi(\bar{x})}^T \varphi(\bar{x}_k) \underbrace{\sum_{l=1}^{N} \alpha_l}_{0} + \hat{\mu}_{\varphi(\bar{x})}^T \hat{\mu}_{\varphi(\bar{x})} \underbrace{\sum_{l=1}^{N} \alpha_l}_{0} \\
&= \sum_{l=1}^{N} \alpha_l \varphi(\bar{x}_l)^T \varphi(\bar{x}_k) - \sum_{l=1}^{N} \alpha_l \varphi(\bar{x}_l)^T \hat{\mu}_{\varphi(\bar{x})} \\
&= \sum_{l=1}^{N} \alpha_l \varphi(\bar{x}_l)^T [\varphi(\bar{x}_k) - \hat{\mu}_{\varphi(\bar{x})}], \quad k = 1, ..., N
\end{aligned} \tag{3.68}$$

by using the previously derived condition for $\bar{\alpha}$ that $\sum_{k=1}^{N} \alpha_k = 0$. Therefore, the kernel PCA eigenproblem given by Eq 3.61 can be written in the following equivalent but simpler form

$$\boxed{\text{D:}}\ solve\ in\ \alpha \quad \Omega_c^* \bar{\alpha} = \lambda \bar{\alpha} \tag{3.69}$$

where

$$\begin{aligned}
\Omega_c^* &= M_c \Omega \\
\Omega_{c_{kl}}^* &= [\varphi(\bar{x}_k) - \hat{\mu}_{\varphi(\bar{x})}]^T \varphi(\bar{x}_l), \qquad k, l = 1, ..., N \\
\Omega_{c_{kl}}^* &= [\varphi(\bar{x}_k) - \hat{\mu}_{\varphi(\bar{x})}]^T \varphi(\bar{x}_l) \\
&= [\varphi(\bar{x}_k) - \frac{1}{N} \sum_{i=1}^{N} \varphi(\bar{x}_i)]^T \varphi(\bar{x}_l) \\
&= \varphi(\bar{x}_k)^T \varphi(\bar{x}_l) - \frac{1}{N} \sum_{i=1}^{N} \varphi(\bar{x}_i)^T \varphi(\bar{x}_l) \\
&= K(\bar{x}_k, \bar{x}_l) - \frac{1}{N} \sum_{i=1}^{N} K(\bar{x}_i, \bar{x}_l) \qquad k, l = 1, ..., N
\end{aligned} \tag{3.70}$$

by using the kernel trick in case of Ω_c^*. M_c is the centering matrix defined by 3.63. The expression for the score variables given by Eq. 3.66 also simplifies to

$$\begin{aligned} z(\bar{x}) &= \bar{w}^T(\varphi(\bar{x}) - \hat{\mu}_{\varphi(\bar{x})}) = \\ &= \left(\sum_{k=1}^{N} \alpha_k(\varphi(\bar{x}_k) - \hat{\mu}_{\varphi(\bar{x})})\right)^T (\varphi(\bar{x}) - \hat{\mu}_{\varphi(\bar{x})}) \\ &= \sum_{k=1}^{N} \alpha_k[\varphi(\bar{x}_k) - \hat{\mu}_{\varphi(\bar{x})}]^T[\varphi(\bar{x}) - \hat{\mu}_{\varphi(\bar{x})}] \\ &= \sum_{l=1}^{N} \alpha_l\varphi(\bar{x}_l)^T[\varphi(\bar{x}_k) - \hat{\mu}_{\varphi(\bar{x})}] \\ &= \sum_{k=1}^{N} \alpha_k \left\{K(\bar{x}_k, \bar{x}) - \frac{1}{N}\sum_{i=1}^{N} K(\bar{x}_k, \bar{x}_i)\right\} \end{aligned} \tag{3.71}$$

by using Eq 3.55, 3.68 and Eq 3.70. It is easy to see that the last two terms in Eq. 3.66 give zero contributions after the summation since $\sum_{k=1}^{N} \alpha_k = 0$ as it was shown by Eq 3.67.

Therefore, the two different forms of the kernel PCA eigenvalue problem given by Eq. 3.61 and 3.69 give identical solutions i.e.

$$\Omega_c\bar{\alpha} = \Omega_c^*\bar{\alpha} = \lambda\bar{\alpha} \tag{3.72}$$

but the expression of the kernel matrix as well as the score variables are simpler in case of using Ω_c^* instead of Ω_c. It is the consequence of making use of the fact that the optimal solutions of the dual problems i.e. the $\bar{\alpha}$ eigenvectors are centered i.e. $\sum_{k=1}^{N} \alpha_k = 0$.

As it was already mentioned in 3.1.2, the number of candidate solutions N (i.e. the number of $\bar{\alpha}$ eigenvectors of the centered kernel matrix Ω_c) can be larger than the dimension of the input space d that is a consequence of solving the dual problem. Therefore, one can select as few score variables ($\bar{\alpha}$ eigenvectors) as possible by minimizing the reconstruction error i.e.

$$\min \sum_{k=1}^{N} \|\bar{x}_k - \tilde{\bar{x}}_k\|_2^2 = \min \sum_{k=1}^{N} \|\bar{x}_k - h(z_k)\|_2^2 \tag{3.73}$$

where $\tilde{\bar{x}} = h(z)$ are the reconstructed input variables obtained from the score variables z by the non-linear mapping $h(z)$[6].

Since the $\bar{\alpha}$ eigenvectors are orthogonal the score variables corresponds to different $\bar{\alpha}$ eigenvectors of the centered kernel matrix Ω_c are decorreleted from each other.

Note, that since the mapped input data $\varphi(\bar{x}) \in \mathbb{R}^{n_h}$ the primal solutions $\bar{w} = \sum_{k=1}^{N} \alpha_k(\varphi(\bar{x}_k) - \hat{\mu}_{\varphi(\bar{x})})$ also $\in \mathbb{R}^{n_h}$ where n_h is the dimension of the feature space that is very often infinite (e.g. using Gaussian kernel). Therefore, solving the primal problem is not feasible in this case. Contrary to the primal problem, the mapped input variables $\varphi(\bar{x})$ appear only in form of dot products in the dual formulation that makes possible the application of the *kernel trick* $K(\bar{x}_k, \bar{x}_l) = \varphi(\bar{x}_k)^T\varphi(\bar{x}_l)$ (using Mercer kernel). Therefore, the explicit mapping of the input data from the d dimensional input space to the high or infinite dimensional feature space $\varphi(\cdot) : \mathbb{R}^d \to \mathbb{R}^{n_h}$ is not necessary when the problem is formulated in the dual space thanks to the *kernel trick*.Thus, the *kernel trick* makes possible to work even in infinite dimensional *feature space* without doing explicit computations in this space.

Automatic centering with bias

The kernel PCA problem including a bias term can be formulated similarly to the linear PCA case discussed in 3.1.4 with difference that the input data are mapped into the feature space by using $\varphi(\cdot) : \mathbb{R}^d \to \mathbb{R}^{n_h}$ Thus the score variables are

$$z_k = z(\bar{x}_k) = \bar{w}^T\varphi(\bar{x}_k) + b \tag{3.74}$$

The maximization problem can be written as

$$\max_{\bar{w}} \sum_{k=1}^{N} [0 - (\bar{w}^T\varphi(\bar{x}_k) + b)]^2 \tag{3.75}$$

The corresponding primal optimization problem can be formulated as

$$\left| \begin{array}{l} \boxed{\text{P:}} \\ \qquad \max_{\bar{w},e} J(\bar{w}, e) \;=\; \gamma\frac{1}{2}\sum_{k=1}^{N} e_k^2 - \frac{1}{2}\bar{w}^T\bar{w} \\ \\ \qquad \textit{such that} \quad e_k \;=\; \bar{w}^T\varphi(\bar{x}_k) + b, \quad k = 1, ..., N \end{array} \right| \tag{3.76}$$

The Lagragian of this constrained optimization problem is

$$\mathcal{L}(\bar{w}, e; \alpha) = \gamma\frac{1}{2}\sum_{k=1}^{N} e_k^2 - \frac{1}{2}\bar{w}^T\bar{w} - \sum_{k=1}^{N} \alpha_k(e_k - \bar{w}^T\varphi(\bar{x}_k) - b) \tag{3.77}$$

with Lagrange multipliers α_k $k = 1, ..., N$. The KKT conditions for optimality give

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \bar{w}} = -\bar{w} + \sum_{k=1}^{N} \alpha_k \varphi(\bar{x}_k) = 0 & \rightarrow \quad \bar{w} = \sum_{k=1}^{N} \alpha_k \varphi(\bar{x}_k) \\ \dfrac{\partial \mathcal{L}}{\partial e_k} = \gamma e_k - \alpha_k = 0 & \rightarrow \quad \alpha_k = \gamma e_k, \qquad k = 1, ..., N \\ \dfrac{\partial \mathcal{L}}{\partial b} = \sum_{k}^{N} \alpha_k = 0 & \rightarrow \quad \sum_{k=1}^{N} \alpha_k = 0 \\ \dfrac{\partial \mathcal{L}}{\partial \alpha_k} = -[e_k - \bar{w}^T \varphi(\bar{x}_k) - b] = 0 & \rightarrow \quad e_k - \bar{w}^T \varphi(\bar{x}_k) - b = 0, \quad k = 1, ..., N \end{cases} \tag{3.78}$$

Eliminating the variables $\bar{w}$ and e_k by substituting the firts and second equations into the fourth in Eqs 3.78 yields

$$\frac{\alpha_k}{\gamma} - \sum_{l=1}^{N} \alpha_l \varphi(\bar{x}_l)^T \varphi(\bar{x}_k) - b = 0, \quad k = 1, ..., N \tag{3.79}$$

Then applying the third KKT condition from Eqs. 3.78 (i.e. $\sum_{k=1}^{N} \alpha_k = 0$) to Eq. 3.79 gives the expression for the bias term b

$$\begin{aligned} &\sum_{k=1}^{N} \frac{\alpha_k}{\gamma} - \sum_{k=1}^{N} \sum_{l=1}^{N} \alpha_l \varphi(\bar{x}_l)^T \varphi(\bar{x}_k) - \sum_{k=1}^{N} b = 0 \\ &\frac{1}{\gamma} \sum_{k=1}^{N} \alpha_k - \sum_{k=1}^{N} \sum_{l=1}^{N} \alpha_l \varphi(\bar{x}_l)^T \varphi(\bar{x}_k) - Nb = 0 \\ &b = -\frac{1}{N} \sum_{k=1}^{N} \sum_{l=1}^{N} \alpha_l \varphi(\bar{x}_l)^T \varphi(\bar{x}_k) \end{aligned} \tag{3.80}$$

Substituting this back into Eq. 3.79 leads to the following set of expressions for $\bar{\alpha}$

$$\frac{\alpha_k}{\gamma} - \sum_{l=1}^{N} \alpha_l \varphi(\bar{x}_l)^T \varphi(\bar{x}_k) + \frac{1}{N} \sum_{k=1}^{N} \sum_{l=1}^{N} \alpha_l \varphi(\bar{x}_l)^T \varphi(\bar{x}_k) = 0, \quad k = 1, ..., N \tag{3.81}$$

The second and third terms on the left hand side can be written as

$$\begin{aligned}\sum_{l=1}^{N}\alpha_l\varphi(\bar{x}_l)^T\varphi(\bar{x}_k) - \frac{1}{N}\sum_{k=1}^{N}\sum_{l=1}^{N}\alpha_l\varphi(\bar{x}_l)^T\varphi(\bar{x}_k) = \\ &= \sum_{l=1}^{N}\alpha_l\varphi(\bar{x}_l)^T\varphi(\bar{x}_k) - \sum_{l=1}^{N}\alpha_l\varphi(\bar{x}_l)^T\frac{1}{N}\sum_{k=1}^{N}\varphi(\bar{x}_k) \\ &= \sum_{l=1}^{N}\alpha_l\varphi(\bar{x}_l)^T\varphi(\bar{x}_k) - \sum_{l=1}^{N}\alpha_l\varphi(\bar{x}_l)^T\hat{\mu}_{\varphi(\bar{x})} \\ &= \sum_{l=1}^{N}\alpha_l\varphi(\bar{x}_l)^T[\varphi(\bar{x}_k) - \hat{\mu}_{\varphi(\bar{x})}]\end{aligned} \tag{3.82}$$

Therefore, Eq. 3.81 can be written as

$$\frac{\alpha_k}{\gamma} - \sum_{l=1}^{N}\alpha_l\varphi(\bar{x}_l)^T[\varphi(\bar{x}_k) - \hat{\mu}_{\varphi(\bar{x})}] = 0, \quad k = 1, ..., N \tag{3.83}$$

by using Eq. 3.82. Note that this equation is identical to Eq. 3.68 obtained previously by using the explicit centering of the feature map instead of the bias term. Therefore, all the equations Eqs. 3.69- 3.71 also hold now with the same remarks. Note, that using a bias term instead of the exlicit centering of the feature map directly leads to the simpler form of the solution due to the fact that the centering of the optimal solutions of the dual problem i.e. $\sum_{k=1}^{N}\alpha_k = 0$ explicitly appears among the KKT conditions given by Eq 3.78 contrary to Eq. 3.59. The identity given by Eq. 3.68 is also true thus the equivalence given by Eq. 3.72

$$\Omega_c\bar{\alpha} = \Omega_c^*\bar{\alpha} = \lambda\bar{\alpha} \tag{3.84}$$

also holds.

The quantity to be maximized in Eq. 3.75 is also equal to the eigenvalue of the centered kernel matrix Ω_c or the simpler matrix $Omega_c^*$

$$\max_{\bar{w}}\sum_{k=1}^{N}[0 - (\bar{w}^T\varphi(\bar{x}_k) + b)]^2 = \sum_{k=1}^{N}e_k^2 = \sum_{k=1}^{N}\frac{\alpha_k^2}{\gamma^2} = \frac{1}{\gamma^2}\sum_{k=1}^{N}\alpha_k^2 = \frac{1}{\gamma^2} = \lambda^2 \tag{3.85}$$

thus similar remarks, concerning the optimal solution of the dual problem, hold now as given previously in the explicit centering case.

Also note, that if the feature map is already centered i.e. $\hat{\mu}_{\varphi(\bar{x})} = \frac{1}{N}\sum_{k=1}^{N}\varphi(\bar{x}_k) = \bar{0}$ then $\Omega_c = \Omega_c^* = \Omega$ and Eq. 3.72 becomes

$$\Omega_c\bar{\alpha} = \Omega_c^*\bar{\alpha} = \Omega\bar{\alpha} = \lambda\bar{\alpha} \tag{3.86}$$

and the expression for the score variables given by Eq. 3.71 further simplifies to

$$z(\bar{x}) = \sum_{k=1}^{N} \alpha_k K(\bar{x}_k, \bar{x}) \tag{3.87}$$

since the last term in Eq. 3.71 that involvs $\hat{\mu}_{\varphi(\bar{x})}$ becomes zero.

3.2.2 SVM formulation of weighted kernel PCA I.

I just put my derivation and I will write the detailed text later.

Let the diagonal weight matrix $V \in \mathbb{R}^{NxN}$ with diagonal elements of $v_i > 0,\ i = 1, ..., N$.

