REGULAR SIMPLEX CRITERION: A NOVEL FEATURE EXTRACTION CRITERION

Quanquan Gu and Jie Zhou

State Key Laboratory on Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology(TNList)
Department of Automation, Tsinghua University, Beijing 100084, China
gqq03@mails.tsinghua.edu.cn,jzhou@tsinghua.edu.cn

ABSTRACT

Feature extraction is an important topic in machine learning. There are two representative criterions for feature extraction, i.e. Fisher Criterion and Maximum Margin Criterion. In this paper, we propose a new criterion, called Regular Simplex Criterion. This criterion requires that samples from the same class are projected to the same point, while samples from different classes have unit distance. Under this criterion, we present a novel dimensionality reduction method, namely Linear Simplex Analysis (LSA). LSA is solved by multivariate linear regression with a specific definition of class indicator matrix which has a strong geometrical interpretation, i.e. each column of this matrix corresponds to a vertex of a regular simplex. Several variants of LSA, e.g. Regularized Simplex Analysis (RSA) and Kernel Simplex Analysis (KSA), are also proposed. Encouraging experimental results on UCI machine learning database indicate that the new criterion as well as the proposed methods are very effective.

Index Terms— Regular Simplex Criterion, Feature Extraction

1. INTRODUCTION

High-dimensional data in the input space is usually not good in practical application due to the *curse of dimensionality*. A common way that attempts to resolve this problem is to use dimensionality reduction, which is an important topic in machine learning. Dimensionality reduction techniques include two types: (1) feature selection: to select a subset of most representative features from the original feature set, and (2) feature extraction: to transform the original feature space to a smaller feature space. Compared with feature selection, feature extraction can not only reduce the dimensionality of the feature space, but also exploit the intrinsic subspace of the original feature space. Feature extraction can be categorized into two types:

1. Unsupervised feature extraction: the most popular unsupervised feature extraction method is Principal Component Analysis (PCA). It aims to find a subspace in which the variance of the projected data is maximum. Since unsupervised feature extraction methods do not take into account the class information, the features extracted is not very suitable for classification.

2. Supervised feature extraction: there are two representative criterions for supervised feature extraction, i.e. *Fisher Criterion* [1] and *Maximum Margin Criterion* [2]. Linear Discriminant Analysis (LDA) [1] is based on Fisher Criterion which aims to maximize the between class variance and minimize the within class variance. Maximum Margin Criterion [2] aims to find a subspace in which a sample is close to those in the same class but far from those in different classes.

In this paper, we propose a new criterion for supervised feature extraction, called *Regular Simplex Criterion*. This criterion requires that samples from the same class are projected to the same point, while samples from different classes have unit distance. Under this criterion, we present a novel dimensionality reduction method, namely Linear Simplex Analysis (LSA). LSA is solved by multivariate linear regression with a specific definition of class indicator matrix which has a strong geometrical interpretation, i.e. each column of this matrix corresponds to a vertex of a c regular simplex, where c is the number of classes. Several variants of LSA, e.g. Regularized Simplex Analysis (RSA) and Kernel Simplex Analysis (KSA), are also proposed. Encouraging experimental results on UCI machine learning database indicate that the new criterion as well as the proposed methods are very effective.

The remainder of this paper is organized as follows. In Section 2, we will review some criterions closely related to ours. In Section 3, we will propose linear simplex analysis. In Section 4, regularized simplex analysis is presented, and in Section 5, kernel simplex analysis is proposed. The experiments on UCI machine learning database are demonstrated in Section 6. Finally, we draw a conclusion in Section 7.

2. RELATED WORK

In this section, we will briefly review Fisher Criterion and MMC Criterion mostly related with ours.

