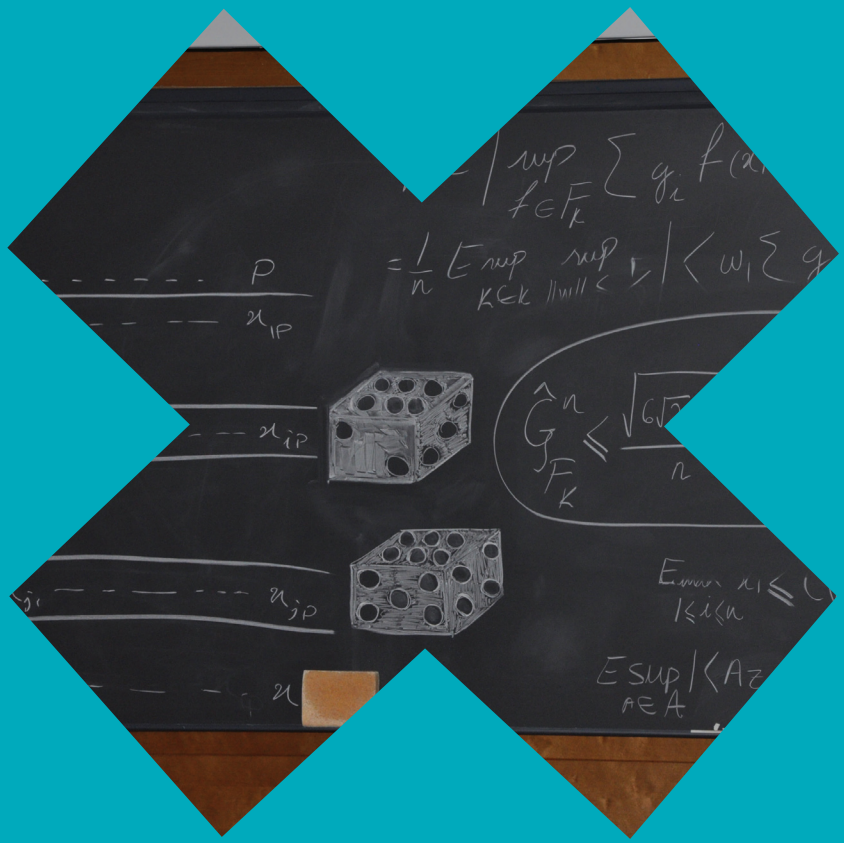
Department of Information and Computer Science

Studies on Kernel Learning and Independent Component Analysis

Nima Reyhani



Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 47/2013

Studies on Kernel Learning and Independent Component Analysis

Nima Reyhani

A doctoral dissertation completed for the degree of Doctor of Science in Technology to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall TU1 of the school on 19 March 2013 at 12.

Aalto University
School of Science
Department of Information and Computer Science

Supervising professor
Prof. Erkki Oja

Thesis advisor
Doc. Ricardo Vigário

Preliminary examiners
Prof. Jyrki Kivinen, University of Helsinki, Finland
Prof. Roman Vershynin, University of Michigan, USA

Opponent
Prof. Masashi Sugiyama, Tokyo Institute of Technology, Japan

Aalto University publication series
DOCTORAL DISSERTATIONS 47/2013

ISBN 978-952-60-5074-4 (printed)
ISBN 978-952-60-5075-1 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-5075-1

Unigrafia Oy
Helsinki 2013

Finland

441 697
Printed matter

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Nima Reyhani

Name of the doctoral dissertation
Studies on Kernel Learning and Independent Component Analysis

Publisher School of Science

Unit Department of Information and Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 47/2013

Field of research Computer and Information Science

Manuscript submitted 9 October 2012 **Date of the defence** 19 March 2013

Permission to publish granted (date) 21 February 2013 **Language** English

☒ **Monograph** ☐ **Article dissertation (summary + original articles)**

Abstract

A crucial step in kernel-based learning is the selection of a proper kernel function or kernel matrix. Multiple kernel learning (MKL), in which a set of kernels are assessed during the learning time, was recently proposed to solve the kernel selection problem. The goal is to estimate a suitable kernel matrix by adjusting a linear combination of the given kernels so that the empirical risk is minimized. MKL is usually a memory demanding optimization problem, which becomes a barrier for large samples.

This study proposes an efficient method for kernel learning by using the low rank property of large kernel matrices which is often observed in applications. The proposed method involves selecting a few eigenvectors of kernel bases and taking a sparse combination of them by minimizing the empirical risk. Empirical results show that the computational demands decrease significantly without compromising classification accuracy, when compared with previous MKL methods.

Computing an upper bound for complexity of the hypothesis set generated by the learned kernel as above is challenging. Here, a novel bound is presented which shows that the Gaussian complexity of such hypothesis set is controlled by the logarithm of the number of involved eigenvectors and their maximum distance, i.e. the geometry of the basis set. This geometric bound sheds more light on the selection of kernel bases, which could not be obtained from previous results.