Formulation with explicit centering

Let the $\sqrt{v}$ *weighted* score variables are

$$z'_k = z'(\bar{x}_k) = \sqrt{v_k}\bar{w}^T(\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}) \tag{3.88}$$

where $\hat{\mu}^v_{\varphi(\bar{x})}$ is the weighted mean of the feature map

$$\hat{\mu}^v_{\varphi(\bar{x})} = \frac{\sum_{k=1}^{N} v_k \varphi(\bar{x}_k)}{\sum_{k=1}^{N} v_k} \tag{3.89}$$

The maximization problem can now be written as

$$\max_{\bar{w}} \sum_{k=1}^{N} [0 - \sqrt{v_k}\bar{w}^T(\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})})]^2 \tag{3.90}$$

The corresponding primal optimization problem can be formulated as

$$\left| \begin{array}{l} \boxed{\text{P:}} \\ \quad \max_{\bar{w},e^*} J(\bar{w}, \bar{e}^*) \;=\; \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^{*2} - \frac{1}{2}\bar{w}^T\bar{w} \\ \quad \textit{such that} \quad e_k^* \;=\; \sqrt{v_k}\bar{w}^T(\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}), \quad k = 1, ..., N \end{array} \right| \tag{3.91}$$

The Lagragian of this constrained optimization problem is

$$\mathcal{L}(\bar{w}, \bar{e}^*; \bar{\alpha}) = \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^{*2} - \frac{1}{2}\bar{w}^T\bar{w} - \sum_{k=1}^{N} \alpha_k [e_k^* - \sqrt{v_k}\bar{w}^T(\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})})] \tag{3.92}$$

with Lagrange multipliers α_k $k = 1, ..., N$. The KKT conditions for optimality give

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \bar{w}} = 0 \quad \rightarrow \quad \bar{w} = \displaystyle\sum_{k=1}^{N} \alpha_k \sqrt{v_k}(\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}) \\ \dfrac{\partial \mathcal{L}}{\partial e^*_k} = 0 \quad \rightarrow \quad \alpha_k = \gamma e^*_k, \qquad\qquad k = 1, ..., N \\ \dfrac{\partial \mathcal{L}}{\partial \alpha_k} = 0 \quad \rightarrow \quad e^*_k - \sqrt{v_k}\bar{w}^T(\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}) = 0, \quad k = 1, ..., N \end{cases} \tag{3.93}$$

Eliminating the variables $\bar{w}$ and e^*_k by substituting the firts and second equations into the third in Eqs 3.93 yields the following set of equations for α_k

$$\frac{\alpha_k}{\gamma} - \sqrt{v_k}\sum_{l=1}^{N} \alpha_l \sqrt{v_l}[\varphi(\bar{x}_l) - \hat{\mu}^v_{\varphi(\bar{x})}]^T[\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}] = 0, \quad k = 1, ..., N \tag{3.94}$$

This system of equations can be written in matrix form by itroducing $\lambda = 1/\gamma$ as

$$\boxed{\text{D:}}\ solve\ in\ \alpha \quad V^{\frac{1}{2}}\Omega_v V^{\frac{1}{2}}\bar{\alpha} = \lambda\bar{\alpha} \tag{3.95}$$

where

$$\Omega_{v_{kl}} = (\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})})^T(\varphi(\bar{x}_l) - \hat{\mu}^v_{\varphi(\bar{x})}), \quad k, l = 1, ..., N.$$

$$\Omega_v = \begin{bmatrix} (\varphi(\bar{x}_1) - \hat{\mu}^v_{\varphi(\bar{x})})^T(\varphi(\bar{x}_1) - \hat{\mu}^v_{\varphi(\bar{x})}) & \cdots & (\varphi(\bar{x}_1) - \hat{\mu}^v_{\varphi(\bar{x})})^T(\varphi(\bar{x}_N) - \hat{\mu}^v_{\varphi(\bar{x})}) \\ (\varphi(\bar{x}_2) - \hat{\mu}^v_{\varphi(\bar{x})})^T(\varphi(\bar{x}_1) - \hat{\mu}^v_{\varphi(\bar{x})}) & \cdots & (\varphi(\bar{x}_2) - \hat{\mu}^v_{\varphi(\bar{x})})^T(\varphi(\bar{x}_N) - \hat{\mu}^v_{\varphi(\bar{x})}) \\ \vdots & \ddots & \vdots \\ (\varphi(\bar{x}_N) - \hat{\mu}^v_{\varphi(\bar{x})})^T(\varphi(\bar{x}_1) - \hat{\mu}^v_{\varphi(\bar{x})}) & \cdots & (\varphi(\bar{x}_N) - \hat{\mu}^v_{\varphi(\bar{x})})^T(\varphi(\bar{x}_N) - \hat{\mu}^v_{\varphi(\bar{x})}) \end{bmatrix} \tag{3.96}$$

$\Omega \in \mathbb{R}^{NxN}$, $\Omega_{kl} = \varphi(\bar{x}_k)^T\varphi(\bar{x}_l) = K(\bar{x}_k, \bar{x}_l)$ (using the kernel trick) is the kernel matrix and $\Omega_v = M_v\Omega M_v^T$ is the weighted centered kernel matrix where $M_v \in \mathbb{R}^{NxN}$ is the weighted centering matrix

$$M_v = I_N - \frac{1}{\sum_{i=1}^{N} v_i}\bar{1}_N\bar{1}_N^T V \tag{3.97}$$

$I_N \in \mathbb{R}^{NxN}$ is the identity matrix and $\bar{\mathbb{1}}_N$ is a column vector of N ones. The elements of the weighted centered kernel matrix can be given

$$\begin{aligned}
\Omega_{v_{kl}} =&[\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}]^T[\varphi(\bar{x}_l) - \hat{\mu}^v_{\varphi(\bar{x})}] \\
=&[\varphi(\bar{x}_k) - \frac{1}{\sum_{i=1}^N v_i}\sum_{i=1}^N v_i\varphi(\bar{x}_i)]^T[\varphi(\bar{x}_l) - \frac{1}{\sum_{i=1}^N v_i}\sum_{i=1}^N v_i\varphi(\bar{x}_i)] \\
=&\varphi(\bar{x}_k)^T\varphi(\bar{x}_l) - \varphi(\bar{x}_k)^T\frac{1}{\sum_{i=1}^N v_i}\sum_{i=1}^N v_i\varphi(\bar{x}_i) \\
&- \varphi(\bar{x}_l)^T\frac{1}{\sum_{i=1}^N v_i}\sum_{i=1}^N v_i\varphi(\bar{x}_i) + \frac{1}{\left[\sum_{i=1}^N v_i\right]^2}\sum_{i=1}^N v_i\sum_{j=1}^N v_j\varphi(\bar{x}_i)^T\varphi(\bar{x}_j) \\
=&K(\bar{x}_k,\bar{x}_l) - \frac{1}{\sum_{i=1}^N v_i}\sum_{i=1}^N v_iK(\bar{x}_k,\bar{x}_i) - \frac{1}{\sum_{i=1}^N v_i}\sum_{i=1}^N v_iK(\bar{x}_l,\bar{x}_i) \\
&+ \frac{1}{\left[\sum_{i=1}^N v_i\right]^2}\sum_{i=1}^N v_i\sum_{j=1}^N v_jK(\bar{x}_i,\bar{x}_j)
\end{aligned} \tag{3.98}$$

by using the kernel trick $K(\bar{x}_k,\bar{x}_l) = \varphi(\bar{x}_k)^T\varphi(\bar{x}_l)$ and assuming that the kernel satisfy the Mercer conditions. According to Eq. 3.95 the pool of candidate solutions is formed by the $\bar{\alpha}$ eigenvectors of the $V^{\frac{1}{2}}\Omega_v V^{\frac{1}{2}} = V^{\frac{1}{2}}M_v\Omega M_v^T V^{\frac{1}{2}}$ matrix. If the normality of the vectors $\bar{\alpha}$ is further assumed i.e. $\bar{\alpha}^T\bar{\alpha} = \sum_{k=1}^N \alpha_k^2 = 1$ the quantity to be maximized in Eq 3.90 is

$$\sum_{k=1}^N[0 - \sqrt{v_k}\bar{w}^T(\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})})]^2 = \sum_{k=1}^N e_k^{*2} = \sum_{k=1}^N\frac{\alpha_k^2}{\gamma^2} = \frac{1}{\gamma^2}\sum_{k=1}^N\alpha_k^2 = \frac{1}{\gamma^2} = \lambda^2 \tag{3.99}$$

Therefore, the optimal solution of the dual problem is the $\bar{\alpha}^{(1)}$ eigenvector of the $V^{\frac{1}{2}}\Omega_v V^{\frac{1}{2}} = V^{\frac{1}{2}}M_v\Omega M_v^T V^{\frac{1}{2}}$ matrix corresponds to the largest eigenvalue

λ_1. The score variables can be given by

$$\begin{aligned}
z(\bar{x}_k) &= \bar{w}^T(\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}) = \\
&= \left(\sum_{l=1}^{N} \alpha_l \sqrt{v_l}(\varphi(\bar{x}_l) - \hat{\mu}^v_{\varphi(\bar{x})})\right)^T (\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}) \\
&= \sum_{l=1}^{N} \alpha_l \sqrt{v_l}[\varphi(\bar{x}_l) - \hat{\mu}^v_{\varphi(\bar{x})}]^T[\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}] \\
&= \sum_{l=1}^{N} \alpha_l \sqrt{v_l} \left\{ K(\bar{x}_l, \bar{x}_k) - \frac{1}{\sum_{i=1}^N v_i} \sum_{i=1}^{N} v_i K(\bar{x}_l, \bar{x}_i) \right. \\
&\left. - \frac{1}{\sum_{i=1}^N v_i} \sum_{i=1}^{N} v_i K(\bar{x}_k, \bar{x}_i) + \frac{1}{\left[\sum_{i=1}^N v_i\right]^2} \sum_{i=1}^{N} v_i \sum_{j=1}^{N} v_j K(\bar{x}_i, \bar{x}_j) \right\}
\end{aligned} \tag{3.100}$$

using Eq 3.88, first equation from Eqs 3.93 and 3.98.

The weighted kernel PCA eigenvalue problem given by Eq 3.95 as well as the expression for the score variables given above by Eq. 3.100 can be further simplified after realising that the optimal solutions of the dual problem i.e. $\bar{\alpha}$ vectors are going to be $\sqrt{v}$ weight centered i.e. $\sum_{k=1}^N \alpha_k \sqrt{v_k} = 0$. This condition is explicitly not among the KKT conditions for optimality given by Eqs 3.93 but it is there implicitly since

$$\begin{aligned}
\sum_{k=1}^{N} \alpha_k \sqrt{v_k} &= \gamma \sum_{k=1}^{N} e^*_k \sqrt{v_k} = \gamma \sum_{k=1}^{N} \sqrt{v_k} \bar{w}^T(\varphi(\bar{x_k}) - \hat{\mu}^v_{\varphi(\bar{x})})\sqrt{v_k} = \\
&= \gamma \bar{w}^T \left\{ \sum_{k=1}^{N} v_k \varphi(\bar{x_k}) - \frac{\sum_{k=1}^N v_k \sum_{i=1}^N v_i \varphi(\bar{x}_i)}{\sum_{i=1}^N v_i} \right\} \\
&= \gamma \bar{w}^T \sum_{k=1}^{N} v_k \varphi(\bar{x_k}) \underbrace{\left\{ 1 - \frac{\sum_{i=1}^N v_i}{\sum_{i=1}^N v_i} \right\}}_{0} = 0
\end{aligned} \tag{3.101}$$

Then the summation in Eqs. 3.94 can be written as

$$\begin{aligned}
&\sum_{l=1}^{N} \alpha_l \sqrt{v_l}[\varphi(\bar{x}_l) - \hat{\mu}^v_{\varphi(\bar{x})}]^T[\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}] = \\
&= \sum_{l=1}^{N} \alpha_l \sqrt{v_l}[\varphi(\bar{x}_l)^T\varphi(\bar{x}_k) - \varphi(\bar{x}_l)^T\hat{\mu}^v_{\varphi(\bar{x})} - \hat{\mu}^{vT}_{\varphi(\bar{x})}\varphi(\bar{x}_k) + \hat{\mu}^{vT}_{\varphi(\bar{x})}\hat{\mu}^v_{\varphi(\bar{x})}] \\
&= \sum_{l=1}^{N} \alpha_l \sqrt{v_l}\varphi(\bar{x}_l)^T\varphi(\bar{x}_k) - \sum_{l=1}^{N} \alpha_l \sqrt{v_l}\varphi(\bar{x}_l)^T\hat{\mu}^v_{\varphi(\bar{x})} \\
&\quad - \hat{\mu}^{vT}_{\varphi(\bar{x})}\varphi(\bar{x}_k) \underbrace{\sum_{l=1}^{N} \alpha_l \sqrt{v_l}}_{0} + \hat{\mu}^{vT}_{\varphi(\bar{x})}\hat{\mu}^v_{\varphi(\bar{x})} \underbrace{\sum_{l=1}^{N} \alpha_l \sqrt{v_l}}_{0} \\
&= \sum_{l=1}^{N} \alpha_l \sqrt{v_l}\varphi(\bar{x}_l)^T\varphi(\bar{x}_k) - \sum_{l=1}^{N} \alpha_l \sqrt{v_l}\varphi(\bar{x}_l)^T\hat{\mu}^v_{\varphi(\bar{x})} \\
&= \sum_{l=1}^{N} \alpha_l \sqrt{v_l}\varphi(\bar{x}_l)^T[\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}]
\end{aligned} \tag{3.102}$$

by using the previously derived condition for $\bar{\alpha}$ that $\sum_{k=1}^{N} \alpha_k \sqrt{v_k} = 0$. Therefore, the weighted kernel PCA eigenproblem given by Eq 3.95 can be written in the following equivalent but simpler form

$$\boxed{\text{D:}}\ \textit{solve in } \alpha \quad V^{\frac{1}{2}}\Omega^*_v V^{\frac{1}{2}}\bar{\alpha} = \lambda\bar{\alpha} \tag{3.103}$$

where

$$\begin{aligned}
\Omega^*_v &= M_v\Omega \\
\Omega^*_{v_{kl}} &= [\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}]^T\varphi(\bar{x}_l), \qquad k,l = 1,...,N \\
\Omega^*_{v_{kl}} &= [\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}]^T\varphi(\bar{x}_l) \\
&= [\varphi(\bar{x}_k) - \frac{1}{\sum_{i=1}^{N} v_i}\sum_{i=1}^{N} v_i\varphi(\bar{x}_i)]^T\varphi(\bar{x}_l) \\
&= \varphi(\bar{x}_k)^T\varphi(\bar{x}_l) - \frac{1}{\sum_{i=1}^{N} v_i}\sum_{i=1}^{N} v_i\varphi(\bar{x}_i)^T\varphi(\bar{x}_l) \\
&= K(\bar{x}_k,\bar{x}_l) - \frac{1}{\sum_{i=1}^{N} v_i}\sum_{i=1}^{N} v_iK(\bar{x}_i,\bar{x}_l) \qquad k,l = 1,...,N
\end{aligned} \tag{3.104}$$

by using the kernel trick in case of Ω_v^*. M_v is the weighted centering matrix defined by 3.97. The expression for the score variables given by Eq. 3.100 also simplifies to

$$\begin{aligned} z(\bar{x}_k) &= \bar{w}^T(\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}) = \\ &= \left(\sum_{l=1}^{N} \alpha_l \sqrt{v_l}(\varphi(\bar{x}_l) - \hat{\mu}^v_{\varphi(\bar{x})})\right)^T (\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}) \\ &= \sum_{l=1}^{N} \alpha_l \sqrt{v_l}[\varphi(\bar{x}_l) - \hat{\mu}^v_{\varphi(\bar{x})}]^T[\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}] \\ &= \sum_{l=1}^{N} \alpha_l \sqrt{v_l}\varphi(\bar{x}_l)^T[\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}] \\ &= \sum_{l=1}^{N} \alpha_l \sqrt{v_l}\left\{K(\bar{x}_l, \bar{x}_k) - \frac{1}{\sum_{i=1}^{N} v_i}\sum_{i=1}^{N} v_i K(\bar{x}_l, \bar{x}_i)\right\} \end{aligned} \tag{3.105}$$

by using Eq 3.88, 3.102 and Eq 3.104. It is easy to see that the last two terms in Eq. 3.100 give zero contributions after the summation since $\sum_{k=1}^{N} \alpha_k \sqrt{v_k} = 0$ as it was shown by Eq 3.102.

Therefore, the two different forms of the weighted kernel PCA eigenvalue problem given by Eq. 3.95 and 3.103 give identical solutions i.e.

$$V^{\frac{1}{2}}\Omega_v V^{\frac{1}{2}}\bar{\alpha} = V^{\frac{1}{2}}\Omega_v^* V^{\frac{1}{2}}\bar{\alpha} = \lambda\bar{\alpha} \tag{3.106}$$

but the expression of the kernel matrix as well as the score variables are simpler in case of using Ω_v^* instead of Ω_v. It is the consequence of making use of the fact that the optimal solutions of the dual problems i.e. the $\bar{\alpha}$ eigenvectors are $\sqrt{v}$ centered i.e. $\sum_{k=1}^{N} \alpha_k \sqrt{v_k} = 0$.

Automatic centering with bias

The $\sqrt{v}$ *weighted* score variables are

$$z'_k = z'(\bar{x}_k) = \sqrt{v_k}(\bar{w}^T\varphi(\bar{x}_k) + b) \tag{3.107}$$

The maximization problem can be written as

$$\max_{\bar{w}} \sum_{k=1}^{N} [0 - \sqrt{v_k}(\bar{w}^T\varphi(\bar{x}_k) + b)]^2 \tag{3.108}$$

The corresponding primal optimization problem can be formulated as

$$\left| \begin{array}{l} \boxed{\text{P:}} \\ \quad \max_{\bar{w},\bar{e}^*} J(\bar{w},\bar{e}^*) = \gamma \frac{1}{2}\sum_{k=1}^{N} e_k^{*2} - \frac{1}{2}\bar{w}^T\bar{w} \\ \quad \textit{such that} \quad e_k^* = \sqrt{v_k}(\bar{w}^T\varphi(\bar{x}_k)+b), \quad k=1,...,N \end{array} \right| \tag{3.109}$$

The Lagragian of this constrained optimization problem is

$$\mathcal{L}(\bar{w},\bar{e}^*;\bar{\alpha}) = \gamma\frac{1}{2}\sum_{k=1}^{N} e_k^{*2} - \frac{1}{2}\bar{w}^T\bar{w} - \sum_{k=1}^{N}\alpha_k(e_k^* - \sqrt{v_k}(\bar{w}^T\varphi(\bar{x}_k)+b)) \tag{3.110}$$

with Lagrange multipliers α_k $k=1,...,N$. The KKT conditions for optimality give

$$\left\{ \begin{array}{lll} \frac{\partial\mathcal{L}}{\partial\bar{w}} = 0 & \rightarrow & \bar{w} = \sum_{k=1}^{N}\alpha_k\sqrt{v_k}\varphi(\bar{x}_k) \\ \frac{\partial\mathcal{L}}{\partial e_k^*} = 0 & \rightarrow & \alpha_k = \gamma e_k^*, \qquad\qquad k=1,...,N \\ \frac{\partial\mathcal{L}}{\partial b} = 0 & \rightarrow & \sum_{k=1}^{N}\alpha_k\sqrt{v_k} = 0 \\ \frac{\partial\mathcal{L}}{\partial\alpha_k} = 0 & \rightarrow & e_k^* - \sqrt{v_k}(\bar{w}^T\varphi(\bar{x}_k)+b) = 0, \quad k=1,...,N \end{array} \right. \tag{3.111}$$

Eliminating the variables $\bar{w}$ and e_k^* by substituting the firts and second equations into the fourth in Eqs 3.111 yields

$$\frac{\alpha_k}{\gamma} - \sqrt{v_k}\left\{\sum_{l=1}^{N}\alpha_l\sqrt{v_l}\varphi(\bar{x}_l)^T\varphi(\bar{x}_k)+b\right\} = 0, \quad k=1,...,N \tag{3.112}$$