Let $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$ be the data set, where $\mathbf{x}_i$ is a d dimensional column vector, and y_i is the label of $\mathbf{x}_i$. $\mathcal{L} = \{1, 2, \ldots, c\}$ is the label set. LDA [1] aims to find a projection matrix $\mathbf{W} \in \mathbb{R}^{d\times m}$ by the Fisher Criterion

$$\max \mathrm{tr}((\mathbf{W}^T\mathbf{S}_w\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{S}_b\mathbf{W})), \tag{1}$$

where $\mathbf{S}_b = \sum_{i=1}^{c} n_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T$ is called between-class scatter matrix, $\mathbf{m}_i$ and n_i are mean vector and size of class i respectively, $\mathbf{m} = \sum_{i=1}^{c} n_i\mathbf{m}_i$ is the overall mean vector, $\mathbf{S}_w = \sum_{i=1}^{c} \mathbf{S}_i$ is the with-in class scatter matrix, $\mathbf{S}_i$ is the covariance matrix of class i.When the size of samples is small, $\mathbf{S}_w$ is singular and Eq.(1) cannot be solved stably. To address this problem, Regularized Discriminant Analysis (RDA) [3] was proposed

$$\max \mathrm{tr}((\mathbf{W}^T(\mathbf{S}_w + \lambda\mathbf{I})\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{S}_b\mathbf{W})), \tag{2}$$

where λ is a small positive regularizer and $\mathbf{I}$ is identity matrix of proper size.

This work was supported by Natural Science Foundation of China under grant 60673106 and 60573062.

Maximum Margin Criterion [2] aims to find a projection matrix $\mathbf{W} \in \mathbb{R}^{d\times m}$ by

$$\max \text{tr}(\mathbf{W}^T(\mathbf{S}_b - \mathbf{S}_w)\mathbf{W}), \tag{3}$$

where $\text{tr}(\cdot)$ denotes the matrix trace and $\mathbf{W}^T\mathbf{W} = \mathbf{I}$. It requires a sample is close to those in the same class but far from those in different classes [2].

3. REGULAR SIMPLEX ANALYSIS

In this paper, we propose a new criterion for supervised feature extraction. This criterion is

1. Samples from the same class are projected to the same point, that is, the with-class distance is zero;

2. Samples from different classes have unit distance.

This criterion can be mathematically formulated as

$$\begin{cases} ||\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)||_2 = 0, & \text{if } y_i = y_j \\ ||\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)||_2 = 1, & \text{if } y_i \neq y_j. \end{cases} \tag{4}$$

The above equation is usually over determined. The intuition is, although obtaining the ideal projection matrix $\mathbf{W}$ is infeasible, we can define an ideal subspace, then we can find a projection matrix which results a subspace aligning to the ideal subspace as close as possible. It is easy to figure out that, when there are two classes, i.e. $c = 2$, the ideal subspace is a line segment with unit length. When $c = 3$, the ideal subspace is a regular triangle with unit edge length. And when $c = 4$, the ideal subspace is a regular tetrahedron with unit edge length. However, when $c > 5$, it is really difficult to imagine. In fact, the ideal subspace is a regular simplex [4]. We will introduce the regular simplex in the following.

3.1. Regular Simplex

In geometry, a simplex or m-simplex is an m-dimensional analogue of a triangle. Specifically, a m-simplex is the convex hull of a set of $m + 1$ affinely independent points in Euclidean space of dimension m. For example, a 0-simplex is a point, a 1-simplex is a line segment, a 2-simplex is a triangle, a 3-simplex is a tetrahedron.

A regular simplex is a special simplex that is also a regular polytope. For example, a 2-simplex is a regular triangle, and a 3-simplex is a regular tetrahedron. It has been proved that all regular m-simplex [4] in $\mathbb{R}^m$ with pairwise distance 1 are congruent. That is, all regular m-simplex with pairwise distance 1 are identical under translation, rotation and reflection. Suppose there are c classes, our goal is to construct a regular $c-1$-simplex as the ideal subspace. We will give the construction procedure as follows.