The rest of this study is a step toward utilizing the statistical learning theory to analyze independent component analysis estimators such as FastICA. This thesis provides a sample convergence analysis for FastICA estimator and shows that the estimations converge in distribution as the number of samples increase. Additionally, similar results for the bootstrap FastICA are established. A direct application of these results is to design a hypothesis testing to study the convergence of the estimates.

Keywords multiple kernel learning, low rank kernel matrix, multiple spectral kernel learning, Gaussian complexity, sparse penalization, source separation, FastICA, bootstrap FastICA, sample convergence analysis

ISBN (printed) 978-952-60-5074-4 **ISBN (pdf)** 978-952-60-5075-1

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Espoo **Location of printing** Helsinki **Year** 2013

Pages 94 **urn** http://urn.fi/URN:ISBN:978-952-60-5075-1

Preface

I enjoyed all of it! Those of us who chose to dedicate some of the best years of their lives to research know that graduate studies have their share of ups and downs. It is challenging and exciting, life changing and life affirming. Yet, it is totally up to us whether to enjoy it or not. Luckily, as I think of my D.Sc. studies, "joy" is among the first things that cross my mind.

I would like to express my deep gratitude to my supervisor, Professor Erkki Oja, for providing me the opportunity to work in his research group. I admire Erkki's enthusiasm to pursue the most challenging questions. I am grateful for his excellent guidance, support, patience, and encouragement specially during difficult times.

I thank Professor Peter J. Bickel for giving me the opportunity to work in his group of outstanding statisticians in UC Berkeley, Statistics Department. I am indebted to Peter for his invaluable instructions in my research and for teaching me mathematical statistics.

I warmly thank Docent Ricardo Vigário for his guidance, professional and significant collaboration, and more than that, for being a trusted friend.

Professor Laurent el Ghaoui and Professor Noureddine el Karoui are warmly thanked for all the valuable scientific discussions and for challenging me with new and deep research questions in optimization and statistics. They greatly improved my problem solving skills.

I thank all of my co-authors for their fruitful collaborations and positive attitude: Docent Ricardo Vigário, Jarkko Ylipaavalniemi, Professor Hideitsu Hino, and Professor Noboru Murata.

I am thankful to the pre-examiners of my thesis, Professor Jyrki Kivinen and Professor Roman Vershynin for their valuable comments, which improved the thesis. Milla Kibble is thanked for editing the language of the thesis.

Most of this work has been carried out in the Adaptive Informatics Research Centre (AIRC) located at the Department of Information and Computer Science, Aalto University. I like to thank Professor Erkki Oja and Professor Olli Simula for providing such an excellent research environment. I am thankful to the staff at ICS department for a pleasant and productive working environment with special thanks to Leila Koivitso, Tarja Pihamaa, and Minna Kauppila. During my studies I spent one year at UC Berkeley, Statistics Department; I wish to thank Choongsoon Bae, Jing Lei, Ying Xu, Sharmodeep Bhattacharyya, Partha Dey, Karl Rohe, Vince Vu, David Purdy, Yulia Gel, Yuval Benjamini, Pegah Jamshidi and Hooshang Jahani for making Berkeley so enjoyable for me during my stay. I specially thank Professor Bin Yu for adopting me in her research group during

my time there and for her special kindness.

I like to thank Intel Research at Berkeley, Common Sense project, for their hospitality during my internship. Ali Rahimi is warmly thanked for his supervision during my stay in Intel research and thereafter; but more than that, for his friendship.

I am very thankful to my friends and colleagues at Aalto with whom I have shared good moments during these years, in and out of the university: Ali Aminian, Ehsan Azmoodeh, Hadi Bordbar, Maryam Borgheie, Kyunghyun Cho, Bahram Dastmalchi, Ramunas Girdziusas, Jussi Gillberg, Nicolau Goncalves, Markus Harva, Javad Hashemi, Hannes Heikinheimo, Ilkka Huopaniemi, Vahid Jafari, Yongnan Ji, Hadi Jorati, Elina Karp, Mahdad Khatibi, Milla Kibble, Gabriella Lener, Arsham Mazaheri, Ali Neissi, Antti Sorjamaa, Tommi Suvitaival, Mehdi Taebnia, Ali Vahdati, and Jarkko Ylipaavalniemi.

Tuomas Jantunen, Gemma Moreno, Denis Deramadi, and Nina von Numers are thanked for their friendship. As I look back to my years in Finland, I realize how important your friendship has been to me.

I acknowledge HECSE for funding my conference trips, and Isaac Newton Institute (Cambridge) for hosting me at their institute for their workshop on Contemporary Frontiers in High-Dimensional Statistical Data Analysis. Funding for the studies was provided by the HUT Department of Computer Science and Engineering and Aalto Department of Information and Computer Science, which are sincerely acknowledged.

I wish to thank my parents and brothers, in law parents, and Tooran for their support and encouragement during all these years.

My deepest and most sincere thank goes to the love of my life, Maral, for endless love, happiness, support, and smart sense of humor.