Then applying the third KKT condition from Eqs. 3.111 (i.e. $\sum_{k=1}^{N} \alpha_k \sqrt{v_k} = 0$) to Eq. 3.112 gives the expression for the bias term b

$$\begin{aligned}
&\sum_{k=1}^{N} \frac{\alpha_k \sqrt{v_k}}{\gamma} - \sum_{k=1}^{N} v_k \sum_{l=1}^{N} \alpha_l \sqrt{v_l} \varphi(\bar{x}_l)^T \varphi(\bar{x}_k) - \sum_{k=1}^{N} v_k b = 0 \\
&\frac{1}{\gamma} \sum_{k=1}^{N} \alpha_k \sqrt{v_k} - \sum_{k=1}^{N} v_k \sum_{l=1}^{N} \alpha_l \sqrt{v_l} \varphi(\bar{x}_l)^T \varphi(\bar{x}_k) - b \sum_{k=1}^{N} v_k = 0 \\
&b = -\frac{\sum_{k=1}^{N} v_k \sum_{l=1}^{N} \alpha_l \sqrt{v_l} \varphi(\bar{x}_l)^T \varphi(\bar{x}_k)}{\sum_{k=1}^{N} v_k}
\end{aligned} \tag{3.113}$$

Substituting this back into Eq. 3.112 leads to the following set of expressions for $\bar{\alpha} = [\alpha_1, ..., \alpha_N]$

$$\frac{\alpha_k}{\gamma} - \sqrt{v_k} \left\{ \sum_{l=1}^{N} \alpha_l \sqrt{v_l} \varphi(\bar{x}_l)^T \varphi(\bar{x}_k) - \frac{\sum_{i=1}^{N} v_i \sum_{l=1}^{N} \alpha_l \sqrt{v_l} \varphi(\bar{x}_l)^T \varphi(\bar{x}_i)}{\sum_{i=1}^{N} v_i} \right\} = 0, \tag{3.114}$$

The expression in the curly brackets can be written as

$$\begin{aligned}
&\sum_{l=1}^{N} \alpha_l \sqrt{v_l} \varphi(\bar{x}_l)^T \varphi(\bar{x}_k) - \frac{\sum_{i=1}^{N} v_i \sum_{l=1}^{N} \alpha_l \sqrt{v_l} \varphi(\bar{x}_l)^T \varphi(\bar{x}_i)}{\sum_{i=1}^{N} v_i} = \\
&= \sum_{l=1}^{N} \alpha_l \sqrt{v_l} \varphi(\bar{x}_l)^T \varphi(\bar{x}_k) - \sum_{l=1}^{N} \alpha_l \sqrt{v_l} \varphi(\bar{x}_l)^T \frac{\sum_{i=1}^{N} v_i \varphi(\bar{x}_i)}{\sum_{i=1}^{N} v_i} \\
&= \sum_{l=1}^{N} \alpha_l \sqrt{v_l} \varphi(\bar{x}_l)^T \varphi(\bar{x}_k) - \sum_{l=1}^{N} \alpha_l \sqrt{v_l} \varphi(\bar{x}_l)^T \hat{\mu}^v_{\varphi(\bar{x})} \\
&= \sum_{l=1}^{N} \alpha_l \sqrt{v_l} \varphi(\bar{x}_l)^T [\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}]
\end{aligned} \tag{3.115}$$

Therefore, Eq. 3.114 can be written as

$$\frac{\alpha_k}{\gamma} - \sqrt{v_k} \sum_{l=1}^{N} \alpha_l \sqrt{v_l} \varphi(\bar{x}_l)^T [\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}] = 0, \quad k = 1, ..., N \tag{3.116}$$

by using Eq. 3.115. Note that these equations are identical to those given by Eq. 3.103 obtained previously by using the explicit centering approach intead of the bias term. Therefore, all the equations Eqs. 3.103- 3.104 also hold now with the same remarks. Note, that using a bias term instead of the exlicit centering of the feature map directly leads to the simpler form of

the solution due to the fact that the $\sqrt{v}$ centering of the optimal solutions of the dual problem i.e. $\sum_{k=1}^{N} \alpha_k \sqrt{v_k} = 0$ explicitly appears among the KKT conditions given by Eq 3.111 contrary to Eq. 3.93. The identity given by Eq. 3.102 is also true thus the equivalence given by Eq. 3.106

$$V^{\frac{1}{2}} \Omega_v V^{\frac{1}{2}} \bar{\alpha} = V^{\frac{1}{2}} \Omega_v^* V^{\frac{1}{2}} \bar{\alpha} = \lambda \bar{\alpha} \tag{3.117}$$

also holds.

The quantity to be maximized in Eq. 3.108 is also equal to the eigenvalue of the $V^{\frac{1}{2}} \Omega_v V^{\frac{1}{2}}$ or the simpler $V^{\frac{1}{2}} \Omega_v^* V^{\frac{1}{2}}$ matrix (assuming that the eigenvectors $\bar{\alpha}$ are normalized i.e. $[\bar{\alpha}^T \bar{\alpha} = \sum_{k=1}^{N} \alpha_k^2 = 1)$

$$\max_{\bar{w}} \sum_{k=1}^{N} [0 - \sqrt{v_k}(\bar{w}^T \varphi(\bar{x}_k) + b)]^2 = \sum_{k=1}^{N} e_k^{*2} = \sum_{k=1}^{N} \frac{\alpha_k^2}{\gamma^2} = \frac{1}{\gamma^2} \sum_{k=1}^{N} \alpha_k^2 = \frac{1}{\gamma^2} = \lambda^2 \tag{3.118}$$

thus similar remarks, concerning the optimal solution of the dual problem, hold now as given previously in the explicit centering case.

Also note, that if the feature map is already weight centered i.e. $\hat{\mu}^v_{\varphi(\bar{x})} = \frac{\sum_{k=1}^{N} v_k \varphi(\bar{x}_k)}{\sum_{k=1}^{N} v_k} = \bar{0}$ then $\Omega_v = \Omega_v^* = \Omega$ and Eq. 3.106 becomes

$$V^{\frac{1}{2}} \Omega_v V^{\frac{1}{2}} \bar{\alpha} = V^{\frac{1}{2}} \Omega_v^* V^{\frac{1}{2}} \bar{\alpha} = V^{\frac{1}{2}} \Omega V^{\frac{1}{2}} \bar{\alpha} = \lambda \bar{\alpha} \tag{3.119}$$

and the expression for the score variables given by Eq. 3.105 further simplifies to

$$z(\bar{x}_k) = \sum_{l=1}^{N} \alpha_l \sqrt{v_l} K(\bar{x}_l, \bar{x}_k) \tag{3.120}$$

since the last term in Eq. 3.105 that involvs $\hat{\mu}^v_{\varphi(\bar{x})}$ becomes zero.

3.2.3 SVM formulation of weighted kernel PCA II.

A slightly different derivation of the weighted kernel PCA will be shown here compared to the the previous section. The score variables were *weighted* there by the square root of the corresponding weight factors (see Eqs. 3.88 and 3.107) and the differences between the zero target value and these weighted projections were maximized as an error function e_k^* (see Eqs. 3.90 and 3.108). Here the differences between the *unweighted* projections and the zero target value will be taken as an error function and this weighted error function will be maximized. As a result of this, the constraints on these errors in the formulation of the primal optimization problem will be

given in a slightly different way that will have some effects on the Lagrange multipliers.

Let the diagonal weight matrix again $V \in \mathbb{R}^{NxN}$ with diagonal elements of $v_i > 0,\ i = 1, ..., N$.

Formulation with explicit centerin

Let the score variables are

$$z_k = z(\bar{x}_k) = \bar{w}^T(\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}) \tag{3.121}$$

where $\hat{\mu}^v_{\varphi(\bar{x})}$ is the weighted mean of the feature map

$$\hat{\mu}^v_{\varphi(\bar{x})} = \frac{\sum_{k=1}^N v_k \varphi(\bar{x}_k)}{\sum_{k=1}^N v_k} \tag{3.122}$$

The maximization problem can now be written as

$$\max_{\bar{w}} \sum_{k=1}^N v_k [0 - \bar{w}^T(\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})})]^2 \tag{3.123}$$

The corresponding primal optimization problem can be formulated as

$$\left| \begin{array}{l} \boxed{\text{P:}} \\ \qquad \max_{\bar{w},\bar{e}} J(\bar{w},\bar{e}) \quad = \quad \gamma \frac{1}{2} \sum_{k=1}^N v_k e_k^2 - \frac{1}{2} \bar{w}^T \bar{w} \\ \qquad \textit{such that} \quad e_k \quad = \quad \bar{w}^T(\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}), \quad k = 1, ..., N \end{array} \right| \tag{3.124}$$

The Lagragian of this constrained optimization problem is

$$\mathcal{L}(\bar{w}, \bar{e}; \bar{\alpha}) = \gamma \frac{1}{2} \sum_{k=1}^N v_k e_k^2 - \frac{1}{2} \bar{w}^T \bar{w} - \sum_{k=1}^N \beta_k [e_k - \bar{w}^T(\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})})] \tag{3.125}$$

with Lagrange multipliers β_k $k = 1, ..., N$. Note, that Eq. 3.123 is equivalent with 3.90 and Eqs. 3.124 can be obtained from Eqs. 3.91 by substituting $e_k^* = \sqrt{v_k} e_k$ that results in the Lagragian given above. The KKT conditions

for optimality give

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \bar{w}} = 0 \quad \rightarrow \quad \bar{w} = \displaystyle\sum_{k=1}^{N} \beta_k (\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}) \\ \dfrac{\partial \mathcal{L}}{\partial e_k} = 0 \quad \rightarrow \quad \beta_k = \gamma v_k e_k, \qquad\qquad k = 1, ..., N \\ \dfrac{\partial \mathcal{L}}{\partial \beta_k} = 0 \quad \rightarrow \quad e_k - \bar{w}^T(\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}) = 0, \quad k = 1, ..., N \end{cases} \tag{3.126}$$

Eliminating the variables $\bar{w}$ and e_k by substituting the firts and second equations into the third in Eqs 3.126 yields the following set of equations for β_k

$$\frac{\beta_k}{\gamma} - v_k \sum_{l=1}^{N} \beta_l [\varphi(\bar{x}_l) - \hat{\mu}^v_{\varphi(\bar{x})}]^T [\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}] = 0, \quad k = 1, ..., N \tag{3.127}$$

This system of equations can be written in matrix form by itroducing $\lambda = 1/\gamma$ as

$$\boxed{\text{D:}}\ solve\ in\ \beta \quad V\Omega_v \bar{\beta} = \lambda \bar{\beta} \tag{3.128}$$

where $\Omega_v = M_v \Omega M_v^T$ as given by Eqs. 3.96- 3.98.

According to Eq. 3.128 the pool of candidate solutions is formed by the $\bar{\beta}$ eigenvectors of the $V\Omega_v = VM_v\Omega M_v^T$ matrix. If the normality of the vectors $V^{-1/2}\bar{\beta}$ is further assumed i.e. $[V^{-1/2}\bar{\beta}]^T[V^{-1/2}\bar{\beta}] = \sum_{k=1}^{N} \beta_k^2/v_k = 1$ the quantity to be maximized in Eq 3.123 is

$$\sum_{k=1}^{N} v_k [0 - \bar{w}^T(\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})})]^2 = \sum_{k=1}^{N} v_k e_k^2 = \sum_{k=1}^{N} \frac{v_k \beta_k^2}{\gamma^2 v_k^2} = \frac{1}{\gamma^2} \sum_{k=1}^{N} \frac{\beta_k^2}{v_k} = \frac{1}{\gamma^2} = \lambda^2 \tag{3.129}$$

Therefore, the optimal solution of the dual problem is the $\bar{\beta}^{(1)}$ eigenvector of the $V\Omega_v = VM_v\Omega M_v^T$ matrix corresponds to the largest eigenvalue λ_1. The

score variables can be given by

$$\begin{aligned}
z(\bar{x}_k) &= \bar{w}^T(\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}) = \\
&= \left(\sum_{l=1}^{N} \beta_l(\varphi(\bar{x}_l) - \hat{\mu}^v_{\varphi(\bar{x})})\right)^T (\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}) \\
&= \sum_{l=1}^{N} \beta_l[\varphi(\bar{x}_l) - \hat{\mu}^v_{\varphi(\bar{x})}]^T[\varphi(\bar{x}_k) - \hat{\mu}^v_{\varphi(\bar{x})}] \\
&= \sum_{l=1}^{N} \beta_l \left\{ K(\bar{x}_l, \bar{x}_k) - \frac{1}{\sum_{i=1}^N v_i} \sum_{i=1}^{N} v_i K(\bar{x}_l, \bar{x}_i) \right. \\
&\left. - \frac{1}{\sum_{i=1}^N v_i} \sum_{i=1}^{N} v_i K(\bar{x}_k, \bar{x}_i) + \frac{1}{\left[\sum_{i=1}^N v_i\right]^2} \sum_{i=1}^{N} v_i \sum_{j=1}^{N} v_j K(\bar{x}_i, \bar{x}_j) \right\}
\end{aligned} \tag{3.130}$$

using Eq 3.121, first equation from Eqs 3.126 and 3.98 as the result of applying the kernel trick on $\Omega_{v_{kl}}$.

The weighted kernel PCA eigenvalue problem given by Eq 3.128 as well as the expression for the score variables given above by Eq. 3.130 can be further simplified after realising that the optimal solutions of the dual problem i.e. the $\bar{\beta}$ vectors are going to be centered i.e. $\sum_{k=1}^N \beta_k = 0$. This condition is explicitly not among the KKT conditions for optimality given by Eqs 3.126 but it is there implicitly since

$$\begin{aligned}
\sum_{k=1}^{N} \beta_k &= \gamma \sum_{k=1}^{N} v_k e_k = \gamma \sum_{k=1}^{N} v_k \bar{w}^T(\varphi(\bar{x_k}) - \hat{\mu}^v_{\varphi(\bar{x})}) = \\
&= \gamma \bar{w}^T \left\{ \sum_{k=1}^{N} v_k \varphi(\bar{x_k}) - \frac{\sum_{k=1}^N v_k \sum_{i=1}^N v_i \varphi(\bar{x}_i)}{\sum_{i=1}^N v_i} \right\} \\
&= \gamma \bar{w}^T \sum_{k=1}^{N} v_k \varphi(\bar{x_k}) \underbrace{\left\{ 1 - \frac{\sum_{i=1}^N v_i}{\sum_{i=1}^N v_i} \right\}}_{0} = 0
\end{aligned} \tag{3.131}$$

Then the summation in Eqs. 3.127 can be written as

$$\begin{aligned}
&\sum_{l=1}^{N}\beta_l[\varphi(\bar{x}_l)-\hat{\mu}^v_{\varphi(\bar{x})}]^T[\varphi(\bar{x}_k)-\hat{\mu}^v_{\varphi(\bar{x})}] = \\
&=\sum_{l=1}^{N}\beta_l[\varphi(\bar{x}_l)^T\varphi(\bar{x}_k)-\varphi(\bar{x}_l)^T\hat{\mu}^v_{\varphi(\bar{x})}-\hat{\mu}^{vT}_{\varphi(\bar{x})}\varphi(\bar{x}_k)+\hat{\mu}^{vT}_{\varphi(\bar{x})}\hat{\mu}^v_{\varphi(\bar{x})}] \\
&=\sum_{l=1}^{N}\beta_l\varphi(\bar{x}_l)^T\varphi(\bar{x}_k)-\sum_{l=1}^{N}\beta_l\varphi(\bar{x}_l)^T\hat{\mu}^v_{\varphi(\bar{x})} \\
&\quad-\hat{\mu}^{vT}_{\varphi(\bar{x})}\varphi(\bar{x}_k)\underbrace{\sum_{l=1}^{N}\beta_l}_{0}+\hat{\mu}^{vT}_{\varphi(\bar{x})}\hat{\mu}^v_{\varphi(\bar{x})}\underbrace{\sum_{l=1}^{N}\beta_l}_{0} \\
&=\sum_{l=1}^{N}\beta_l\varphi(\bar{x}_l)^T\varphi(\bar{x}_k)-\sum_{l=1}^{N}\beta_l\varphi(\bar{x}_l)^T\hat{\mu}^v_{\varphi(\bar{x})} \\
&=\sum_{l=1}^{N}\beta_l\varphi(\bar{x}_l)^T[\varphi(\bar{x}_k)-\hat{\mu}^v_{\varphi(\bar{x})}]
\end{aligned} \tag{3.132}$$

by using the previously derived condition for $\bar{\beta}$ that $\sum_{k=1}^{N}\beta_k = 0$. Therefore, the weighted kernel PCA eigenproblem given by Eq 3.128 can be written in the following equivalent but simpler form

$$\boxed{\text{D:}}\ solve\ in\ \beta \quad V\Omega^*_v\bar{\beta} = \lambda\bar{\beta} \tag{3.133}$$

where $\Omega^*_v = M_v\Omega$ as given by 3.104. The expression for the score variables given by Eq. 3.130 also simplifies to

$$\begin{aligned}
z(\bar{x}_k) &= \bar{w}^T(\varphi(\bar{x}_k)-\hat{\mu}^v_{\varphi(\bar{x})}) = \\
&=\left(\sum_{l=1}^{N}\beta_l(\varphi(\bar{x}_l)-\hat{\mu}^v_{\varphi(\bar{x})})\right)^T(\varphi(\bar{x}_k)-\hat{\mu}^v_{\varphi(\bar{x})}) \\
&=\sum_{l=1}^{N}\beta_l[\varphi(\bar{x}_l)-\hat{\mu}^v_{\varphi(\bar{x})}]^T[\varphi(\bar{x}_k)-\hat{\mu}^v_{\varphi(\bar{x})}] \\
&=\sum_{l=1}^{N}\beta_l\varphi(\bar{x}_l)^T[\varphi(\bar{x}_k)-\hat{\mu}^v_{\varphi(\bar{x})}] \\
&=\sum_{l=1}^{N}\beta_l\left\{K(\bar{x}_l,\bar{x}_k)-\frac{1}{\sum_{i=1}^{N}v_i}\sum_{i=1}^{N}v_iK(\bar{x}_l,\bar{x}_i)\right\}
\end{aligned} \tag{3.134}$$

by using Eq 3.121, 3.132 and Eq 3.104. It is easy to see that the last two terms in Eq. 3.130 give zero contributions after the summation since $\sum_{k=1}^{N} \beta_k = 0$ as it was shown by Eq 3.132.