Let $\mathbf{s}_i \in \mathbb{R}^{c-1}, i = 1, 2, \ldots, c$, be the vertex of one regular $c-1$-simplex and denote $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_c]$. One can construct $\mathbf{S}$ recursively. The elements in $\mathbf{S}$ can be calculated as

$$\begin{aligned} \mathbf{s}_1 &= [1, 0, \ldots, 0]^T \\ \mathbf{s}_{i,1} &= \frac{-1}{c-1}, i = 2, \ldots, m \\ \mathbf{s}_{i+1,i+1} &= \sqrt{1 - \sum_{l=1}^{i} \mathbf{s}_{l,i}^2} \\ \mathbf{s}_{i+1,j} &= -\frac{\mathbf{s}_{i+1,i+1}}{c-i-1}, j = i+2, \ldots, c \\ \mathbf{s}_{j,i+1} &= 0, j = i+2, \ldots, c-1. \end{aligned} \tag{5}$$

This recursive calculation is repeated until $i = c - 2$ and all the vertices $\mathbf{s}_i$ will be obtained.

It is easy to check that

$$\begin{aligned} \sum_i \mathbf{s}_i &= \mathbf{0}, i = 1, 2, \ldots, c \\ \mathbf{s}_i^T\mathbf{s}_i &= 1, i = 1, 2, \ldots, c \\ ||\mathbf{s}_i - \mathbf{s}_j||_2 &= \sqrt{2 - 2\mathbf{s}_i^T\mathbf{s}_j} = \sqrt{2 + \frac{2}{c-1}}, \forall i \neq j. \end{aligned} \tag{6}$$

In other words, the vertices of regular $c-1$-simplex have zero mean, unit norm and equal pairwise distance.

3.2. Linear Simplex Analysis

Given the ideal subspace for any $c > 1$, our criterion in Eq.(4) is up to a scale $\sqrt{2 + \frac{2}{c-1}}$ equivalent to

$$\mathbf{W}^T\mathbf{x}_i = \mathbf{s}_j, \text{if } y_i = j. \tag{7}$$

Since the criterion has a close relation with regular simplex, we call this criterion as *Regular Simplex Criterion*.

From Eq.(7), we find that Regular Simplex Criterion can be approximated by least square regression,

$$\sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{W}^T\mathbf{x}_i)^2 = \text{tr}(\mathbf{Y} - \mathbf{W}^T\mathbf{X}), \tag{8}$$

where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbb{R}^{c-1\times n}$ is defined as

$$\mathbf{y}_i = \mathbf{s}_j, y_i = j. \tag{9}$$

The solution of Eq.(8) is

$$\mathbf{W} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{Y}^T, \tag{10}$$

where $\mathbf{W}$ is the project we pursue. We call this supervised feature extraction method Linear Simplex Analysis (LSA). It should be noted that the projection learned by LSA is usually not orthogonal, while the projections learned by LDA and MMC are orthogonal.

We summarize the LSA method in Algorithm 1.

Algorithm 1 Linear Simplex Analysis

Input:Training set $\{\mathbf{x}_i, y_i\}_{i=1}^n$,
Output:$\mathbf{W} \in \mathbb{R}^{d\times c}$;
1.Construct the regular $c-1$-simplex $\mathbf{S}$ by Eq.(5);
2.**For** $i = 1$ to n **Do**
  $\mathbf{Y}(:, i) = \mathbf{S}(:, y_i)$;
End For;
3. Calculate the projection $\mathbf{W} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{Y}^T$.

3.3. Relation with LDA

It is very interesting to show that LDA can also be formulated as least square regression [5] as in Eq.(8) where $\mathbf{Y} \in \mathbb{R}^{c\times n}$ is defined as $\mathbf{Y}_{ji} = \begin{cases} \sqrt{\frac{n}{n_j}} - \sqrt{\frac{n_j}{n}}, & \text{if } y_i = j \\ -\sqrt{\frac{n_j}{n}}, & \text{otherwise} \end{cases}$. It is easy to check that $\mathbf{Y}$ also has zero mean. So LDA aims to find a projection to align the subspace to a c-simplex with vertex $\mathbf{s}_j = [-\sqrt{\frac{n_j}{n}}, \ldots, \sqrt{\frac{n}{n_j}} -$

$\sqrt{\frac{n_j}{n}}, \ldots, -\sqrt{\frac{n_j}{n}}]^T$. However, this simplex is not regular, so it is not congruent, i.e. it is not identical under translation, rotation and reflection.