Helsinki, 10 November, 2012,

Nima Reyhani

Contents

List of Symbols

List of Abbreviations

ERM	Empirical Risk Minimization
fMRI	functional Magnetic Resonance Imaging
ICA	Independent Component Analysis
ISA	Independent Subspace Analysis
LDA	Linear Discriminant Analysis
KKT	Karush-Kuhn-Tucker
MKL	Multiple Kernel Learning
Multiple SKL	Multiple Spectral Kernel Learning
NGCA	Non-Gaussian Subspace Analysis
RBF	Radial basis functions
SDP	Semi-definite programming
SILP	Semi-infinite linear programming
SVM	Support Vector Machines

1. Introduction

1.1 Background

Recent advances in computer networks and storage technologies allow to collect, store and share a significant amount of data with high resolution and fine details. Financial markets, news articles, online social networks, online computer games, network traffic, clinical trials records, and weather data are some examples of major data sources. Such data allows us to have a better understanding of our surroundings. Analysis tools, for example statistical and machine learning devices, become paramount to cope with the considerable amount of data.

A typical statistical data analysis includes making assumptions or building a model for a particular data (model selection), fitting data to the model, and extra model tuning. Having wide applications in data analysis and modeling, regression and classification models are two major topics in machine learning, see e.g. (Lehmann and Casella, 1998; Schölkopf and Smola, 2002; Davison, 2003; Steinwart and Christmann, 2008; Izenman, 2008).

Kernel methods, such as Support Vector Machines (SVM) (Cortes and Vapnik, 1995; Vapnik, 1999), have been successfully applied to many regression and classification problems and have become a standard solution for a significant number of data analysis and modeling problems. For example, see (Burbidge et al., 2001; Sebastiani, 2002; Guyon et al., 2002; Furey et al., 2002; Leopold and Kindermann, 2002; Kim, 2003; Lanckriet et al., 2004a; Evgeniou et al., 2006; Tripathi et al., 2006; Lu et al., 2009; Wilks, 2011).

The core of kernel methods is transferring the samples from the original Euclidean space to a higher dimensional space using a non-linear mapping, called feature map, and then finding the classification/regression function in that new space using linear models (Vapnik, 1999; Schölkopf and Smola, 2002; Steinwart and Christmann, 2008). This transfer improves the accuracy of prediction by allowing to construct nonlinear decision functions in the original space. Many classification methods use the concept of similarity between samples. For the transferred samples, this can be simply computed using a so called kernel function in the original space. The kernel function can be constructed using the feature maps.

The performance of kernel methods depends on the selected feature map or, practically, the kernel function, in addition to other tuning parameters. How-

ever, the selection of a suitable kernel is often left up to the user (Cortes and Vapnik, 1995; Vapnik and Chapelle, 2000; Shawe-Taylor and Kandola, 2002; Vapnik, 1999). One approach to select a suitable kernel function is to use resampling techniques over a set of kernel functions, which does not scale well with the number of kernels or the number of samples. Therefore, the applicability of such approaches is limited, for example see (Chapelle et al., 2002; Sonnenburg et al., 2006; Rakotomamonjy et al., 2008).

Recently, kernel learning and in particular multiple kernel learning (MKL) were proposed to solve the kernel selection problem. MKL searches for a kernel function among a linear combination of given kernels such that it minimizes the empirical risk of the plugged-in SVM. Most of the proposed algorithms for MKL focus on improving the optimization steps involved in MKL. The size of a kernel matrix grows as the number of samples increases, leading to larger memory requirement and a concomitant increase in computational load. This naturally limits the use of MKL methods in very large sample settings.

This thesis focuses on improving the efficiency of kernel learning by approximating the kernel bases with a set of rank one matrices. This type of approximation has been considered in the literature previously, but usually the final kernel learning over the approximated set is handled using ordinary MKL techniques. In particular, we show that, with some approximations, the MKL over this class takes the Basis pursuit (Chen et al., 2001) optimization form, depending on the loss function. For general loss functions we also provide similar results.

In machine learning and statistics, it is important to study the parameters that influence the performance of a model over new sample points. In statistical learning theory, for example (Vapnik, 1999; Koltchinskii, 2011; Mendelson, 2012), it is shown that the complexity of the hypothesis set controls the generalization error, i.e. the prediction error for samples that are out of the training set. In the MKL setting, the hypothesis set consists of linear functions that can be separately constructed by the members of the kernel set. We show that this complexity depends on the geometry (diameter) of the kernel set as well as on its size. In addition, we provide an analysis of the approximation involved in the proposed MKL approach.

Another important model in statistical data analysis is independent component analysis (ICA), which is the focus in the rest of the thesis. In ICA we assume that the samples are linear combinations of a set of independent random variables, called independent components (Hyvärinen, 1999) or source signals. The goal in ICA is to estimate the mixing or de-mixing coefficients, given only the mixed samples. A set of algorithms have been proposed to solve the ICA problem (Hyvärinen et al., 2001). For instance, FastICA algorithm relies on central limit theorem (Van der Vaart and Wellner, 1996), and searches for a vector such that its inner product with the samples vector is as non-Gaussian as possible.