Therefore, the two different forms of the weighted kernel PCA eigenvalue problem given by Eq. 3.128 and 3.133 give identical solutions i.e.

$$V\Omega_v\bar{\beta} = V\Omega_v^*\bar{\beta} = \lambda\bar{\beta} \tag{3.135}$$

but the expression of the kernel matrix as well as the score variables are simpler in case of using Ω_v^* instead of Ω_v. It is the consequence of making use of the fact that the optimal solutions of the dual problems i.e. the $\bar{\beta}$ eigenvectors are centered i.e. $\sum_{k=1}^{N} \beta_k = 0$.

Automatic centering with bias

The score variables are

$$z_k = z(\bar{x}_k) = \bar{w}^T\varphi(\bar{x}_k) + b \tag{3.136}$$

The maximization problem can be written as

$$\max_{\bar{w}} \sum_{k=1}^{N} v_k[0 - (\bar{w}^T\varphi(\bar{x}_k) + b)]^2 \tag{3.137}$$

The corresponding primal optimization problem can be formulated as

$$\left|\begin{array}{l} \boxed{\text{P:}} \\ \quad \max_{\bar{w},\bar{e}} J(\bar{w},\bar{e}) \;=\; \gamma\frac{1}{2}\sum_{k=1}^{N} v_k e_k^2 - \frac{1}{2}\bar{w}^T\bar{w} \\ \quad \textit{such that} \quad e_k \;=\; \bar{w}^T\varphi(\bar{x}_k) + b, \quad k = 1,...,N \end{array}\right| \tag{3.138}$$

The Lagragian of this constrained optimization problem is

$$\mathcal{L}(\bar{w},\bar{e};\bar{\beta}) = \gamma\frac{1}{2}\sum_{k=1}^{N} v_k e_k^2 - \frac{1}{2}\bar{w}^T\bar{w} - \sum_{k=1}^{N}\beta_k(e_k - (\bar{w}^T\varphi(\bar{x}_k) + b)) \tag{3.139}$$

with Lagrange multipliers β_k $k = 1,...,N$. Note, that Eq. 3.137 is equivalent with 3.108 and Eqs. 3.138 can be obtained from Eqs. 3.109 by substituting

$e_k^* = \sqrt{v_k} e_k$ that results in the Lagragian given above. The KKT conditions for optimality give

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \bar{w}} = 0 \quad \rightarrow \quad \bar{w} = \displaystyle\sum_{k=1}^{N} \beta_k \varphi(\bar{x}_k) \\ \dfrac{\partial \mathcal{L}}{\partial e_k} = 0 \quad \rightarrow \quad \beta_k = \gamma v_k e_k, \qquad\qquad k = 1, ..., N \\ \dfrac{\partial \mathcal{L}}{\partial b} = 0 \quad \rightarrow \quad \displaystyle\sum_{k=1}^{N} \beta_k = 0 \\ \dfrac{\partial \mathcal{L}}{\partial \beta_k} = 0 \quad \rightarrow \quad e_k - (\bar{w}^T \varphi(\bar{x}_k) + b) = 0, \quad k = 1, ..., N \end{cases} \tag{3.140}$$

Eliminating the variables $\bar{w}$ and e_k by substituting the firts and second equations into the fourth in Eqs 3.140 yields

$$\frac{\beta_k}{\gamma} - v_k \left\{ \sum_{l=1}^{N} \beta_l \varphi(\bar{x}_l)^T \varphi(\bar{x}_k) + b \right\} = 0, \quad k = 1, ..., N \tag{3.141}$$

Then applying the third KKT condition from Eqs. 3.140 (i.e. $\sum_{k=1}^{N} \beta_k = 0$) to Eq. 3.141 gives the expression for the bias term b

$$\begin{aligned} &\sum_{k=1}^{N} \frac{\beta_k}{\gamma} - \sum_{k=1}^{N} v_k \sum_{l=1}^{N} \beta_l \varphi(\bar{x}_l)^T \varphi(\bar{x}_k) - \sum_{k=1}^{N} v_k b = 0 \\ &\frac{1}{\gamma} \sum_{k=1}^{N} \beta_k - \sum_{k=1}^{N} v_k \sum_{l=1}^{N} \beta_l \varphi(\bar{x}_l)^T \varphi(\bar{x}_k) - b \sum_{k=1}^{N} v_k = 0 \\ &b = - \frac{\sum_{k=1}^{N} v_k \sum_{l=1}^{N} \beta_l \varphi(\bar{x}_l)^T \varphi(\bar{x}_k)}{\sum_{k=1}^{N} v_k} \end{aligned} \tag{3.142}$$

Substituting this back into Eq. 3.141 leads to the following set of expressions for $\bar{\beta} = [\beta_1, ..., \beta_N]$

$$\frac{\beta_k}{\gamma} - v_k \left\{ \sum_{l=1}^{N} \beta_l \varphi(\bar{x}_l)^T \varphi(\bar{x}_k) - \frac{\sum_{i=1}^{N} v_i \sum_{l=1}^{N} \beta_l \varphi(\bar{x}_l)^T \varphi(\bar{x}_i)}{\sum_{i=1}^{N} v_i} \right\} = 0, \tag{3.143}$$

The expression in the curly brackets can be written as

$$\begin{aligned}
&\sum_{l=1}^{N}\beta_l\varphi(\bar{x}_l)^T\varphi(\bar{x}_k)-\frac{\sum_{i=1}^{N}v_i\sum_{l=1}^{N}\beta_l\varphi(\bar{x}_l)^T\varphi(\bar{x}_i)}{\sum_{i=1}^{N}v_i}=\\
&=\sum_{l=1}^{N}\beta_l\varphi(\bar{x}_l)^T\varphi(\bar{x}_k)-\sum_{l=1}^{N}\beta_l\varphi(\bar{x}_l)^T\frac{\sum_{i=1}^{N}v_i\varphi(\bar{x}_i)}{\sum_{i=1}^{N}v_i}\\
&=\sum_{l=1}^{N}\beta_l\varphi(\bar{x}_l)^T\varphi(\bar{x}_k)-\sum_{l=1}^{N}\beta_l\varphi(\bar{x}_l)^T\hat{\mu}^v_{\varphi(\bar{x})}\\
&=\sum_{l=1}^{N}\beta_l\varphi(\bar{x}_l)^T[\varphi(\bar{x}_k)-\hat{\mu}^v_{\varphi(\bar{x})}]
\end{aligned}\tag{3.144}$$

Therefore, Eq. 3.143 can be written as

$$\frac{\beta_k}{\gamma}-v_k\sum_{l=1}^{N}\beta_l\varphi(\bar{x}_l)^T[\varphi(\bar{x}_k)-\hat{\mu}^v_{\varphi(\bar{x})}]=0,\quad k=1,...,N \tag{3.145}$$

by using Eq. 3.144. Note that these equations are identical to those given by Eq. 3.133 obtained previously by using the explicit centering approach intead of the bias term. Therefore, the equations Eq. 3.133 and 3.134 also hold now with the same remarks. Note, that using a bias term instead of the exlicit centering of the feature map directly leads to the simpler form of the solution due to the fact that the centering of the optimal solutions of the dual problem i.e. $\sum_{k=1}^{N}\beta_k=0$ explicitly appears among the KKT conditions given by Eq 3.140 contrary to Eq. 3.126. The identity given by Eq. 3.132 is also true thus the equivalence given by Eq. 3.135

$$V\Omega_v\bar{\beta}=V\Omega_v^*\bar{\beta}=\lambda\bar{\beta} \tag{3.146}$$

also holds.

The quantity to be maximized in Eq. 3.137 is also equal to the eigenvalue of the $V\Omega_v$ or the simpler $V\Omega_v^*$ matrix (assuming that the vectors $V^{-1/2}\bar{\beta}$ are normalized i.e. $[V^{-1/2}\bar{\beta}]^T[V^{-1/2}\bar{\beta}]=\sum_{k=1}^{N}\beta_k^2/v_k=1$)

$$\max_{\bar{w}}\sum_{k=1}^{N}v_k[0-(\bar{w}^T\varphi(\bar{x}_k)+b)]^2=\sum_{k=1}^{N}v_ke_k^2=\sum_{k=1}^{N}\frac{v_k\beta_k^2}{\gamma^2v_k^2}=\frac{1}{\gamma^2}\sum_{k=1}^{N}\frac{\beta_k^2}{v_k}=\frac{1}{\gamma^2}=\lambda^2 \tag{3.147}$$

thus similar remarks, concerning the optimal solution of the dual problem, hold now as given previously in the explicit (weighted) centering case.

Also note, that if the feature map is already weight centered i.e. $\hat{\mu}^v_{\varphi(\bar{x})} = \frac{\sum_{k=1}^N v_k \varphi(\bar{x}_k)}{\sum_{k=1}^N v_k} = \bar{0}$ then $\Omega_v = \Omega_v^* = \Omega$ and Eq. 3.135 becomes

$$V\Omega_v\bar{\beta} = V\Omega_v^*\bar{\beta} = V\Omega\bar{\beta} = \lambda\bar{\beta} \tag{3.148}$$

and the expression for the score variables given by Eq. 3.134 further simplifies to

$$z(\bar{x}_k) = \sum_{l=1}^{N} \beta_l K(\bar{x}_l, \bar{x}_k) \tag{3.149}$$

since the last term in Eq. 3.134 that involvs $\hat{\mu}^v_{\varphi(\bar{x})}$ becomes zero.

3.2.4 Remarks on the different formulations of the weighted kernel PCA

It was shown in section 3.2.2 that the optimal solutions of the dual problem are the leading eigenvectors of the $V^{\frac{1}{2}}\Omega_v V^{\frac{1}{2}} = V^{\frac{1}{2}} M_v \Omega M_v^T V^{\frac{1}{2}}$ real *symmetric* p.s.d. matrix thus it has N real nonzero eigenvalues and the corresponding eigenvectors are orthogonal. Furthermore, it was also shown that these $\bar{\alpha}$ eigenvectors are *weight centered* i.e. $V^{\frac{1}{2}}\bar{\alpha} = \sum_{k=1}^N \bar{\alpha}_k \sqrt{v_k} = 0$ thus they are also the leading eigenvectors of the $V^{\frac{1}{2}}\Omega_v^* V^{\frac{1}{2}} = V^{\frac{1}{2}} M_v \Omega V^{\frac{1}{2}}$ matrix resulting in a simpler form of the solution. Therefore, the projections onto the different $\bar{\alpha}$ eigenvectors i.e. the different score variables as given by Eq. 3.100 or the simpler Eq. 3.105 are uncorrelated. Contrary to 3.2.2 the optimal solutions of the dual problem are expressed as leading eigenvectors of the $V\Omega_v = VM_v\Omega M_v^T$ *non-symmetric* p.s.d. matrix in section 3.2.3. Furthermore, it was also shown that these $\bar{\beta}$ eigenvectors are centered i.e. $\sum_{k=1}^N \bar{\beta}_k = 0$ thus they are also the leading eigenvectors of the $V\Omega_v^* = VM_v\Omega$ *non-symmetric* matrix resulting in a simpler form of the score variables given by Eq. 3.134 istead of simpler Eq. 3.130.

Note, that the explicit centering of the feature map and the bias term approach give exactly the same results in both cases but using a bias term directly leads to the simplest form. Furthermore, both approaches automatically lead to the weighted centering of the kernel matrix.

As it was already mentioned at Eqs. 3.125 and 3.139 the two formulations of the weighted kernel PCA problem are not independent. This can be easily seen by multiplying both side of the eigenproblems given by Eq. 3.135 with $V^{-\frac{1}{2}}$ and doing some algebra using the diagonality of the weight matrix V

as

$$V^{-\frac{1}{2}}V^{\frac{1}{2}}V^{\frac{1}{2}}\Omega_v V^{\frac{1}{2}}V^{-\frac{1}{2}}\bar{\beta} = \lambda V^{-\frac{1}{2}}\bar{\beta}$$

$$V^{-\frac{1}{2}}V^{\frac{1}{2}}V^{\frac{1}{2}}\Omega_v^* V^{\frac{1}{2}}V^{-\frac{1}{2}}\bar{\beta} = \lambda V^{-\frac{1}{2}}\bar{\beta} \tag{3.150}$$

$$V^{\frac{1}{2}}\Omega_v V^{\frac{1}{2}}\bar{\alpha} = \lambda\bar{\alpha}$$

$$V^{\frac{1}{2}}\Omega_v^* V^{\frac{1}{2}}\bar{\alpha} = \lambda\bar{\alpha}$$

where $\bar{\alpha} = V^{-\frac{1}{2}}\bar{\beta}$ was used. Therefore, the eigenproblems for the optimal solutions of the dual problem Eqs. 3.135 can be transformed to the eigenproblems Eqs. 3.106 by multiplying both sides with $V^{-\frac{1}{2}}$ and substituting $\bar{\beta} = V^{\frac{1}{2}}\bar{\alpha}$ and vice versa by multiplying both sides of Eqs. 3.106 with $V^{\frac{1}{2}}$ and substituting $\bar{\alpha} = V^{-\frac{1}{2}}\bar{\beta}$.

Also note, that the extra condition (when a bias term is used instead of the explicit centering) $\sum_{k=1}^{N}\beta_k = 0$ as well as the normality assumption $\sum_{k=1}^{N}\beta_k^2/v_k = 1$ of section 3.2.3 transform to the corresponding condition $\sum_{k=1}^{N}\beta_k = \sum_{k=1}^{N}\alpha_k\sqrt{v_k} = 0$ and normality assumption $\sum_{k=1}^{N}\beta_k^2/v_k = \sum_{k=1}^{N}\alpha_k^2 v_k/v_k = \sum_{k=1}^{N}\alpha_k^2 = 1$ of section 3.2.2 and vice versa. Furthermore, since $\bar{\alpha}$ vectors are eigenvectors of a real symmetric p.s.d. matrix the N eigenvectors are mutually orthogonal i.e.

$$[\bar{\alpha}^{(i)}]^T[\bar{\alpha}^{(j)}] = 0,\ \forall\, i,j \quad i \neq j, \quad i,j = 1,..,N \tag{3.151}$$

as it was already mentioned above. Thus the $V^{-\frac{1}{2}}\bar{\beta}$ vectors are going to be orthogonal insted of the $\bar{\beta}$ eigenvectors that can be seen by substituting $\bar{\alpha} = V^{-\frac{1}{2}}\bar{\beta}$ above. This observation can be directly derived after converting the $V\Omega_v\bar{\beta} = \lambda\bar{\beta}$ *ordinary non-symmetric* eigenvalue problem into the following *generalized* eigenvalue problem by multiplying with V^{-1} from the left

$$\Omega_v\bar{\beta} = \lambda V^{-1}\bar{\beta} \tag{3.152}$$

Since Ω_v and V^{-1} are a real symmetric, Ω_v is p.s.d. and V^{-1} is p.d. this problem corresponds to the usual real symmetric eigenproblem [7]. It means that there are N real non-negative eigenvalues of this problem and the corresponding $\bar{\beta}$ eigenvectors form an orthogonal set with respect to the V^{-1}-*inner product* defined as

$$(\bar{x},\bar{y})_{V^{-1}} = (V^{-1}\bar{x},\bar{y}) \tag{3.153}$$

and the V^{-1}-*norm* is given by

$$\|\bar{x}\|_{V^{-1}} = (V^{-1}\bar{x},\bar{x})^{-1/2} \tag{3.154}$$

Note, that although the matrix $V\Omega_v$ of the standard eigenproblem is non-symmetric with respect to the Euclidean inner product, it is symmetric with respect to the V^{-1}-*inner product* since

$$(V\Omega_v\bar{x},\bar{y})_{V^{-1}} = (\bar{x},V\Omega_v\bar{y})_{V^{-1}},\ \forall\ \bar{x},\bar{y} \tag{3.155}$$

The V^{-1}-*orthogonality* of the $\bar{\beta}$ eigenvectors mentioned before Eq. 3.153 yields

$$[V^{-1}\bar{\beta}^{(i)}]^T[\bar{\beta}^{(j)}] = [V^{-1/2}\bar{\beta}^{(i)}]^T[V^{-1/2}\bar{\beta}^{(j)}] = 0,\ \forall\ i,j \quad i\neq j, \quad i,j=1,..,N \tag{3.156}$$

using the V^{-1}-*inner product* defined above by Eq. 3.153 and the fact that V^{-1} matrix is diagonal. Note that is exactly the orthogonality of the $V^{-\frac{1}{2}}\bar{\beta}$ vectors derived above by using the orthogonality of the $\bar{\alpha}$ eigenvectors and the transformation between $\bar{\alpha}$ and $\bar{\beta}$. Furthermore, the normality condition for the $\bar{\beta}$ eigenvectors i.e.

$$\sum_{k=1}^{N}\beta_k^2/v_k = \bar{\beta}^TV^{-1}\bar{\beta} = 1 \tag{3.157}$$

is exactly the V^{-1}-*normality* of the $\bar{\beta}$ eigenvectors according to the definition of V^{-1}-*norm* given by Eq. 3.154. The meaning of Eqs 3.156 and 3.157 is that the $\bar{\beta}$ vectors i.e. the N eigenvectors of the $V\Omega_v$ matrix form a mutually V^{-1}-*orthonormal* set. Morover, it can be seen that the $\frac{1}{\sum_{k=1}^{N}v_k}[v_1,...,v_N]$ (V^{-1}-*normalized*) vector is one of the N eigenvectors of the $V\Omega_v$ matrix with eigenvalue $\lambda=0$. Therefore, the remaining $N-1$ eigenvectors must be V^{-1}-*orthogonal* to this eigenvector that leads to the centering of the $N-1$ remaining $\bar{\beta}$ eigenvectors of the $V\Omega_v$ matrix

$$0 = \frac{1}{\sum_{k=1}^{N}v_k}\bar{\beta}^TV^{-1}[v_1,...,v_N]^T = \frac{1}{\sum_{k=1}^{N}v_k}\bar{\beta}^T\bar{\mathbb{1}}_N = \frac{1}{\sum_{k=1}^{N}v_k}\sum_{k=1}^{N}\beta_k \tag{3.158}$$

The optimal solution of the maximization problem given in section 3.2.3 is given by the $K-1$ ($K\leq N$) leading eigenvectors of the $V\Omega_v$ matrix as it was shown by Eq. 3.129. Since the $\frac{1}{\sum_{k=1}^{N}v_k}[v_1,...,v_N]$ vector is one of the N eigenvectors of the $V\Omega_v$ matrix with (a minimal) $\lambda=0$ eigenvalue the remaining $N-1$ centered eigenvectors are going to be selected before this. Therefore, all the $K-1$ ($K\leq N$) leading eigenvectors of the $V\Omega_v$ matrix are going to be centered. Furthermore, it was also shown in section 3.2.3 that these $N-1$ eigenvectors of the $V\Omega_v$ matrix are also eigenvectors of the $V\Omega_v^*$

matrix with the same eigenvalue due to the fact that they are centered.