4. REGULARIZED SIMPLEX ANALYSIS

Due to limited training examples, the variance of the estimated $\mathbf{W}$ by least square regression may be large and thus the estimation is not reliable. Especially, when the number of features d is larger than the number of samples n, $\mathbf{X}\mathbf{X}^T$ is singular that the least square regression is ill-posed. An effective way to overcome this problem is to penalize the norm of $\mathbf{W}$, e.g. L_2 norm. Linear regression with L_2 norm regularization is known as ridge regression [1]. The objective function of the multivariate ridge regression is

$$\begin{aligned} & \sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{W}^T\mathbf{x}_i)^2 + \lambda||\mathbf{W}||_F^2 \\ = \quad & \text{tr}(\mathbf{Y}\mathbf{Y}^T - 2\mathbf{W}^T\mathbf{X}\mathbf{Y}^T + \mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W}) + \lambda\text{tr}(\mathbf{W}^T\mathbf{W}), \end{aligned} \tag{11}$$

where λ is a regularizer controls balance between the model complexity and the empirical loss, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ is defined in Eq.(9). Taking the derivative with respect to $\mathbf{W}$ and set it to zero, we have the solution of the multivariate ridge regression as

$$\mathbf{W} = (\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{X}\mathbf{Y}^T. \tag{12}$$

We call this regularized linear simplex analysis as Regularized Simplex Analysis (RSA).

Given a testing data set $\mathbf{X}^t = [\mathbf{x}_1^t, \ldots, \mathbf{x}_l^t]$, the projection can be calculated by

$$\mathbf{Z}^t = \mathbf{W}^T\mathbf{X}^t = \mathbf{Y}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{X}^t. \tag{13}$$

5. KERNEL SIMPLEX ANALYSIS

The methods proposed above are all linear methods. They may fail to discover the intrinsic geometry when the data is highly nonlinear. In this section, we utilize the kernel trick to generalize the RSA to Reproducing Kernel Hilbert Space (RKHS) [6], namely KSA.

We consider the problem in a feature space $\mathcal{F}$ induced by some nonlinear mapping $\phi : \mathbb{R}^d \to \mathcal{F}$. For a proper chosen ϕ, the inner product $\langle , \rangle$ in $\mathcal{F}$ is defined as

$$\langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle = K(\mathbf{x}, \mathbf{y}), \tag{14}$$

where $K(,)$ is a positive semi-definite kernel function. The mostly used kernel functions include:

1. **Polynomial Kernel**:$K(\mathbf{x}, \mathbf{y}) = (1 + \langle \mathbf{x}, \mathbf{y} \rangle)^d$;
2. **Gaussian Kernel**:$K(\mathbf{x}, \mathbf{y}) = \exp(-\frac{||\mathbf{x}-\mathbf{y}||^2}{\sigma^2})$.

Before we give KSA, we first present a theorem in the following.

Theorem 1 $(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{X}\mathbf{Y}^T = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{Y}^T$.

Proof: See Appendix A.

By Theorem 1, Eq.(12) equals to

$$\mathbf{W} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{Y}^T. \tag{15}$$

Let $\Phi = [\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \ldots, \phi(\mathbf{x}_n)]$ denote the data matrix in RKHS, then Eq.(15) can be written as follows:

$$\mathbf{W} = \Phi(\Phi^T\Phi + \lambda\mathbf{I})^{-1}\mathbf{Y}^T = \Phi(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{Y}^T, \tag{16}$$

where $\mathbf{K}$ is the kernel matrix with element $\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$.

Given a testing data set $\mathbf{X}^t = [\mathbf{x}_1^t, \ldots, \mathbf{x}_l^t]$, the projection of KSA is

$$\mathbf{Z}^t = \mathbf{W}^T\Phi^t = \mathbf{Y}(\mathbf{K} + \lambda\mathbf{I})^{-1}\Phi^T\Phi^t = \mathbf{Y}(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{K}^t, \tag{17}$$

where $\Phi^t = [\phi(\mathbf{x}_1^t), \phi(\mathbf{x}_2^t), \ldots, \phi(\mathbf{x}_l^t)]$ and $\mathbf{K}^t$ is the kernel matrix with element $\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j^t)$.