The result of FastICA algorithm may vary with the change of initialization or sampling (Hyvärinen et al., 2001). Nevertheless, there is no straightforward way to find confidence interval for the estimations. One possibility to circumvent this issue is to subsample from the given data set for a number of iterations and then combine the results coming out of FastICA runs. This setup/algorithm is called Bootstrap FastICA. It has been shown that the bootstrapping or similar randomization improves the data exploration in an ICA setting, for example see (Reyhani et al., 2011). Similar technique applies to other ICA algorithms.

In the rest of the thesis, the convergence of the FastICA and bootstrap FastICA

algorithms are studied. In addition, the relations between the rates of convergence to, for instance, the number of independent variables and the sample size are characterized. This is an attempt to establish similar results as of statistical learning theory (Vapnik, 1999) for FastICA. Note that, the convergence of FastICA in population has been studied in literature. However, the connection to sample analysis is missing and the technique of the proof in previous works can hardly be extended for sample analysis, nor for the bootstrap version. Here, applying the empirical process theory, we provide an analysis of FastICA and bootstrap FastICA estimators.

1.2 Contributions

The main contributions of the thesis are summarized as follows. The results appeared as a series of publications (Reyhani and Bickel, 2009; Hino et al., 2010; Reyhani et al., 2011; Ogawa et al., 2011; Reyhani and Oja, 2011; Hino et al., 2012; Reyhani, in print; Ylipaavalniemi et al., 2012).

-Multiple Spectral Kernel Learning

We derive an efficient numerical solution for utilizing the low rank property of kernel matrices in multiple kernel learning framework through adjusting the spectrum of rank one approximations, hence we call it multiple spectral kernel learning (Multiple SKL).

-Geometrical Gaussian Complexity

A geometric Gaussian complexity for the multiple spectral class is derived, by which we show that the complexity of the resulting hypothesis set depends not only on the dictionary size, but also on the diameter of the dictionary.

-Consistency and Asymptotic Normality of FastICA and Bootstrap FastICA

We establish the consistency and asymptotic normality of the FastICA algorithm. In addition, we show that the bootstrap FastICA is asymptotically normal. The results also contain a sample convergence analysis of this algorithm.

1.3 List of related publications

I. Reyhani, N. & Bickel, P. (2009). Nonparametric ICA for nonstationary instantaneous mixtures. In *Proceedings of Workshop on Learning from non IID Data: Theory, Algorithms and Practice*, ECML–PKDD.

II. Hino, H., Reyhani, N., & Murata, N. (2010). Multiple kernel learning by conditional entropy minimization. In *Machine Learning and Applications (ICMLA), 2010 IEEE Ninth International Conference on*, 223-228.

III. Ogawa, T., Hino, H., Reyhani, N., Murata, N., & Kobayashi, T. (2011). Speaker recognition using multiple kernel learning based on conditional entropy minimization. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2204-2207.

IV. Reyhani, N., Ylipaavalniemi, J., Vigário, R., & Oja, E. (2011). Consistency

and asymptotic normality of FastICA and bootstrap FastICA. *Signal Processing*, 92(8), 1767-1778.

V. Reyhani, N., Hino, H., & Vigário, R. (2011). New Probabilistic Bounds on Eigenvalues and Eigenvectors of Random Kernel Matrices. In *Proceedings of the Twenty-Seventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-11)*, AUAI Press, 627-634.

VI. Reyhani, N., & Oja, E. (2011). Non-Gaussian component analysis using Density Gradient Covariance matrix. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, IEEE, 966-972.

VII. Ylipaavalniemi, J., Reyhani, N., & Vigário, R. (2012). Distributional convergence of subspace estimates in FastICA: a bootstrap study. In *Proceedings of the 10th International Conference on Latent Variable Analysis and Source Separation, LVA/ICA 2012*, 123-130.

VIII. Hino, H., Reyhani, N., & Murata, N. (2012). Multiple kernel learning with Gaussianity measures. *Neural Computation*, 24(7), 1853-1881.

IX. Reyhani, N. (2013). Multiple spectral kernel learning and a gaussian complexity computation. *Neural Computation*, in print.

1.4 Organization of the thesis

The thesis is organized as follows. We shortly introduce basic results in empirical process theory and concentration inequalities in Chapter 2. Chapter 3 contains a summary on the empirical risk minimization, kernel selection, multiple kernel learning, and our developed framework for the kernel learning. Chapter 4 provides a summary on previous results on Rademacher complexity of hypothesis class of multiple kernels and their relation to proposed framework together with the new bound for the Gaussian complexity. Chapter 5 contains a short overview on the ICA model, some available algorithmic solutions such as FastICA, bootstrap FastICA, and a sample analysis of FastICA. The statistical convergence analysis of the FastICA and Bootstrap FastICA algorithms is provided in Chapter 6. The conclusions are drawn in Chapter 7.

2. Background theory

This chapter provides some definitions and results on empirical process theory and concentration inequalities that will be used in other chapters.