The conditions for the $\bar{\alpha}$ eigenvectors in the case of problem formulation given in section 3.2.2 can be given similarly by changing the V^{-1}-*norm* used above to the standard Euclidean norm since the optimal solution of the maximization problem in terms of dual variables is now given by the leading $\bar{\alpha}$ eigenvectors of the symmetric $V^{\frac{1}{2}}\Omega_v V^{\frac{1}{2}}$ matrix. The normalized N eigenvectors of this matrix form a mutually orthonormal set

$$[\bar{\alpha}^{(i)}]^T[\bar{\alpha}^{(j)}] = \delta_{ij}, \quad \forall \quad i,j = 1,..,N \tag{3.159}$$

where δ_{ij} is the Kronecker delta. It can be seen that the $\frac{1}{\sum_{k=1}^N v_k}[\sqrt{v_1},...,\sqrt{v_N}]$ (normalized) vector is one of the N eigenvectors of the $V^{\frac{1}{2}}\Omega_v V^{\frac{1}{2}}$ matrix with zero eigenvalue. Thus all the remaining $N-1$ eigenvectors must be orthogonal to this vector that leads to the $\sqrt{v}$ weighted centering of these remaining $N-1$ $\bar{\alpha}$ eigenvectors of the $V^{\frac{1}{2}}\Omega_v V^{\frac{1}{2}}$ matrix

$$0 = \frac{1}{\sum_{k=1}^N v_k}\bar{\alpha}^T[\sqrt{v_1},...,\sqrt{v_N}]^T = \frac{1}{\sum_{k=1}^N v_k}\sum_{k=1}^N \bar{\alpha}_k\sqrt{v_k} \tag{3.160}$$

The optimal solution of the maximization problem given in section 3.2.2 is given by the $K-1$ ($K \leq N$) leading eigenvectors of the $V^{\frac{1}{2}}\Omega_v V^{\frac{1}{2}}$ matrix as it was shown by Eq. 3.99. Since the $\frac{1}{\sum_{k=1}^N v_k}[\sqrt{v_1},...,\sqrt{v_N}]$ vector is one of the N eigenvectors of the $V^{\frac{1}{2}}\Omega_v V^{\frac{1}{2}}$ matrix with (a minimal) $\lambda = 0$ eigenvalue the remaining $N-1$ $\sqrt{v}$ weight centered eigenvectors are going to be selected before this. Therefore, all the $K-1$ ($K \leq N$) leading eigenvectors of the $V^{\frac{1}{2}}\Omega_v V^{\frac{1}{2}}$ matrix are going to be $\sqrt{v}$ weight centered. Furthermore, it was also shown in section 3.2.2 that these $N-1$ eigenvectors of the $V^{\frac{1}{2}}\Omega_v V^{\frac{1}{2}}$ matrix are also eigenvectors of the $V^{\frac{1}{2}}\Omega_v^* V^{\frac{1}{2}}$ matrix with the same eigenvalue due to the fact that they are $\sqrt{v}$ weight centered.

It can also be seen, that the projections onto the $\bar{\alpha}$ eigenvectors given by Eq. 3.105 (as well as Eq. 3.100) in section 3.2.2 transform to the projections onto the $\bar{\beta}$ eigenvectors given by Eq. 3.130 (as well as Eq. 3.134) in section 3.2.3 after substituting $\alpha_l = \beta_l/\sqrt{v_l}$ into Eq. 3.105 (or into Eq. 3.100) and simplifying with $\sqrt{v_l}$. Similary, the projections onto the $\bar{\beta}$ eigenvectors given by Eq. 3.134 (or Eq. 3.130) in section 3.2.3 transform the projections onto the $\bar{\alpha}$ eigenvectors given by Eqs. 3.105 (or Eq. 3.100) in section 3.2.2 after substituting $\beta_l = \alpha_l\sqrt{v_l}$ into Eq. 3.134 (or into Eq. 3.130).

Furthermore, if the weight matrix V is the identity matrix i.e. $V = I$ then

$\Omega_v = \Omega_c$ and $\Omega_v^* = \Omega_c^*$ that transform both formulations of the weighted kernel PCA to the (unweighted) kernel PCA formulation given in section 3.2.1.

3.3 The relation of weighted kerenl PCA to the basic KSC problems

Before showing the relation between the previously derived weighted kernel PCA and the basic spectral clustering algorithms discussed in Chapter 1 first we recall the ingredients of the basic kernel spectral clustering algorithms. The *unnormalized*, *normalized symmetric* and *normalized random walk* graph Laplacians are

$$\begin{array}{lcl} L & := & D - W \\ L_{sym} & := & D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \\ L_{rw} & := & D^{-1} L = I - D^{-1} W \end{array} \tag{3.161}$$

respectively. W is the *weighted adjacency matrix* associated with the undirected *symilarity graph* of the input data points obtained by using some non-negative valued *similarity function* and D is the diagonal *degree matrix* related to the *similarity graph* with diagonal elements equal to the row sums of the *weighted adjacency matrix* W. Furthermore, it is assumed that the similarity function used to produce the weighted adjacency matrix W corresponds to a Mercer kernel i.e. $W \equiv \Omega$ and the graph Laplacians given by Eqs. 3.161 will be used in the form of $I - L_{(.)}$ in the corresponding eigenproblem.

3.3.1 The symmetric normalized graph Laplacian kernel spectral clustering algorithm

The kernel spectral clustering algorithm related to the *symmetric* normalized graph Laplacian is given by CH1 **??** and the core problem can be formulated as an eigenvalue problem involving the second graph Laplacian from Eq. 3.161 as

$$\boxed{D^{-\frac{1}{2}} \Omega D^{-\frac{1}{2}} \bar{\alpha} = \lambda \bar{\alpha}} \tag{3.162}$$

It can be seen that this eigenproblem is similar to the *weighted* kernel PCA eigenproblem especially to that form of the *weighted* kernel PCA that was derived in section 3.2.2. It was shown there that the formulations of the weighted kernel PCA using both the explicit *weighted* centering of the feature

map and the bias term approach result in the same eigenvalue problem given by Eq. 3.106 as

$$V^{\frac{1}{2}}\Omega_v V^{\frac{1}{2}}\bar{\alpha} = V^{\frac{1}{2}}\Omega_v^* V^{\frac{1}{2}}\bar{\alpha} = \lambda\bar{\alpha} \tag{3.163}$$

where $\Omega_v = M_v\Omega M_v$, $\Omega_v^* = M_v\Omega$ and M_v is the weighted centering matrix defined by Eq. 3.97. It was also mentioned at the end of section 3.2.2 that these weight centered kernel matrices becomes the kernel matrix i.e. $\Omega_v = \Omega_v^* = \Omega$ if the feature map is already weight centered i.e. $\hat{\mu}^v_{\varphi(\bar{x})} = \frac{\sum_{k=1}^N v_k\varphi(\bar{x}_k)}{\sum_{k=1}^N v_k} = \bar{0}$ and the eigenproblem becomes

$$V^{\frac{1}{2}}\Omega_v V^{\frac{1}{2}}\bar{\alpha} = V^{\frac{1}{2}}\Omega_v^* V^{\frac{1}{2}}\bar{\alpha} = V^{\frac{1}{2}}\Omega V^{\frac{1}{2}}\bar{\alpha} = \lambda\bar{\alpha} \tag{3.164}$$

as given by Eq. 3.119. Choosing the diagonal weight matrix being equal to the inverse degree matrix of the similarity graph i.e. $V = D^{-1}$ results in the downweighting of the contributions comming from weakly connected points i.e. from outliers to the summation in the maximization problem given by Eqs. 3.90 and 3.108. Thus, this obvious choice of the weight matrix makes the weighted kernel PCA robust against outliers. Furthermore, the weighted kernel PCA eigenproblem given by Eq. 3.164 becomes identical with Eq. 3.162 after substituting $V = D^{-1}$ into Eq. 3.164. Therefore, the kernel spectral clustering core problem related to the *symmetric* normalized graph Laplacian is identical with the *weighted* kernel PCA problem chosing the weight matrix being equal to the inverse degree matrix i.e. $V = D^{-1}$ and supposing that the feature map is already weight centered i.e. $\hat{\mu}^v_{\varphi(\bar{x})} = \frac{\sum_{k=1}^N v_k\varphi(\bar{x}_k)}{\sum_{k=1}^N v_k} = \bar{0}$. Furthermore, the leading $i = 1, ..., K-1,\ K \leq N$ $\bar{\alpha}$ eigenvectors are going to be $\sqrt{v}$ centered in this case i.e. $\sum_{k=1}^N \alpha_k^{(i)}\sqrt{v_k} = 0$.

Keeping the assumption that $V = D^{-1}$ but discarding that the feature map is already weight centered i.e. $\hat{\mu}^v_{\varphi(\bar{x})} = \frac{\sum_{k=1}^N v_k\varphi(\bar{x}_k)}{\sum_{k=1}^N v_k} \neq \bar{0}$ the weighted kernel PCA eigenproblems given by Eq. 3.163 hold insted of Eq. 3.164. Therefore, the weight centered kernel matrices Ω_v or Ω_v^* will play the role of the kernel matrix Ω and both eigenproblems given by Eq. 3.163 will give identical solutions. Furthermore, the only difference between the results concerning both the eigenvalues and eigenvectors in Eq. 3.162 and Eq. 3.163 is that the leading $i = 1, ..., K-1,\ K \leq N$ $\bar{\alpha}$ eigenvectors are going to be $\sqrt{v}$ weight centered in Eq. 3.163 i.e. $\sum_{k=1}^N \alpha_k^{(i)}\sqrt{v_k} = 0$.

3.3.2 The random walk normalized graph Laplacian kernel spectral clustering algorithm

The kernel spectral clustering algorithm related to the *random walk* normalized graph Laplacian is given by CH1 **??** and the core problem can be formulated as an eigenvalue problem involving the third graph Laplacian from Eq. 3.161 as

$$\boxed{D^{-1}\Omega\bar{\beta} = \lambda\bar{\beta}} \tag{3.165}$$

It can be seen that this eigenproblem is similar to the *weighted* kernel PCA eigenproblem especially to that form of the *weighted* kernel PCA that was derived in section 3.2.3. It was shown there that the formulations of the weighted kernel PCA using both the explicit *weighted* centering of the feature map and the bias term approach result in the same eigenvalue problem given by Eq. 3.135 as

$$V^1\Omega_v\bar{\beta} = V^1\Omega_v^*\bar{\beta} = \lambda\bar{\beta} \tag{3.166}$$

where $\Omega_v = M_v\Omega M_v$, $\Omega_v^* = M_v\Omega$ and M_v is the weighted centering matrix defined by Eq. 3.97. It was also mentioned at the end of section 3.2.3 that these weight centered kernel matrices becomes the kernel matrix i.e. $\Omega_v = \Omega_v^* = \Omega$ if the feature map is already weight centered i.e. $\hat{\mu}^v_{\varphi(\bar{x})} = \frac{\sum_{k=1}^N v_k\varphi(\bar{x}_k)}{\sum_{k=1}^N v_k} = \bar{0}$ and the eigenproblem becomes

$$V^1\Omega_v\bar{\beta} = V^1\Omega_v^*\bar{\beta} = V^1\Omega\bar{\beta} = \lambda\bar{\beta} \tag{3.167}$$

as given by Eq. 3.148. Choosing the diagonal weight matrix being equal to the inverse degree matrix of the similarity graph i.e. $V = D^{-1}$ results in the downweighting of the contributions comming from weakly connected points i.e. from outliers to the summation in the maximization problem given by Eqs. 3.123 and 3.137. Thus, this obvious choice of the weight matrix makes the weighted kernel PCA robust against outliers. Furthermore, the weighted kernel PCA eigenproblem given by Eq. 3.167 becomes identical with Eq. 3.165 after substituting $V = D^{-1}$ into Eq. 3.167. Therefore, the kernel spectral clustering core problem related to the *random walk* normalized graph Laplacian is identical with the *weighted* kernel PCA problem chosing the weight matrix being equal to the inverse degree matrix i.e. $V = D^{-1}$ and supposing that the feature map is already weight centered i.e. $\hat{\mu}^v_{\varphi(\bar{x})} = \frac{\sum_{k=1}^N v_k\varphi(\bar{x}_k)}{\sum_{k=1}^N v_k} = \bar{0}$. Furthermore, the leading $i = 1, ..., K-1,\ K \leq N$ $\bar{\beta}$ eigenvectors are going to be centered in this case i.e. $\sum_{k=1}^N \beta_k^{(i)} = 0$.

Keeping the assumption that $V = D^{-1}$ but discarding that the feature map

is already weight centered i.e. $\hat{\mu}^v_{\varphi(\bar{x})} = \frac{\sum_{k=1}^N v_k \varphi(\bar{x}_k)}{\sum_{k=1}^N v_k} \neq \bar{0}$ the weighted kernel PCA eigenproblems given by Eq. 3.166 hold insted of Eq. 3.167. Therefore, the weight centered kernel matrices Ω_v or Ω_v^* will play the role of the kernel matrix Ω and both eigenproblems given by Eq. 3.166 will give identical solutions. Furthermore, the only difference between the results concerning both the eigenvalues and eigenvectors in Eq. 3.165 and Eq. 3.166 is that the leading $i = 1, ..., K-1,\ K \leq N$ $\bar{\beta}$ eigenvectors are going to be centered in Eq. 3.166 i.e. $\sum_{k=1}^N \beta_k^{(i)} = 0$.

3.4 Weighted kernel K-means and weighted kernel PCA

The relation between classical K-means algorithm and linear PCA was investigated by *Ding et al.* [8]. It was shown that principal componets are the continuous relaxed solutions to the discrete cluster membership indicators for the K-means clustering. The spectral clustering objective of normalized cut was derived from the weighted kernel K-means objective function by *Dhillon et al.* [9] as a special case of weightening making the connection between weighted kernel K-means and kernel spectral clustering based on the symmetric normalized graph Laplacian. Here we will show the relation of weighted kernel K-means to the weighted kernel PCA given in sections 3.2.2 and 3.2.3 by using a formulation of the K-means objective similar to that used in [8] but including a weightening and explicit weighted centering of the feature map. Formulating the K-means objective this way will result in the weighted kernel PCA problems discussed in sections 3.2.2 and 3.2.3. According to the link between the weighted kernel PCA and normalized graph Laplacian spectral clustering algorithms established in sections 3.3.1 and 3.3.2 it immediately leads to a *centered* version of the normalized cut algorithm similarly to the results of *Dhillon et al.* [9]. Furthermore, it results in the link between the (kernel) K-means and (kernel) PCA shown by *Ding et al.* [8] when choosing the diagonal weight matrix being equal to the identity matrix. Therfore, a connection can be made between weighted kernel K-means, weighted kernel PCA and normalized cut spectral clustering algorithms as well as between kernel K-means and kernel PCA.

The aim of the classical K-means clustering algorithm is to find K proto-

types, the cluster centroids

$$\bar{\mu}_k = \frac{1}{n_k} \sum_{i \in C_k} \bar{x}_i, \quad k = 1, ..., K \tag{3.168}$$

by *minimizing* the following objective function

$$J_K = \sum_{k=1}^{K} \sum_{i \in C_k} (\bar{x}_i - \bar{\mu}_k)^2 \tag{3.169}$$

where n_k is the number of data points assigned to the $k-th$ cluster in the current partition of the $\{\bar{x}_j\}_{j=1}^N \in \mathcal{D}$, $\bar{x} \in \mathbb{R}^d$ input data set. Data pints are assigned to the $k-th$ cluster in the current step of iteration if the squared distance from the current $k-th$ cluster centorid is minimal.