6. EXPERIMENTS

In this experiment, we compare the proposed methods with Fisher Criterion and Maximum Margin Criterion on the UCI [7] machine learning benchmark database (e.g. wine, glass, yeast, image, ionosphere, waveform). We choose Nearest Neighbor as the classification algorithm. For each method, we project the data to $1 \sim \min(c-1, d)$ dimensions, and choose the best dimensionality corresponding to best classification accuracy. We use Gaussian kernel for all the kernel based methods. The hyperparameters in the methods, e.g. regularizer λ in RDA and RSA, σ in Gaussian kernel, are tuned by 5-fold cross validation on the training set. We randomly choose $\{30\%, 40\%, 50\%, 60\%, 70\%\}$ of the data for training and the rest for testing. Since the training samples are randomly chosen, we repeat this experiment 10 times and calculate the average result.

Table.1 summarizes the characteristics of the subset of UCI machine learning database used in this experiment.

Table 1. Description of a subset of UCI database

Datasets	#samples (n)	#features (d)	#classes (c)
wine	178	13	3
glass	214	9	6
yeast	1484	8	10
image	210	19	7
ionosphere	351	34	2
waveform	5000	21	3

In order to compare Regular Simplex Criterion with Fisher Criterion and Maximum Margin Criterion clearly, we list the classification result of Fisher Criterion (LDA, RDA), Maximum Margin Criterion (MMC) and Regular Simplex Criterion (LSA, RSA) in Fig. 1. We can find that on wine, yeast, ionosphere and waveform data, the performance of Regular Simplex Criterion is nearly the same as Fisher Criterion and Maximum Margin Criterion. On image data, Regular Simplex Criterion is the best. And on glass data, Maximum Margin Criterion is the best. It should be noted that the reason that LSA performs not very well on image and ionosphere is due to the singularity of $\mathbf{X}\mathbf{X}^T$ in Eq.(10). Thus RSA is needed. In summary, Regular Simplex Criterion is as good as Fisher Criterion and Maximum Margin Criterion, and it is even better than Fisher Criterion and Maximum Margin Criterion at some time.

Table 2 lists the classification results of LDA, RDA, Kernel Discriminant Analysis (KDA), MMC, LSA, RSA and KSA of 50% for training and the rest for testing. The performance of methods under Regular Simplex Criterion usually takes the first place. And KSA is especially good on all the datasets.

In summary, Regular Simplex Criterion may provide an alternative of Fisher Criterion and Maximum Margin Criterion. At least, users in machine learning and other related areas have another candidate criterion, i.e. Regular Simplex Criterion, for algorithm design or straightforward application.