2.1 Notation

The set of natural numbers, real numbers, and real positive numbers are denoted by $\mathbb{N}$, $\mathbb{R}$, and $\mathbb{R}_+$. We denote scalars with lower case letters (e.g. x and λ), and vectors with bold face letters (e.g. $\boldsymbol{x}$ and $\boldsymbol{\lambda}$). We use capital letters for matrices (e.g. K). The subscript for matrix, vectors, or scalars denotes the index. The entries of a vector are denoted by lower case letters with subscript (e.g. x_i for the i-th entry of $\boldsymbol{x}$). The (i, j) entry of a matrix T_l is denoted by $[T_l]_{i,j}$. I_p denotes the identity matrix in $\mathbb{R}^p$. The inner product between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is denoted by $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$. In addition, $\langle \boldsymbol{x}, \boldsymbol{x}' \rangle = \boldsymbol{x}^\top \boldsymbol{x}'$ denotes the Euclidean inner product. The unit ball in $\mathbb{R}^p$ is denoted by $\mathcal{S}^{p-1}$.

For random variable x or random vector $\boldsymbol{x}$, their expectation is denoted by $\mathbb{E}x$ and $\mathbb{E}\boldsymbol{x}$, respectively. The symbol $\perp\!\!\!\perp$ denotes statistical independence. Conditional expectation with respect to x given y is denoted by $\mathbb{E}_{x|y}$ or $\mathbb{E}\{x|y\}$. Similar notation is used for conditional probability. For a sequence of real numbers $\boldsymbol{x} = (x_1, x_2, \dots)$, the ℓ_r-norm, $r \geq 1$, is denoted by $\|\cdot\|_{\ell_r}$ and $\|\boldsymbol{x}\|_{\ell_r} = (\sum_{i=1}^{\infty} x_i^r)^{1/r}$. The ℓ_r space is the space of sequences with finite ℓ_r-norm. We denote the Euclidean norm by $\|\cdot\|$. The space of functions with bounded L_r-norm, i.e. $\{f : \|f\|_{P,r} := \left(\int_{\mathcal{X}} |f|^r \mathrm{d}P\right)^{\frac{1}{r}} < \infty, f : \mathcal{X} \to \mathbb{R}\}$, is denoted by $L_r(\mathcal{X}, P)$, where P denotes the probability measure. The norm in Hilbert space $\mathcal{H}$ is denoted by $\|\cdot\|_{\mathcal{H}}$. The Lipschitz constant for a function $f : \mathbb{R}^p \to \mathbb{R}$ is denoted by $|f|_L$, which is defined as a constant c such that $|f(\boldsymbol{x}) - f(\boldsymbol{x}')| \leq c\|\boldsymbol{x} - \boldsymbol{y}\|$ holds, for all $\boldsymbol{x}, \boldsymbol{x}'$ in domain of f. For a metric space $(\mathcal{A}, d)$, the diameter of a set is denoted by diam, diam $\mathcal{A} := \sup_{x,y \in \mathcal{A}} d(x, y)$ when the sup exists, otherwise diam $\mathcal{A} = \infty$.

2.2 Concentration inequalities

Concentration inequalities characterize the distance between the average of a finite number of samples drawn independently from a particular distribution and the mean or the median of that distribution as a function of, for example, the number of samples. In the asymptotic case, the result matches the law of

large numbers (Van der Vaart and Wellner, 1996). Here, we state Hoeffding's inequality and bounded difference inequality.

Theorem 2.2.1 (Hoeffding's inequality, (Ledoux, 2001; Steinwart and Christmann, 2008)). *Let $x_1, \ldots, x_n$, $n \geq 1$, be random variables distributed independently according to some distribution P, x_i has values in $[a, b], \forall i \leq n$, and $a < b$. Then, we have*

$$P\left\{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mathbb{E}x_i) \geq (b-a)\sqrt{\frac{\tau}{2n}}\right\} \leq \exp(-\tau),$$

for all $\tau > 0$.

Further, let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be independent and identically distributed random variables with values in a Hilbert space $\mathcal{H}$ satisfying $\|\boldsymbol{x}_i\|_\infty \leq B, i = 1, \ldots, n$. Then, for all $\tau > 0$, we have

$$P\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i - \mathbb{E}\boldsymbol{x}\right\|_{\mathcal{H}} \geq B\left(\sqrt{\frac{2\tau}{n}} + \sqrt{\frac{1}{n}} + \frac{4B\tau}{3n}\right)\right\} \leq \exp(-\tau).$$

Theorem 2.2.2 (Bounded difference inequality, (Ledoux, 2001)). *Let us suppose that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ with values in $\mathcal{X} \subseteq \mathbb{R}^p$ are independent, and $f : \mathcal{X}^n \to \mathbb{R}$. Let $c_1, \ldots, c_n$ satisfy*

$$\sup_{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{x}_i'} \left|f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i-1}, \boldsymbol{x}_i, \boldsymbol{x}_{i+1}, \ldots, \boldsymbol{x}_n) - f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i-1}, \boldsymbol{x}_i', \boldsymbol{x}_{i+1}, \ldots, \boldsymbol{x}_n)\right| \leq c_i,$$

for $i = 1, \ldots, n$, and $\boldsymbol{x}_i'$ is an independent copy of $\boldsymbol{x}_i$. Then, for $\tau > 0$ we have

$$P\{|f - \mathbb{E}f| \geq \tau\} \leq 2\exp\left(-\frac{2\tau^2}{\sum_{i=1}^{n} c_i^2}\right).$$

2.3 Empirical process theory

In this section, we briefly bring definitions and results from probability theory, which are mainly borrowed from (Van der Vaart and Wellner, 1996) and (Van der Vaart, 1998).