In the *kernel* version of this classical K-means problem the d dimensional input data $\bar{x} \in \mathbb{R}^d$ are mapped to a high n_h dimensional feature space by using the $\varphi(\cdot) : \mathbb{R}^d \to \mathbb{R}^{n_h}$ non-linear mapping (where n_h can even be inifite). Let introduce $v_i \in \mathbb{R}^+$ weights for each data points and let $\hat{\mu}^v_{\varphi(\bar{x})}$ denote the weighted mean of the feature map

$$\hat{\mu}^v_{\varphi(\bar{x})} = \frac{\sum_{j=1}^N v_j \varphi(\bar{x}_j)}{\sum_{j=1}^N v_j} \tag{3.170}$$

The contribution of data points assigned to the $k-th$ cluster J_k to the *weight centered weighted kernel* version of the classical K-means objective given by Eq. 3.169 can be written as

$$J_k = \sum_{i \in C_k} v_i (\varphi^*(\bar{x}_i) - \hat{\mu}^v_k)^2 \tag{3.171}$$

where $\varphi^*(\bar{x})$ is the weight centered feature map of the corresponding input data point $\bar{x}$ i.e.

$$\varphi^*(\bar{x}) = \varphi(\bar{x}) - \hat{\mu}^v_{\varphi(\bar{x})} \tag{3.172}$$

and

$$\hat{\mu}^v_k = \frac{\sum_{i \in C_k} v_i \varphi^*(\bar{x}_i)}{\sum_{i \in C_k} v_i} \tag{3.173}$$

is the weighted mean of the weight centered feature map of those data that assigned to the $k-th$ cluster.

Eq. 3.171 can be written as

$$
\begin{aligned}
J_k &= \sum_{i\in C_k} v_i(\varphi^*(\bar{x}_i) - \hat{\mu}_k^v)^2 = \sum_{i\in C_k} v_i {\varphi^*}^T(\bar{x}_i)\varphi^*(\bar{x}_i) + \sum_{i\in C_k} v_i \hat{\mu}_k^{v^T}\hat{\mu}_k^v \\
&- 2\sum_{i\in C_k} v_i {\varphi^*}^T(\bar{x}_i)\hat{\mu}_k^v = \sum_{i\in C_k} v_i {\varphi^*}^T(\bar{x}_i)\varphi^*(\bar{x}_i) \\
&+ \sum_{i\in C_k} v_i \frac{\sum_{j\in C_k} v_j {\varphi^*}^T(\bar{x}_j)\sum_{l\in C_k} v_l\varphi^*(\bar{x}_l)}{\left[\sum_{i\in C_k} v_i\right]^2} \\
&- 2\sum_{i\in C_k} v_i {\varphi^*}^T(\bar{x}_i)\frac{\sum_{j\in C_k} v_j {\varphi^*}^T(\bar{x}_j)}{\sum_{i\in C_k} v_i} = \sum_{i\in C_k} v_i {\varphi^*}^T(\bar{x}_i)\varphi^*(\bar{x}_i) \\
&+ \frac{\sum_{i,j\in C_k} v_i v_j {\varphi^*}^T(\bar{x}_i)\varphi^*(\bar{x}_j)}{\sum_{i\in C_k} v_i} - 2\frac{\sum_{i,j\in C_k} v_i v_j {\varphi^*}^T(\bar{x}_i)\varphi^*(\bar{x}_j)}{\sum_{i\in C_k} v_i} \\
&= \sum_{i\in C_k} v_i {\varphi^*}^T(\bar{x}_i)\varphi^*(\bar{x}_i) - \frac{\sum_{i,j\in C_k} v_i v_j {\varphi^*}^T(\bar{x}_i)\varphi^*(\bar{x}_j)}{\sum_{i\in C_k} v_i} \\
&= \sum_{i\in C_k} v_i[\varphi(\bar{x}_i) - \hat{\mu}^v_{\varphi(\bar{x})}]^T[\varphi(\bar{x}_i) - \hat{\mu}^v_{\varphi(\bar{x})}] \\
&- \frac{\sum_{i,j\in C_k} v_i v_j[\varphi(\bar{x}_i) - \hat{\mu}^v_{\varphi(\bar{x})}]^T[\varphi(\bar{x}_j) - \hat{\mu}^v_{\varphi(\bar{x})}]}{\sum_{i\in C_k} v_i}
\end{aligned}
\tag{3.174}
$$

Summing up the contributions of the K different clusters gives the *weight centered weighted kernel* version of the classical K-means objective

$$
\begin{aligned}
J_K = \sum_{k=1}^{K} \Bigg\{ &\sum_{i\in C_k} v_i[\varphi(\bar{x}_i) - \hat{\mu}^v_{\varphi(\bar{x})}]^T[\varphi(\bar{x}_i) - \hat{\mu}^v_{\varphi(\bar{x})}] \\
&- \frac{\sum_{i,j\in C_k} v_i v_j[\varphi(\bar{x}_i) - \hat{\mu}^v_{\varphi(\bar{x})}]^T[\varphi(\bar{x}_j) - \hat{\mu}^v_{\varphi(\bar{x})}]}{\sum_{i\in C_k} v_i} \Bigg\}
\end{aligned}
\tag{3.175}
$$

The summation of the first term can be immediately written as the trace of the $V\Omega_v$ matrix

$$
\sum_{i=1}^{N} v_i[M_v\Omega M_v^T]_{ii} = \sum_{i=1}^{N} v_i[\Omega_v]_{ii} = Tr[V\Omega_v] \tag{3.176}
$$

where Ω, Ω_v and M_v are the kernel matrix, weight centered kernel matrix and weighted centering matrix given by Eqs. 3.62- 3.64. Since the trace of the $V\Omega_v$ matrix is a constant the minimization of the *weight centered*

weighted kernel K-means objective given by Eq 3.175 is equivalent with the maximization of the second (negative) term in the summation. In order to be able to write this second term as a trace maximization let indroduce the $H_K \in \mathbb{R}^{NxK}_{\geq 0}$ orthonormal matrix with columns of $\bar{h}_k$

$$\bar{h}_k = [0, ..., 0, \underbrace{\underbrace{\sqrt{v_i}, ..., \sqrt{v_j}}_{i...j \in C_k}}_{n_k}, 0, ...0]^T / \sqrt{w_k} \tag{3.177}$$

cluster membership indicator vectors where $w_k = \sum_{i \in C_k} v_i$ and $w = \sum_{k=1}^{K} w_k = \sum_{i=1}^{N} v_i$. It should be noted here that these cluster membership indicators are given in a slightly different form than the usual

$$\bar{f}_k = [0, ..., 0, \underbrace{1, ..., 1}_{n_k}, 0, ...0]^T \tag{3.178}$$

$k = 1, ..., K$ mutually orthogonal vectors. One can interpret the vectors given in Eq. 3.177 as the normalized columns of the $V^{\frac{1}{2}} F_K$ orthogonal matrix where the K columns of F_K are the usual orthogonal cluster indicator vectors. For the shake of simplicity cluster membership indicators are going to be used in the form of Eq. 3.177 hereafter. It can be easily seen that one of these K $\bar{h}_k$ vectors can be written as the linear combination of the others since

$$\sum_{k=1}^{K} \sqrt{w_k} \bar{h}_k = [\sqrt{v_1}, ..., \sqrt{v_N}]^T = V^{\frac{1}{2}} \bar{\mathbb{1}}_N \tag{3.179}$$

always holds. This redundancy can be removed by performing a linear transformation T on H_K into $A_k = H_K T$ where $A_K \in \mathbb{R}^{NxK}$ and $T \in \mathbb{R}^{KxK}$ orthonormal matrix with last column of $t_K = [\sqrt{w_1}, ..., \sqrt{w_K}]^T / \sqrt{w}$. The result of choosing the last column of T this way is that the last column of A_K i.e. $\bar{\alpha}_K$ will be equal to

$$\bar{\alpha}_K = \frac{\sqrt{w_1}}{\sqrt{w}} \bar{h}_1 + ... + \frac{\sqrt{w_K}}{\sqrt{w}} \bar{h}_K = \sum_{k=1}^{K} \frac{\sqrt{w_k}}{\sqrt{w}} \bar{h}_k = [\sqrt{v_1}, ..., \sqrt{v_N}]^T / \sqrt{w} \tag{3.180}$$

The columns of both H_K and T are mutually orthonormal i.e. $H_K^T H_K = I_K$ and $T^T T = I_K$. Furthermore, it is easy to see that the colomns of A_K are also mutually orthonormal i.e.

$$A_K^T A_K = [H_K T]^T [H_K T] = \underbrace{T^T \underbrace{H_K^T H_K}_{I_K} T}_{I_K} = I_K \tag{3.181}$$

that can be written as

$$\bar{\alpha}^{T(i)}\bar{\alpha}^{(j)} = \delta_{ij}, \quad \forall \quad i,j = 1,...,K \leq N \tag{3.182}$$

Since the last column of A_K i.e. $\bar{\alpha}_K$ is always $\bar{\alpha}_K = [\sqrt{v_1},...,\sqrt{v_N}]^T/\sqrt{w}$ the remaining $K-1$ $\bar{\alpha}_k$ vectors must be orthogonal to this

$$\begin{aligned} 0 =& \bar{\alpha}^{(i)T}\bar{\alpha}^{(K)} = \frac{1}{\sqrt{w}}\bar{\alpha}^{(i)T}[\sqrt{v_1},...,\sqrt{v_N}]^T = \frac{1}{\sqrt{w}}\sum_{j=1}^{N}\alpha_j^{(i)}\sqrt{v_j}, \\ & \forall \quad i = 1,...,K-1, \quad K \leq N \end{aligned} \tag{3.183}$$

that results in the $\sqrt{v}$ centering of these $K-1$ $\bar{\alpha}_k$ vectors. The second term of the summation in Eq 3.175 can be now written as

$$\begin{aligned} &\sum_{k=1}^{K}\frac{\sum_{i,j\in C_k} v_i v_j[\varphi(\bar{x}_i)-\hat{\mu}^v_{\varphi(\bar{x})}]^T[\varphi(\bar{x}_j)-\hat{\mu}^v_{\varphi(\bar{x})}]}{\sum_{i\in C_k} v_i} = Tr[H_K^T V^{\frac{1}{2}} M_v \Omega M_v^T V^{\frac{1}{2}} H_K] \\ &= Tr[[A_K T^T]^T V^{\frac{1}{2}} M_v \Omega M_v^T V^{\frac{1}{2}}[A_K T^T]] = Tr[A_K^T V^{\frac{1}{2}} M_v \Omega M_v^T V^{\frac{1}{2}} A_K] \\ &= Tr[A_{K-1}^T V^{\frac{1}{2}} M_v \Omega M_v^T V^{\frac{1}{2}} A_{K-1}] + \frac{1}{w}\underbrace{\sum_{i=1}^{N}\sum_{j=1}^{N} v_i v_j [\Omega_v]_{ij}}_{0} \end{aligned} \tag{3.184}$$

by using the mutually orthonormal columns of H_K i.e. the cluster membership indicators and $A_K = H_K T$ orthonormal matrix. The last equation was derived by making use of the last $(K-th)$ column of A_K matrix given by Eq. 3.180 and the weight centered property of Ω_v matrix that gives

$$\begin{aligned} &\sum_{i=1}^{N}\sum_{j=1}^{N} v_i v_j [\Omega_v]_{ij} = \sum_{i=1}^{N}\sum_{j=1}^{N} v_i v_j \varphi(\bar{x}_i)^T\varphi(\bar{x}_j) - \frac{1}{w}\sum_{i=1}^{N}\sum_{j=1}^{N} v_i v_j \varphi(\bar{x}_i)^T \sum_{l}^{N} v_l \varphi(\bar{x}_l) \\ &- \frac{1}{w}\sum_{i=1}^{N}\sum_{j=1}^{N} v_i v_j \sum_{l=1}^{N} v_l \varphi(\bar{x}_l)^T\varphi(\bar{x}_j) + \frac{1}{w^2}\sum_{i=1}^{N}\sum_{j=1}^{N} v_i v_j \sum_{l=1}^{N}\sum_{t=1}^{N} v_l v_t \varphi(\bar{x}_l)^T\varphi(\bar{x}_t) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} v_i v_j \varphi(\bar{x}_i)^T\varphi(\bar{x}_j) - \frac{1}{\cancel{w}}\sum_{i=1}^{N}\sum_{l=1}^{N} v_i v_l \varphi(\bar{x}_i)^T\varphi(\bar{x}_l)\overbrace{\sum_{j=1}^{N} v_j}^{\cancel{w}} \\ &- \frac{1}{\cancel{w}}\sum_{j=1}^{N}\sum_{l=1}^{N} v_j v_l \varphi(\bar{x}_l)^T\varphi(\bar{x}_j)\overbrace{\sum_{i=1}^{N} v_i}^{\cancel{w}} + \frac{1}{\cancel{w^2}}\sum_{l}^{N}\sum_{t}^{N} v_l v_t \varphi(\bar{x}_l)^T\varphi(\bar{x}_t)\overbrace{\sum_{i=1}^{N} v_i \sum_{j=1}^{N} v_j}^{\cancel{w^2}} = 0 \end{aligned} \tag{3.185}$$

Therefore, the minimization of the *weight centered weighted kernel* version of the classical K-means objective given by Eq. 3.175 can be written as the following trace maximization problem

$$\begin{aligned} \max_{A_{K-1}} \quad & Tr[A_{K-1}^T V^{\frac{1}{2}} M_v \Omega M_v^T V^{\frac{1}{2}} A_{K-1}] \\ s.t.: \quad & A_{K-1}^T A_{K-1} = I_{K-1}, \quad A_{K-1} \in \mathbb{R}^{NxK-1}, \\ & 0 = \frac{1}{\sqrt{w}} \bar{\alpha}^{(k)T} [\sqrt{v_1}, ..., \sqrt{v_N}]^T = \frac{1}{\sqrt{w}} \sum_{i=1}^{N} \alpha_i^{(k)} \sqrt{v_i}, \\ & \forall \quad k = 1, ..., K-1, \quad K \leq N \end{aligned} \tag{3.186}$$

The Ky Fan theorem [10] provides a global solution of the relaxed version of this problem by allowing A_{K-1} to be any orthonormal matrix that satisfies the above constraints. According to the Ky Fan theorem the optimal solution of this relaxed trace maximization problem is given by the $K-1$ leading eigenvectors of the $V^{\frac{1}{2}} M_v \Omega M_v^T V^{\frac{1}{2}}$ matrix i.e. the $\bar{\alpha}$ eigenvectors correspond to the first $K-1$ highest eigenvalues λ

$$V^{\frac{1}{2}} M_v \Omega M_v^T V^{\frac{1}{2}} \bar{\alpha} = V^{\frac{1}{2}} \Omega_v V^{\frac{1}{2}} \bar{\alpha} = \lambda \bar{\alpha} \tag{3.187}$$

Note, that this eigenproblem is exactly the same as the *weighted kernel* PCA eigenproblem discussed in section 3.2.2. Therefore, *weight centered weighted kernel* K-means can be interpreted as *weighted kernel* PCA. Also note, that the constraints (orthonormality and $\sqrt{v}$ centered property) on the optimal solutions are also the same in both cases, therefore the identity between the eigenproblems involving Ω_v and Ω_v^* discussed in section 3.2.2 also holds for the $k = 1, ..., K-1$ leading eigenvectors.

The non-symmetric version of the *weighted kernel* PCA problem discussed in section 3.2.3 can be derived from the *weight centered weighted kernel* K-means objective given by Eq. 3.175 in a similar way by using the relation between the optimal solutions of the symmetric $\bar{\alpha}$ and non-symmetric $\bar{\beta}$ problems discussed in section 3.2.4. It was shown there, that the key point in the transformation of the properties of $\bar{\alpha}$ vectors (optimal solutions of the symmetric dual problem) to the properties of $\bar{\beta}$ vectors (optimal solutions of the non-symmetric dual problem) is the replacement of the standard Euclidean norm by the V^{-1}-norm defined by Eq. 3.154. Let change the orthonormal

cluster membership indicators given by Eq. 3.177 to

$$\bar{h}_k = [0, ..., 0, \underbrace{\underbrace{v_i, ..., v_j}_{i...j \in C_k}}_{n_k}, 0, ...0]^T / \sqrt{w_k} \tag{3.188}$$

V^{-1}-orhonormal cluster membership indicators. Therefore, the $H_K \in \mathbb{R}^{NxK}_{\geq 0}$ matrix with columns of $\bar{h}_k$ is now a V^{-1}-orhonormal matrix i.e. $H_K^T V^{-1} H_K = I_k$. It should be noted here that these cluster membership indicators are given in a slightly different form than the usual

$$\bar{f}_k = [0, ..., 0, \underbrace{1, ..., 1}_{n_k}, 0, ...0]^T \tag{3.189}$$

$k = 1, ..., K$ mutually orthogonal vectors. One can interpret the vectors given in Eq. 3.188 as the V^{-1}-normalized columns of the VF_K orthogonal matrix where the K columns of F_K are the usual orthogonal cluster indicator vectors. For the shake of simplicity cluster membership indicators are going to be used in the form of Eq. 3.188 hereafter. One of the columns of this H_K matrix i.e. one of the $\bar{h}_k$ vectors can be expressed as the linear combination of the others since

$$\sum_{k=1}^{K} \sqrt{w_k} \bar{h}_k = [v_1, ..., v_N]^T = V \bar{\mathbb{1}}_N \tag{3.190}$$

always holds. This redundancy can be removed similarly to the previous case i.e. by performing a linear transformation T on H_K into $B_k = H_K T$ where $B_K \in \mathbb{R}^{NxK}$ and $T \in \mathbb{R}^{KxK}$ is an orthonormal matrix with last column of $t_K = [\sqrt{w_1}, ..., \sqrt{w_K}]^T / \sqrt{w}$. The result of choosing the last column of T this way is that the last column of B_K i.e. $\bar{\beta}_K$ will be equal to

$$\bar{\beta}_K = \frac{\sqrt{w_1}}{\sqrt{w}} \bar{h}_1 + ... + \frac{\sqrt{w_K}}{\sqrt{w}} \bar{h}_K = \sum_{k=1}^{K} \frac{\sqrt{w_k}}{\sqrt{w}} \bar{h}_k = [v_1, ..., v_N]^T / \sqrt{w} \tag{3.191}$$