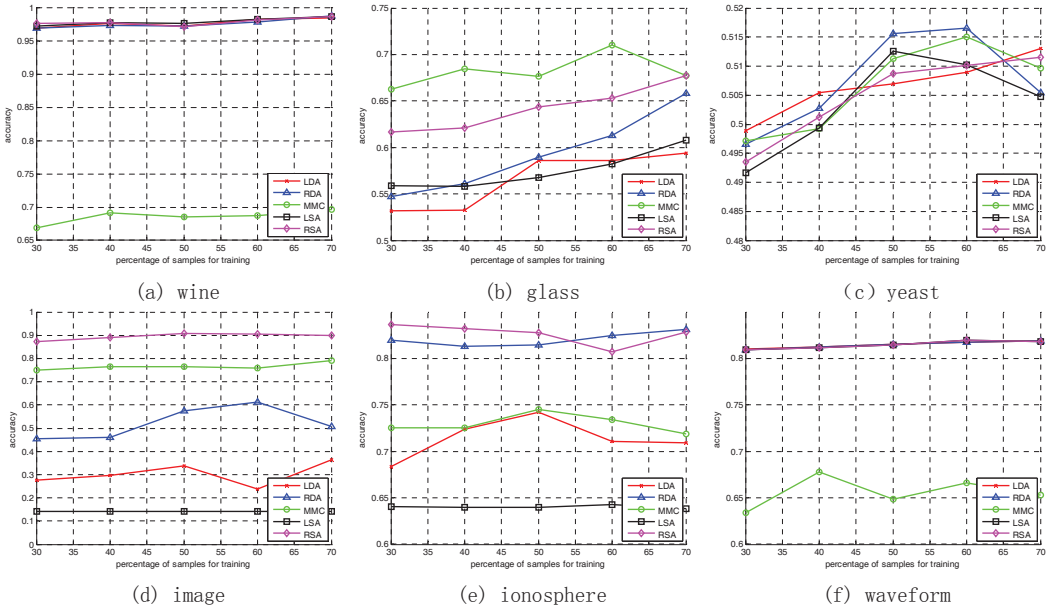

Fig. 1. Classification accuracy on the ORL database with $30\%, 40\%, 50\%, 60\%, 70\%$ samples randomly selected for training and the rest for testing.

Table 2. Classification accuracy on UCI database of 50% for training and the rest for testing

	wine	glass	yeast	image	iono	wave
LDA	97.27	58.57	50.69	33.81	74.23	81.54
RDA	97.27	58.95	51.56	57.52	81.43	81.52
KDA	96.59	62.86	52.77	87.62	93.14	81.55
MMC	68.52	67.62	51.12	76.48	74.51	64.85
LSA	**97.95**	58.76	50.70	14.29	64.00	81.53
RSA	97.16	64.38	50.87	90.86	82.74	81.47
KSA	96.59	**94.29**	**96.14**	**100.0**	**99.43**	**99.88**

7. CONCLUSIONS

The contributions of this paper include three folds: (1) We propose a new criterion, called Regular Simplex Criterion for feature extraction; (2) We present a novel dimensionality reduction method, namely Linear Simplex Analysis (LSA) under the Regular Simplex Criterion; (3) Several variants of LSA, e.g. RSA and KSA, are also proposed. Encouraging experimental results on UCI machine learning database indicate that the new criterion as well as the proposed methods are very effective.

8. REFERENCES

[1] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*, Springer, 2001.

[2] Haifeng Li, Tao Jiang, and Keshu Zhang, "Efficient and robust feature extraction by maximum margin criterion," in *NIPS*, 2003.

[3] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165–175, 1989.

[4] Felix Lazebnik, "On a regular simplex in rn," Tech. Rep., Department of Mathematical Sciences, University of Delaware,Newark, 2004.

[5] Jieping Ye, "Least squares linear discriminant analysis," in *ICML*, 2007, pp. 1087–1093.

[6] John Shawe Taylor and Nello Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.

[7] A. Asuncion and D.J. Newman, "UCI machine learning repository," 2007.

Appendix A

Proof: Since $\mathbf{W} = (\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{X}\mathbf{Y}^T$. then

$$\begin{aligned} & (\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})\mathbf{W} = \mathbf{X}\mathbf{Y}^T \\ \Rightarrow \quad & \mathbf{X}\mathbf{X}^T\mathbf{W} + \lambda\mathbf{W} = \mathbf{X}\mathbf{Y}^T \\ \Rightarrow \quad & \mathbf{W} = \frac{1}{\lambda}\mathbf{X}(\mathbf{Y}^T - \mathbf{X}^T\mathbf{W}) \end{aligned}$$

Let $\mathbf{Z} = \frac{1}{\lambda}(\mathbf{Y}^T - \mathbf{X}^T\mathbf{W})$ then

$$\begin{aligned} & \mathbf{W} = \mathbf{X}\mathbf{Z} \\ \Rightarrow \quad & \lambda\mathbf{Z} = \mathbf{Y}^T - \mathbf{X}^T\mathbf{W} = \mathbf{Y}^T - \mathbf{X}^T\mathbf{X}\mathbf{Z} \\ \Rightarrow \quad & \mathbf{Y}^T = (\mathbf{X}^T\mathbf{X} + \lambda)\mathbf{Z} \\ \Rightarrow \quad & \mathbf{Z} = (\mathbf{X}^T\mathbf{X} + \lambda)^{-1}\mathbf{Y}^T \end{aligned}$$

Therefore

$$\mathbf{W} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda)^{-1}\mathbf{Y}^T$$