A sequence of random vectors $\boldsymbol{x}_n$ converges to $\boldsymbol{x}$ in probability, if

$$P\{d(\boldsymbol{x}_n, \boldsymbol{x}) > \epsilon\} \to 0, \quad \forall \epsilon > 0,$$

where the function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ denotes a metric on $\mathbb{R}^p$. We denote the convergence in probability by $\boldsymbol{x}_n \xrightarrow{P} \boldsymbol{x}$. We also use $x_n = o_P(r_n)$ to denote that $x_n = y_n r_n$, with $y_n \xrightarrow{P} 0$.

We say that $\boldsymbol{x}_n$ converges almost surely to $\boldsymbol{x}$ if

$$P\left\{\lim_{n\to\infty} d(\boldsymbol{x}_n, \boldsymbol{x}) = 0\right\} = 1,$$

A sequence $\boldsymbol{x}_n$ converges in distribution if

$$P\{\boldsymbol{x}_n \leq \boldsymbol{y}\} \to P\{\boldsymbol{x} \leq \boldsymbol{y}\},$$

for every $\boldsymbol{y}$ at which the limit distribution $\boldsymbol{y} \mapsto P\{\boldsymbol{x} \leq \boldsymbol{y}\}$ is continuous. We denote the convergence in distribution by $\boldsymbol{x}_n \rightsquigarrow \boldsymbol{x}$.

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be a set of independent random variables with probability distribution P. The empirical measure is the discrete uniform measure on the observations, denoted by $P_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{\boldsymbol{x}_i}$, where $\delta_{\boldsymbol{x}_i}$ is delta-Dirac that puts mass $\frac{1}{n}$ at $\boldsymbol{x}_i, i = 1, \ldots, n$.

Given a measurable function $f : \mathcal{X} \to \mathbb{R}$, we denote the expectation of f under P_n by $P_n f$ and the expectation of f under P by Pf:

$$Pf := \int_{\mathcal{X}} f dP \quad \text{and} \quad P_n f := \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i).$$

A stochastic process is an indexed collection $\{\boldsymbol{x}_t : t \in T\}$ of random variables defined on the same probability space. An empirical process is a stochastic process that is indexed by a function class $\mathcal{F}$ and is defined by

$$\sqrt{n}\,|(P_n - P)f|\,.$$

In analyzing asymptotic behavior of statistical estimators such as maximum likelihood estimator or in studying the generalization error of kernel methods, usually the supremum of the above term appears, i.e.

$$\sqrt{n}\sup_{f\in\mathcal{F}}(P_n - P)f, \tag{2.1}$$

where $\mathcal{F}$ is the set of hypothesis functions consisting of hypothetical estimators/predictors. Most of the results in empirical process theory are about studying convergence of (2.1) or its characterizations depending on the size/entropy of the class $\mathcal{F}$.

The rest of this chapter presents basic concepts and results from empirical process theory, which are essential to study the generalization error of kernel learning method, Chapter 4, and asymptotic normality of FastICA, which is established in Chapter 5 and Chapter 6.

2.3.1 Rademacher and Gaussian Complexities

One of the classical approaches to study the supremum of empirical processes is to reduce this process to a Rademacher process (see below), using the symmetrization device.

Lemma 2.3.1 (Symmetrization lemma (Van der Vaart and Wellner, 1996; Koltchinskii, 2011))**.** *For every nondecreasing, convex function $\psi : \mathbb{R} \to \mathbb{R}$, and class of measurable functions $\mathcal{F}$,*

$$\mathbb{E}\psi\left(\sup_{f\in\mathcal{F}}|(P_n - P)f|\right) \leq 2\mathbb{E}\psi\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(\boldsymbol{x}_i)\right|\right), \tag{2.2}$$

where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are independent and identically distributed random vectors with values in $\mathcal{X} \subseteq \mathbb{R}^p$. Also, $\epsilon_1, \ldots, \epsilon_n$ are independent random variables that take values in $\{-1, +1\}$ with probability $\frac{1}{2}$. We assume that ϵ_i are independent to $\boldsymbol{x}_j$, $\forall 1 \leq i, j \leq n$. The process $X_f = \sum_{i=1}^{n}\epsilon_i f(\boldsymbol{x}_i), \forall f \in \mathcal{F}$ is called Rademacher process.