Since the columns of H_k are V^{-1}-orthonormal and T is an orthonormal matrix i.e. $H_K^T V^{-1} H_K = I_k$ and $T^T T = I_K$ the columns of B_K are going to be V^{-1}-orthonormal

$$B_K^T V^{-1} B_K = [H_K T]^T V^{-1} [H_K T] = \underbrace{T^T \underbrace{H_K^T V^{-1} H_K}_{I_K} T}_{I_K} = I_K \tag{3.192}$$

that can be written as

$$\bar{\beta}^{T(i)}\bar{\beta}^{(j)} = \delta_{ij}, \quad \forall \quad i,j = 1,...,K \leq N \tag{3.193}$$

Since the last column of B_K i.e. $\bar{\beta}_K$ is always $\bar{\beta}_K = [v_1,...,v_N]^T/\sqrt{w}$ the remaining $K-1$ $\bar{\beta}_k$ vectors must be V^{-1}-orthogonal to this

$$\begin{aligned} 0 =& \bar{\beta}^{(i)T}V^{-1}\bar{\beta}^{(K)} = \frac{1}{\sqrt{w}}\bar{\beta}^{(i)T}\underbrace{V^{-1}[v_1,...,v_N]^T}_{\bar{\mathbb{1}}_N} = \frac{1}{\sqrt{w}}\sum_{j=1}^{N}\beta_j^{(i)}, \\ & \forall \quad i = 1,...,K-1, \quad K \leq N \end{aligned} \tag{3.194}$$

that results in the centering of these $K-1$ $\bar{\beta}_k$ vectors i.e. $\bar{\mathbb{1}}_N^T\bar{\beta}_k = 0$. The second term of the summation in Eq 3.175 can be now written as

$$\begin{aligned} &\sum_{k=1}^{K}\frac{\sum_{i,j\in C_k}v_iv_j[\varphi(\bar{x}_i)-\hat{\mu}^v_{\varphi(\bar{x})}]^T[\varphi(\bar{x}_j)-\hat{\mu}^v_{\varphi(\bar{x})}]}{\sum_{i\in C_k}v_i} = Tr[H_K^TM_v\Omega M_v^TH_K] \\ &= Tr[[B_KT^T]^TM_v\Omega M_v^T[B_KT^T]] = Tr[B_K^TM_v\Omega M_v^TB_K] \\ &= Tr[B_{K-1}^TM_v\Omega M_v^TB_{K-1}] + \frac{1}{w}\underbrace{\sum_{i=1}^{N}\sum_{j=1}^{N}v_iv_j[\Omega_v]_{ij}}_{0} \end{aligned} \tag{3.195}$$

by using the V^{-1}-orthonormal columns of H_K i.e. the cluster membership indicators and $B_K = H_KT$ V^{-1}-orthonormal matrix. The last equation was derived by making use of the last $(K-th)$ column of B_K matrix given by Eq. 3.191 and the weight centered property of Ω_v matrix. Therefore, the minimization of the *weight centered weighted kernel* version of the classical K-means objective given by Eq. 3.175 can be now written as the following trace maximization problem

$$\begin{aligned} &\max_{B_{K-1}} \quad Tr[B_{K-1}^TM_v\Omega M_v^TB_{K-1}] \\ &s.t.: \quad B_{K-1}^TV^{-1}B_{K-1} = I_{K-1}, \quad B_{K-1} \in \mathbb{R}^{NxK-1}, \\ &\qquad 0 = \bar{\mathbb{1}}_N^T\bar{\beta}_k, \ \forall \quad k = 1,...,K-1, \quad K \leq N \end{aligned} \tag{3.196}$$

Since both $M_v\Omega M_v^T$ and V^{-1} are symmetric and V^{-1} is positive definite the the extended Ky Fan theorem [11] provides a global solution of the relaxed version of this trace maximization problem by allowing B_{K-1} to be

any orthonormal matrix that satisfies the above constraints. It says that the optimal solution of this relaxed trace maximization problem is given by the $K-1$ leading eigenvectors of the following eigenvalue problem

$$M_v \Omega M_v^T \bar{\beta} = \Omega_v \bar{\beta} = \lambda V^{-1} \bar{\beta} \tag{3.197}$$

that can be written as

$$V \Omega_v \bar{\beta} = \lambda \bar{\beta} \tag{3.198}$$

since V is a diagonal matrix. Note, that this eigenproblem is exactly the same as the non-symmetric *weighted kernel* PCA eigenproblem discussed in section 3.2.3. Also note, that the constraints (V^{-1}-orthonormality and centered property) on the optimal solutions are also the same in both cases, therefore the identity between the eigenproblems involving Ω_v and Ω_v^* discussed in section 3.2.3 also holds for the $k = 1, ..., K-1$ eigenvectors.

It was shown in sections 3.3.1 and 3.3.2 that a link can be established between the classical kernel spectral clustering algorithms based on the normalized graph Laplacian and the different versions of weighted kernel PCA by choosing the diagonal weight matrix V being equal to the inverse degree matrix D^{-1} of the similarity graph. In doing so the current result obtained for the symmetric problem is very similar to that obtained by *Dhillon et al.* [9] concerning the link between weighted kernel K-means and normalized cut kernel spectral clustering with the difference that in our case the weight centered kernel matrix will play the role of the kernel matrix. Furthermore, by choosing the diagonal weight matrix being equal to the identity matrix the weight centered kernel K-means objective given by Eq. 3.175 transforms to a (unweighted) centered version of the kernel K-means and both formulations of the weighted kernel PCA transform to the (unweighted) kernel PCA problem discussed in section 3.2.1 as it was shown in section 3.2.4. In this case the current result for the symmetric problem (and formulation) is identical with that obtained by *Ding et al.* [8] concerning the relation of (kernel) K-means and (kernel) PCA. Therfore, a connection can be made between weighted kernel K-means, weighted kernel PCA and normalized cut spectral clustering algorithms as well as between kernel K-means and kernel PCA by appropriately choosing the weight matrix.

3.5 Kernel spectral clustering through weighted kernel PCA

The close relation of weighted kernel PCA to the basic kernel spectral clustering algorithms was shown in the previous section. The complete kernel

spectral clustering algorithm applicable for multile clusters through weighted kernel PCA will be given in this section. A straightforward *out-of-sample extension* offerd by the formulation as well as a clustering quality measure or model selection criteria, the *balanced line fit* (BLF) will also be presented. The content of this section is mainly based on [13].

3.5.1 Multiway kernel spectral clustering as weighted kernel PCA

Two slightly different formulations of weighted kernel PCA was given in sections 3.2.2 and 3.2.3 using different weightening and their relation was discussed in section 3.2.4. The optimal solution of the dual weighted kernel PCA problem was shown to be the $k = 1, ..., K,\ K \leq N$ eigenvectors of the corresponding matrices. Weighted kernel PCA was formulated both by the explicit weighted centering of the feature map and by using a bias term in all cases. Furthermore, it was shown that finally both approaches lead to the same results but this simplest form of the solution can be directly obtained by using the bias term approach. The multiway version of weighted kernel PCA will be formulated here by using this bias term approach and a weightening applied in section 3.2.3 using matrix notations. The multiway version of results obtained in section 3.2.2 can be formulated similarly or can be derived from the current results by using the relation between the two different form of solution discussed in section 3.2.4.

Let the non-linear mapping of the d dimensional input data to a high n_h dimenzional feature space $\varphi(\cdot) : \mathbb{R}^d \to \mathbb{R}^{n_h}$. The $k-th$ score variable that corresponds to the $l-th$ input data can be written as

$$z_l^{(k)} = z^{(k)}(\bar{x}_l) = \bar{w}^{(k)^T}\varphi(\bar{x}_l) + b^{(k)} \tag{3.199}$$

and the weighted kernel PCA maximization problem with $V \in \mathbb{R}^+$ diagonal weight matrix is

$$\max_{\bar{w}^{(1)},...,\bar{w}^{(K)}} \sum_{k=1}^{K}\sum_{l=1}^{N} v_l[0 - (\bar{w}^{(k)^T}\varphi(\bar{x}_l) + b^{(k)})]^2 \tag{3.200}$$

similarly to section 3.2.3 but now we are looking for the $K \leq N$ directions that maximize the variance of the $K \leq N$ projected variables where N is the number of input data points. This maximization problem can be written as

$$\max_{\bar{w}^{(1)},...,\bar{w}^{(K)}} \sum_{k=1}^{K} \bar{e}^{(k)^T} V \bar{e}^{(k)} \tag{3.201}$$

by introducing $\Phi \in \mathbb{R}^{Nxn_h}$ feature map matrix with elements of

$$\Phi = \begin{bmatrix} \varphi^T(\bar{x}_1) \\ \vdots \\ \varphi^T(\bar{x}_N) \end{bmatrix} \tag{3.202}$$

and $\bar{e}^{(k)} \in \mathbb{R}^{Nx1}$ (error) vectors

$$\begin{aligned} \bar{e}^{(k)} &= \Phi \bar{w}^{(k)} + b^{(k)} \\ \bar{e}^{(k)} &= \begin{bmatrix} e_1^{(k)} \\ \vdots \\ e_N^{(k)} \end{bmatrix} = \begin{bmatrix} \bar{w}^{(k)^T} \varphi(\bar{x}_1) + b^{(k)} \\ \vdots \\ \bar{w}^{(k)^T} \varphi(\bar{x}_N) + b^{(k)} \end{bmatrix} \end{aligned} \tag{3.203}$$

The corresponding primal optimization problem can be formulated as

$$\boxed{\text{P:}} \quad \begin{aligned} \max_{\bar{w}^{(1)},...,\bar{w}^{(K)}} J(\bar{w}, \bar{e}, b) &= \frac{1}{2} \sum_{k=1}^{K} \gamma^{(k)} \bar{e}^{(k)^T} V \bar{e}^{(k)} - \frac{1}{2} \sum_{k=1}^{K} \bar{w}^{(k)^T} \bar{w}^{(k)} \\ s.t. \quad \bar{e}^{(k)} &= \Phi \bar{w}^{(k)} + b^{(k)}, \quad k = 1, ..., K \end{aligned} \tag{3.204}$$

The Lagragian of this constrained optimization problem is

$$\begin{aligned} \mathcal{L}(\bar{w}^{(k)}, \bar{e}^{(k)}, b^{(k)}; \bar{\beta}^{(k)}) = & \frac{1}{2} \sum_{k=1}^{K} \gamma^{(k)} \bar{e}^{(k)^T} V \bar{e}^{(k)} - \frac{1}{2} \sum_{k=1}^{K} \bar{w}^{(k)^T} \bar{w}^{(k)} \\ & - \sum_{k=1}^{K} \bar{\beta}^{(k)^T} (\bar{e}^{(k)} - \Phi \bar{w}^{(k)} - b^{(k)} \bar{\mathbb{1}}_N) \end{aligned} \tag{3.205}$$

with Lagrange multiplier vectors $\bar{\beta}^{(k)} \in \mathbb{R}^{Nx1}$ and $k = 1, ..., K$. The KKT conditions for optimality give

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \bar{w}^{(k)}} = -\bar{w}^{(k)} + \Phi^T \bar{\beta}^{(k)} = 0 & \rightarrow \quad \bar{w}^{(k)} = \Phi^T \bar{\beta}^{(k)} \\ \dfrac{\partial \mathcal{L}}{\partial \bar{e}^{(k)}} = \dfrac{1}{2} 2\gamma^{(k)} V \bar{e}^{(k)} - \bar{\beta}^{(k)} = 0 & \rightarrow \quad \bar{\beta}^{(k)} = \gamma^{(k)} V \bar{e}^{(k)} \\ \dfrac{\partial \mathcal{L}}{\partial b^{(k)}} = \bar{\mathbb{1}}_N^T \bar{\beta}^{(k)} = 0 & \rightarrow \quad \bar{\mathbb{1}}_N^T \bar{\beta}^{(k)} = 0 \\ \dfrac{\partial \mathcal{L}}{\partial \bar{\beta}^{(k)}} = -\bar{e}^{(k)} + \Phi \bar{w}^{(k)} + b^{(k)} \bar{\mathbb{1}}_N = 0 & \rightarrow \quad \bar{e}^{(k)} - \Phi \bar{w}^{(k)} - b^{(k)} \bar{\mathbb{1}}_N = 0 \end{cases} \tag{3.206}$$

for $k = 1, ..., K$. Eliminating the primal variables $\bar{w}^{(k)}$ and $\bar{e}^{(k)}$ by substituting the first and second ($\bar{e}^{(k)} = \frac{V^{-1}\bar{\beta}^{(k)}}{\gamma^{(k)}}$) equations into the fourth yields

$$\frac{V^{-1}\bar{\beta}^{(k)}}{\gamma^{(k)}} - \Phi\Phi^T \bar{\beta}^{(k)} - \bar{\mathbb{1}}_N b^{(k)} = 0, \quad k = 1, ..., K \tag{3.207}$$

Applying the third KKT condition on this gives the expression for the last primal variable $b^{(k)}$

$$\begin{aligned} \bar{\mathbb{1}}_N^T \bar{\beta}^{(k)} &= \gamma^{(k)} \bar{\mathbb{1}}_N^T V \Phi\Phi^T \bar{\beta}^{(k)} + \gamma^{(k)} \bar{\mathbb{1}}_N^T V \bar{\mathbb{1}}_N b^{(k)} = 0 \\ \bar{\mathbb{1}}_N^T V \bar{\mathbb{1}}_N b^{(k)} &= -\bar{\mathbb{1}}_N^T V \Phi\Phi^T \bar{\beta}^{(k)} \\ b^{(k)} &= -\frac{\bar{\mathbb{1}}_N^T V \Phi\Phi^T \bar{\beta}^{(k)}}{\bar{\mathbb{1}}_N^T V \bar{\mathbb{1}}_N} \end{aligned} \tag{3.208}$$

for $k = 1, ..., K$. Substituting this back into Eq. 3.207 gives the following equations for the optimality expressed in terms of dual variables $\bar{\beta}^{(k)}$

$$\begin{aligned} &\frac{V^{-1}\bar{\beta}^{(k)}}{\gamma^{(k)}} - \Phi\Phi^T \bar{\beta}^{(k)} + \frac{\bar{\mathbb{1}}_N \bar{\mathbb{1}}_N^T V \Phi\Phi^T \bar{\beta}^{(k)}}{\bar{\mathbb{1}}_N^T V \bar{\mathbb{1}}_N} = 0, \quad k = 1, ..., K \\ &\frac{V^{-1}\bar{\beta}^{(k)}}{\gamma^{(k)}} - \left[I_N - \frac{\bar{\mathbb{1}}_N \bar{\mathbb{1}}_N^T V}{\bar{\mathbb{1}}_N^T V \bar{\mathbb{1}}_N} \right] \Phi\Phi^T \bar{\beta}^{(k)} = 0, \quad k = 1, ..., K \end{aligned} \tag{3.209}$$

Multiplying both sides by the diagonal weight matrix V from the left and substituting $\lambda^{(k)} = 1/\gamma^{(k)}$ the optimal solution can be expressed in terms of

dual variables $\bar{\beta}^{(k)}$ as the following eigenvalue equations involving the dual variables as eigenvectors

$$VM_v\Omega\bar{\beta}^{(k)} = \lambda^{(k)}\bar{\beta}^{(k)}, \quad k = 1, ..., K \tag{3.210}$$

Note that the expression in the square brackets in Eq. 3.209 is the M_v weighted centering matrix by definition (see Eq. 3.63) and $\Omega = \Phi\Phi^T$ where $\Omega_{ij} = \varphi(\bar{x}_i)^T\varphi(\bar{x}_j)$ is the kernel matrix. According to Eq. 3.210 the pool of candidate solutions is formed by the eigenvectors of the $VM_v\Omega$ matrix. The objective of the maximization problem given by Eq. 3.201 can be now written as

$$\sum_{k=1}^{K}\bar{e}^{(k)^T}V\bar{e}^{(k)} = \sum_{k=1}^{K}\frac{\bar{\beta}^{(k)^T}V^{-1}}{\gamma^{(k)}}V\frac{V^{-1}\bar{\beta}^{(k)}}{\gamma^{(k)}} = \sum_{k=1}^{K}\frac{1}{\gamma^{(k)^2}}\bar{\beta}^{(k)^T}V^{-1}\bar{\beta}^{(k)} = \sum_{k=1}^{K}\lambda^{(k)^2} \tag{3.211}$$

using $\bar{\beta}^{(k)^T}V^{-1}\bar{\beta}^{(k)} = 1$ that is the V^{-1}-normality of the $\bar{\beta}^{(k)}$ eigenvectors of the $VM_v\Omega$ matrix discussed in section 3.2.4. Therefore, the dual solution of the weighted kernel PCA maximization problem can be given by the K eigenvectors of the $VM_v\Omega$ matrix correspond to the K highest eigenvalues. This point offers a nice way to reformulate the optimal solution of the weighted kernel PCA as a trace maximization resulting in a form that is equivalent with the weighted kernel K-means trace maximization problem obtained in the previous section.