In the above lemma the random variables $\epsilon_1, \ldots, \epsilon_n$ are called Rademacher random variables. The symmetrization device has been also presented in different forms. For example, taking the assumption of the Lemma 2.3.1 and for

$\varepsilon > 0, n \geq \frac{8v^2}{\varepsilon^2}$, the following holds:

$$P\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i) - Pf\right| > \varepsilon\right\} \leq 4P\left\{\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\epsilon_i f(\boldsymbol{x}_i)\right| > \frac{n\varepsilon}{4}\right\}, \tag{2.3}$$

where $v^2 = \sup_{f\in\mathcal{F}} \mathrm{var}(f)$. For the proof see (Van der Vaart and Wellner, 1996; Mendelson, 2003a).

The expectation of the term with supremum in right hand side of the inequality (2.2) is commonly used as a notion of complexity of a class of functions, which is formally defined below.

Definition 2.3.2 (Rademacher and Gaussian complexity (Bartlett and Mendelson, 2003; Koltchinskii and Panchenko, 2005))**.** *Let us assume that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are independent and identically distributed random vectors with values in $\mathcal{X} \subseteq \mathbb{R}^p$. For a class of measurable functions $\mathcal{F}$ the Rademacher complexity is defined by*

$$\mathcal{R}_{\mathcal{F}}^n = \mathbb{E}\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(\boldsymbol{x}_i)\right|\right\}, \tag{2.4}$$

where $\epsilon_1, \ldots, \epsilon_n$ are Rademacher random variables. The expectation is with respect to both ϵ_i and $\boldsymbol{x}_i$. Similarly the Gaussian complexity is defined by

$$\mathcal{G}_{\mathcal{F}}^n = \mathbb{E}\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} g_i f(\boldsymbol{x}_i)\right|\right\}, \tag{2.5}$$

where $g_1, \ldots, g_n$ are independent normal random variables, i.e. $g_i \sim \mathcal{N}(0,1), i = 1, \ldots, n$. The expectation is with respect to both g_i and $\boldsymbol{x}_i$. Both Normal and Rademacher random variables are independent to $\boldsymbol{x}_i, i = 1, \ldots, n$.

Both the Rademacher and Gaussian complexities of a function class $\mathcal{F}$ measure the supremum of the correlation between any $f \in \mathcal{F}$ and pure independent noise, described either as independent normal or Rademacher random variables. In machine learning, both the Rademacher and Gaussian complexity are interpreted as an index of how likely the prediction class taken from $\mathcal{F}$ may learn the noise rather than learning the data. These two complexity measures are related as the theorem below states.

Theorem 2.3.3. (Tomczak-Jaegermann, 1989) *There are absolute constants c and C such that, for every class $\mathcal{F}$ and every integer n, $c\mathcal{R}_{\mathcal{F}}^n \leq \mathcal{G}_{\mathcal{F}}^n \leq C \ln n \mathcal{R}_{\mathcal{F}}^n$.*

It is usually hard to compute the value of $\mathcal{G}_{\mathcal{F}}^n$. However, it is possible to expand the expectation into conditional expectation as below

$$\mathcal{G}_{\mathcal{F}}^n = \mathbb{E}\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} g_i f(\boldsymbol{x}_i)\right|\right\} = \mathbb{E}\mathbb{E}\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} g_i f(\boldsymbol{x}_i)\right| \,|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right\},$$

and find an upper bound or approximate the conditional expectation. This conditional expectation is called empirical Gaussian complexity and is denoted by $\widehat{\mathcal{G}}_{\mathcal{F}}^n$:

$$\widehat{\mathcal{G}}_{\mathcal{F}}^n := \mathbb{E}\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} g_i f(\boldsymbol{x}_i)\right| \,|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right\}. \tag{2.6}$$

Similarly, the empirical Rademacher complexity, $\widehat{\mathcal{R}}_{\mathcal{F}}^n$ is defined by

$$\widehat{\mathcal{R}}_{\mathcal{F}}^n := \mathbb{E}\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(\boldsymbol{x}_i)\right| |\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\right\}. \tag{2.7}$$

Using concentration inequalities, we can compute $\mathcal{G}_{\mathcal{F}}^n$ ($\mathcal{R}_{\mathcal{F}}^n$) from $\widehat{\mathcal{G}}_{\mathcal{F}}^n$ ($\widehat{\mathcal{R}}_{\mathcal{F}}^n$) when the empirical complexity is bounded. For example, by bounded difference inequality, we have

$$P\left\{\left|\widehat{\mathcal{R}}_{\mathcal{F}}^n - \mathcal{R}_{\mathcal{F}}^n\right| \geq \epsilon\right\} \leq 2\exp\left(-\frac{n\epsilon^2}{8}\right),$$

where we assume that $\forall f \in \mathcal{F}, |f| \leq 1$. Similar holds for the empirical Gaussian complexity (Bartlett and Mendelson, 2003).