It was shown in the previous section (3.4) that minimizing the weighted kernel K-means objective in case of weight centered feature map (weighted centering of the kernel matrix) is equivalent with the weighted kernel PCA maximization problem. This result was derived there by reformulating the weighted kernel K-means minimization problem as a trace maximization and the optimal solution of that trace maximization was shown to be the leading $K-1$ eigenvectors (in case of K true clusters) of exactly the same eigenproblem as the dual weighted kernel PCA eigenproblem. The optimal solution of the weighted kernel PCA in the dual space can be now given as a trace maximization by using Eqs. 3.210 and 3.211. The third KKT condition for optimality in Eq. 3.206 says that the optimal solutions of the dual problem i.e. the $\bar{\beta}^{(k)}$ eigenvectors of the $VM_v\Omega$ matrix are centered. It was shown in section 3.2.3 that Eq. 3.210 can also be written as

$$VM_v\Omega\bar{\beta}^{(k)} = VM_v\Omega M_v^T\bar{\beta}^{(k)} = \lambda^{(k)}\bar{\beta}^{(k)} \tag{3.212}$$

in this case since the extra terms introduced in the weighted centering of the kernel matrix by M_v^T will give zero contributions. Suppose that the $K \leq N$

directions are looked for in the weighted kerenel PCA problem that maximize the variance of the projected variables. According to Eqs. 3.210- 3.212 the optimal solution of this maximization problem is given by the K eigenvectors of the $VM_v\Omega M_v^T$ matrix corresponds to the K highest eigenvalues. Let introduce $B_k \in \mathbb{R}^{NxK}$ matrix and $\Lambda_K \in \mathbb{R}^{KxK}$ diagonal matrix where columns of B_K are formed by distinct eigenvectors of the $VM_v\Omega M_v^T$ matrix and the corresponding eigenvalues are the diagonal elements of Λ_K. Then Eq. 3.212 can be written with K candidate solutions as

$$VM_v\Omega M_v^T B_K = B_K\Lambda_K \tag{3.213}$$

Multiplying both sides of this by $B_K^T V^{-1}$ from the left results in

$$\begin{aligned} B_K^T \underbrace{V^{-1}V}_{I_N} M_v\Omega M_v^T B_K &= \underbrace{B_K^T V^{-1} B_K}_{I_K}\Lambda_K \\ B_K^T M_v\Omega M_v^T B_K &= \Lambda_K \end{aligned} \tag{3.214}$$

where the V^{-1}-orthonormality of the $\bar{\beta}^{(k)}$ eigenvectors discussed in section 3.2.4 was used. According to Eq. 3.211 the optimal solution for the weighted kernel PCA objective maximization can be written as

$$\begin{aligned} max \quad & Tr[\Lambda_K] = Tr[B_K^T M_v\Omega M_v^T B_K] \\ s.t. \quad & B_K^T V^{-1} B_K = I_K \\ & \bar{\mathbb{1}}_N^T \bar{\beta}^{(k)} = 0, \quad \forall\, k = 1, ..., K \end{aligned} \tag{3.215}$$

Note that this is exactly the same trace maximization problem as obtained in section 3.4 by reformulating the minimization of the weighted kerenel K-means objective as a trace maximization problem (see Eq. 3.196). Therefore, the optimal solution of that weighted kerenel K-means trace maximization problem is identical with the optimal solution of the weighted kernel PCA problem. The other formulation of the weighted kerenel K-means trace maximization given by Eq. 3.186 can be derived similarly by first transforming the eigenproblem Eq. 3.212 as shown by Eqs. 3.150 in section 3.2.4 then using the same steps as above but changing from V^{-1}-norm to the Euclidean norm.

Returning back to the multiway formulation of the weighted kernel PCA problem, the $k-th$ score variable that corresponds to the $l-th$ input data

can be written as

$$\begin{aligned} z_l^{(k)} =& z^{(k)}(\bar{x}_l) = \bar{w}^{(k)^T}\varphi(\bar{x}_l) + b^{(k)} = \bar{\beta}^{(k)^T}\Phi\varphi(\bar{x}_l) + b^{(k)} = \\ & \sum_{i=1}^{N}\beta_i^{(k)}\varphi^T(\bar{x}_i)\varphi(\bar{x}_l) + b^{(k)} = \sum_{i=1}^{N}\beta_i^{(k)}K(\bar{x}_i,\bar{x}_l) + b^{(k)} \end{aligned} \tag{3.216}$$

by using Eq. 3.199, the first KKT condition for optimality given by Eqs. 3.206 and the kernel trick. The $k-th$ score variable $\bar{z}^{(k)} \in \mathbb{R}^{Nx1}$ can be written as

$$\begin{aligned} \bar{z}^{(k)} =& \Phi\bar{w}^{(k)} + \bar{\mathbb{1}}_N b^{(k)} = \Phi\Phi^T\bar{\beta}^{(k)} - \frac{\bar{\mathbb{1}}_N\bar{\mathbb{1}}_N^T V\Phi\Phi^T\bar{\beta}^{(k)}}{\bar{\mathbb{1}}_N^T V\bar{\mathbb{1}}_N} = \\ & \left[I_N - \frac{\bar{\mathbb{1}}_N\bar{\mathbb{1}}_N^T V}{\bar{\mathbb{1}}_N^T V\bar{\mathbb{1}}_N}\right]\Phi\Phi^T\bar{\beta}^{(k)} = M_v\Omega\bar{\beta}^{(k)} = \lambda^{(k)}V^{-1}\bar{\beta}^{(k)} \end{aligned} \tag{3.217}$$

that shows two important properties of the score variables: (i) every entry in the $k-th$ score variable has the same sign as the corresponding entry in the $k-th$ eigenvector (since $\lambda^{(k)}$ and the weights are pozitive); (ii) the score variables $\bar{z}^{(k)}$ are orthogonal to the $[v_1,...,v_N]^T$ vector i.e. the weighted mean is removed from the score variables

$$\sum_{i=1}^{N} z_i^{(k)}v_i = \bar{z}^{(k)^T}[v_1,...,v_N]^T = \lambda^{(k)}[V^{-1}\bar{\beta}^{(k)}]^T[v_1,...,v_N]^T = \lambda^{(k)}\underbrace{\bar{\beta}^{(k)^T}\bar{\mathbb{1}}_N}_{0} = 0 \tag{3.218}$$

where the third KKT condition for optimality was used from Eqs. 3.206.

Piecewise constant eigenvectors

The relation of weighted kernel PCA to the kernel spectral clustering based on the random walk normalized graph Laplacian was discussed in section 3.3.2. A close connection was shown between these algorithms when the diagonal weight matrix is chosen to be the inverse degree matrix of the graph i.e. $V = D^{-1}$, $diag\ D \in \mathbb{R}^{NxN}$ $[D]_{ii} = d_i = \sum_{j=1}^{N}[\Omega]_{ij}$. The kernel spectral clustering eigenvalue problem is

$$D^{-1}\Omega\bar{\alpha} = \lambda\bar{\alpha} \tag{3.219}$$

and the corresponding weighted kernel PCA eigenvalue problem

$$D^{-1}M_D\Omega\bar{\beta} = \lambda\bar{\beta} \tag{3.220}$$

as given by Eq. 3.210 above. M_D is the weighted centering matrix as defined by Eq. 3.63 and substituting $V = D^{-1}$.

Let assume that there are K well formed clusters in the N input data and the corresponding partition of these $i = 1, ..., N$ data points is $\Delta = \{C_1, ..., C_K\}$. A vector $\bar{x}$ is *piecewise constant* relative to a partition Δ of the $i = 1, ..., N$ datapoints if and only if $x_i = x_j,\ \forall i, j \in C_k,\ k = 1, ..., K$ i.e. if the $i, j - th$ points belong to the same set [12]. The piecewise constant property (with respect to the partition Δ) of the K eigenvectors of Eq 3.219 that belong to the eigenvalue of one was discussed in [12]. It was stated that these eigenvectors of the $D^{-1}\Omega$ matrix are piecewise constant with respect to the partition $\Delta = \{C_1, ..., C_K\}$ when the corresponding K clusters are well formed ($[\Omega]_{ij} \neq 0 \iff i, j \in C_k\ \forall k = 1, ..., K$) due to the *block-stochastic* property of $D^{-1}\Omega$. A matrix P indexed by $i = 1, ..., N$ is block-stochastic with respect to a partition $\Delta = \{C_1, ..., C_K\}$ of the indices if $P_{ik} = \sum_{j \in C_k} P_{ij}$ are constant $\forall\ i \in C_{k'},\ \forall k, k' = 1, ..., K$ [12] and the $R \in \mathbb{R}^{KxK}$ matrix with $P_{k'k}$ entries are non-singular.

The $D^{-1}\Omega$ matrix is a row stochastic matrix i.e. $\sum_{j=1}^{N} [\Omega]_{ij} = 1,\ \forall i = 1, ..., N$ because $diag\ D \in \mathbb{R}^{NxN}$ $[D]_{ii} = d_i = \sum_{j=1}^{N} [\Omega]_{ij}$. Furthermore, if the K clusters are well formed ($[\Omega]_{ij} \neq 0 \iff i, j \in C_k\ \forall k = 1, ..., K$) then the i, j entries of $D^{-1}\Omega$ are different than zero only if the $i, j - th$ points belong to the same $k - th$ cluster i.e. $[D^{-1}\Omega]_{ij} \neq 0 \iff i, j \in C_k\ \forall k = 1, ..., K$. Therefore, the $D^{-1}\Omega$ matrix is block-stochastic since $\forall i \in C_k$

$$\sum_{j=1}^{N} [D^{-1}\Omega]_{ij} = \sum_{j \in C_{k'}} [D^{-1}\Omega]_{ij} = \begin{cases} 1 & if\ k = k' \\ 0 & if\ k \neq k' \end{cases} \quad \forall k, k' = 1, ..., K \qquad (3.221)$$

Thus, the piecewise constant vectors (relative to the Δ partition) are eigenvectors of the $D^{-1}\Omega$ matrix with eigenvalue of one. If these eigenvectors are formed by the $\xi_k,\ k = 1, ..., K$ constants the structure of $D^{-1}\Omega$ matrix expressed by Eq. 3.221 results in K independent conditions for the K ξ_k ($\xi_k = \xi_k$). Therefore, the dimension of the space of these piecewise constant eigenvectors (with eigenvalue of one) is K i.e. there are K orthonormal eigenvectors with eigenvalue of one that are piecewise constant with respect to the partition (perfect clustering). More precisely, the eigenspace of $D^{-1}\Omega$ associated to $\lambda = 1$ eigenvalue is a K-dimensional space and spanned by the K orthogonal eigenvectors corresponds to $\lambda = 1$. Since these orthogonal eigenvectors are piecewise constants on the clusters they serves as cluster indicatorts to the respective disconnected components. Thus the clustering is trivial due to the fact that each cluster is represented by these K orthonormal eigenvectors as a single poin in $\mathbb{R}^K$.

Although the $D^{-1}M_D\Omega$ matrix is not block-stochastic its eigenvectors correspond to eigenvalue of one are piecewise constants with respect to the well formed clusters. According to Eq. 3.104, multiplying Ω by M_v has the effect that removes from each columns of Ω the weighted mean of the corresponding column and the $i, j - th$ elements of $VM_v\Omega$ matrix can be written as

$$\begin{aligned}
[VM_v\Omega]_{ij} &= v_i[\varphi(\bar{x}_i) - \hat{\mu}^v_{\varphi(\bar{x})}]^T\varphi(\bar{x}_j) \\
&= v_i[\varphi(\bar{x}_i) - \frac{\sum_{l=1}^N v_l\varphi(\bar{x}_l)}{\sum_{l=1}^N v_l}]^T\varphi(\bar{x}_j) \\
&= v_i\varphi(\bar{x}_i)^T\varphi(\bar{x}_j) - v_i\frac{\sum_{l=1}^N v_l\varphi(\bar{x}_l)^T\varphi(\bar{x}_j)}{\sum_{l=1}^N v_l}
\end{aligned} \tag{3.222}$$

The sum of the $i - th$ row is equal to

$$\sum_{j=0}^N [VM_v\Omega]_{ij} = v_i\sum_{j=0}^N \varphi(\bar{x}_i)^T\varphi(\bar{x}_j) - v_i\sum_{j=0}^N \frac{\sum_{l=1}^N v_l\varphi(\bar{x}_l)^T\varphi(\bar{x}_j)}{\sum_{l=1}^N v_l} \tag{3.223}$$

that becomes

$$\sum_{j=0}^N [D^{-1}M_D\Omega]_{ij} = \frac{\sum_{j=0}^N [\Omega]_{ij}}{d_i} - \frac{1}{d_i\epsilon}\sum_{l=1}^N \frac{\sum_{j=0}^N [\Omega]_{lj}}{d_l} \tag{3.224}$$

by substituting $V = D^{-1}$, $diag\ D \in \mathbb{R}^{NxN}$, $[D]_{ii} = d_i = \sum_{j=1}^N [\Omega]_{ij}$, $\epsilon = \sum_{i=1}^N 1/d_i$ and $[\Omega]_{ij} = \varphi(\bar{x}_i)^T\varphi(\bar{x}_j)$. Let the $i - th$ data point belong to the $k - th$ cluster ($i \in C_k$). Dividing up the summation of the $i - th$ row according to the (perfect) particion $\Delta = \{C_1, ..., C_K\}$ of the $i, j = 1, ..., N$ data points and making use of the well formed property of the clusters i.e. $[\Omega]_{ij} \neq 0 \iff i, j \in C_k\ \forall k = 1, ..., K$ one can get

$$\sum_{j=0}^N [D^{-1}M_D\Omega]_{ij} = \underbrace{1 - \frac{n_{k'}}{d_i\epsilon}}_{=\sum_{j\in C_{k'},\ k=k'}} - \underbrace{\sum_{k'=1,k'\neq k}^K \frac{n_{k'}}{d_i\epsilon}}_{\sum_{j\in C_{k'},\ k\neq k'}} \tag{3.225}$$

It shows that unlike the $D^{-1}\Omega$ the $D^{-1}M_D\Omega$ does *not* fulfil the block-stochastic property since d_i, $i = 1, ..., N$ are different for the different rows. However, this result also shows that the vectors being piecewise constant ξ_k, $k = 1, ..., K$ with respect to the well formed particion $\Delta = \{C_1, ..., C_K\}$ are eigenvectors of the $D^{-1}M_D\Omega$ matrix with eigenvalue of one if and only if the ξ_k

constants satisfy the

$$n_1\xi_1 + ... + n_k\xi_k + ... + n_K\xi_K = \sum_{k=1}^{K} n_k\xi_k = 0 \tag{3.226}$$

condition. According to Eq. 3.226, $K-1$ out of these K ξ_k constants can be chosen independently. Therefore, the dimension of the space of these piecewise constant eigenvectors (with eigenvalue of one) is $K-1$ i.e. there are $K-1$ orthonormal eigenvectors with eigenvalue of one that are piecewise constant with respect to the partition (perfect clustering) and satisfy Eq. 3.226. More precisely, the eigenspace of $D^{-1}M_D\Omega$ associated to $\lambda = 1$ eigenvalue is a $K-1$-dimensional space and spanned by the $K-1$ orthogonal eigenvectors corresponds to $\lambda = 1$. Note, that the condition Eq. 3.226 ensures that these $K-1$ orthogonal eigenvectors are going to be centered. Therefore, the cluster membership information is coded in the signs of the corresponding components of these $K-1$ piecewise constant eigenvectors and each cluster is represented as a single point in different orthant in $\mathbb{R}^{(K-1)}$ (in the space spanned by the $K-1$ eigenvectors with eigenvalue of one).

Chapter 4

Sparse kernel spectral clustering algorithms

This schapter will be on sparse ksc algorithms: why we need sparse algorithms mentioning the two ways (sparse iterative eigensolvers vs out of semple extension)?,THEN FIRTS the random training point selection algorithms: basic Nyström, one of the improved Nyström (fast and memory friendly), then the basic wkpca algorithm for multiple clusters including the quality measures (BLF). AFTER the clever selection of the training points i.e. Rényi entropy maxim. and Cholesky and the corresponding sparse wkpca algorithms,Bla...bla...blaa.....

4.1 First section

Chapter 5

Kernel spectral clustering on climate data

This schapter will be on the application,Bla...bla...blaa.....

5.1 First section

Bibliography

[1] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, A support vector machine formulation to PCA analysis and its kernel version. IEEE Transaction on Neural Networks, 11(2), (2003), 447-450.

[2] I.T. Jolliffe, Principal Component Analysis, 2nd Edn., New York: Springer-Verlag (2002)

[3] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, Least Squares Support Vector Machines, World Scientific, Singapore, (2002) (ISBN 981-238-151-1)

[4] J.A.K. Suykens, C. Alzate,K. Pelckmans, Primal and dual model representations in kernel-based learning, Statist. Surv. Volume 4 (2010), 148-183.

[5] B. Schölkopf, A. Smola, K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural computation, 10(5), (1998), 1299–1319

[6] B. Schölkopf, S. Mika, C.J.C. Burges, P. Knirsch, K.-R. Müller, G. Rätsch, A. Smola, Input space versus feature space in kernel-based methods, IEEE Transactions on Neural Networks, 10(5), (1999), 1000-1017

[7] Y. Saad, Numerical methods for large eigenvalue problems, 2nd Edn., Manchester University Press, (2011) (ISBN 978-1-611970-72-2)

[8] C. Ding, X. He, K-means clustering via principal component analysis, In Proceedings of the twenty-first international conference on Machine learning, Banff, Alberta, Canada, (2004), ACM. 225–232

[9] I.S. Dhillon, Y. Guan, B. Kulis, Kernel k-means: spectral clustering and normalized cuts. KDD. (2004), 551–556

[10] K. Fan, On a theorem of Weyl concerning the eigenvalues of linear transformations. Proceedings of the National Academy of the Sciences of U.S. A 35, (1949), 652-655.

[11] A. H. Sameh, J. A. Wisniewski, A trace minimization algorithm for the generalized eigenvalue problem, SIAM J. Numer. Anal. 19, (1982), 1243-1259

[12] M. Meila, J. Shi, A random walks view of spectral segmentation. In Proceedings AISTATAS. (2001)

[13] C. Alzate, J.A.K. Suykens, Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA, IEEE Transactions on Pattern Analysis and Machine Intelligence 32(2), (2010), 335-347