2.3.2 Donsker and Glivenko-Cantelli classes of functions

The Rademacher process $X_f := \sum_{i=1}^n \epsilon_i f(\boldsymbol{x}_i), f \in \mathcal{F}, f$ bounded, is a sub-Gaussian process. In other words, the tails of the density of distances between X_f and X_g for $f, g \in \mathcal{F}$ are not heavier than a Gaussian's, i.e.

$$P\{|X_f - X_g| > x\} \leq 2\exp\left(-\frac{1}{2}\frac{x^2}{d(f,g)}\right),$$

for every $f, g \in \mathcal{F}, x > 0$ and semi-metric d on the index set $\mathcal{F}$. This property can be established, for example, by bounded difference inequality. Then, we can apply Maximal inequalities results, (Van der Vaart and Wellner, 1996)-Chapter 2.2, for sub-Gaussian processes, which provides an upper bound for the supremum of X_f over $\mathcal{F}$. This upper bound depends on the complexity or the size of the function class $\mathcal{F}$, which can be measured by covering number or bracketing number.

In this section, we present basic results on some properties of the empirical processes based on the bracketing number, which is defined below.

Definition 2.3.4 (Bracketing number and bracketing integral, (Van der Vaart and Wellner, 1996; Van der Vaart, 1998))**.** *Given functions l and u, both in $L_r(P)$, the bracket $[l,u]$ is the set of all functions f with $l \leq f \leq u$. An ε-bracket in $L_r(P)$ is a bracket $[l,u]$ with $P(u-l)^r < \varepsilon^r$. The bracketing functions l, u must have finite $L_r(P)$-norm but it is not required that they belong to $\mathcal{F}$.*

The bracketing number, $\mathcal{N}_{[]}(\varepsilon, \mathcal{F}, L_r(P))$, is the minimum number of ε-brackets required to cover the set $\mathcal{F}$. The bracketing entropy is $\ln \mathcal{N}_{[]}(\varepsilon, \mathcal{F}, L_r(P))$. The bracketing integral is defined by

$$J_{[]}(\delta, \mathcal{F}, L_2(P)) = \int_0^\delta \sqrt{\ln \mathcal{N}_{[]}(\varepsilon, \mathcal{F}, L_2(P))}d\varepsilon.$$

A class of functions $\mathcal{F}$ is called *P-Glivenko-Cantelli*, if the following holds

$$\sup_{f\in\mathcal{F}}|(P_n - P)f| \to 0 \quad P\text{-almost surely.} \tag{2.8}$$

In the same line, a class of functions $\mathcal{F}$ is called *P-Donsker* if the limit of the sequence of processes $\{\sqrt{n}(P_n - P)f : f \in \mathcal{F}\}$ is a P-Brownian bridge; for details see (Van der Vaart and Wellner, 1996; Koltchinskii, 2011).

There are two important theorems which state the necessary conditions for establishing the convergence of (2.1), stated below.

Theorem 2.3.5 (Glivenko-Cantelli Theorem (Van der Vaart, 1998)-Theorem 19.4)**.** *Every class $\mathcal{F}$ of measurable functions such that $\mathcal{N}_{[]}(\varepsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\varepsilon > 0$ is P-Glivenko-Cantelli.*

Theorem 2.3.6 (Donsker Theorem (Van der Vaart, 1998)-Theorem 19.5)**.** *Every class $\mathcal{F}$ of measurable functions such that $J_{[]}(1, \mathcal{F}, L_2(P)) < \infty$ is P-Donsker.*

Note that, a class of functions that satisfies the Donsker-Theorem's condition, also satisfies the Glivenko-Cantelli-Theorem's condition.

Remark 2.3.7 (Bracketing entropy of Lipschitz parametric functions (Van der Vaart and Wellner, 1996; Van der Vaart, 1998))**.** *Let us consider the function class $\mathcal{F}$ defined by*

$$\mathcal{F} = \left\{ f_\theta : f_\theta : \mathcal{X} \to \mathbb{R}, \theta \in \Theta \right\},$$

which consist of measurable functions that are indexed by a bounded set $\Theta \subset \mathbb{R}^p$, and p is a positive integer constant.

In machine learning and statistics, Θ is usually called the parameter set, and p denotes the dimension of the parameter set. For example, in sparse linear regression Lasso (Tibshirani, 1996) the function class is defined by

$$\mathcal{F} = \left\{ f_{\boldsymbol{\theta}} : f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{x}, \|\boldsymbol{\theta}\|_{\ell_1} \leq C, \boldsymbol{\theta} \in \mathbb{R}^p \right\},$$

for some $C > 0$. Recall that the Lasso regression is defined by $\widehat{y} = \widehat{\boldsymbol{\theta}}^\top \boldsymbol{x}$, where $\widehat{\boldsymbol{\theta}}$ is the solution of

$$\underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\operatorname{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} \|y_i - \boldsymbol{\theta}^\top \boldsymbol{x}_i\| + \lambda \|\boldsymbol{\theta}\|_{\ell_1},$$

and $\lambda > 0$ is the penalization term.

Suppose that there exists a measurable function $\dot{f}$, such that

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq \dot{f}(x) \, \|\theta_1 - \theta_2\| \,, \tag{2.9}$$

where $\theta_1, \theta_2 \in \Theta$ are in a small neighborhood of $\theta_\circ \in \Theta$. Also, assume that $P\|\dot{f}\|_2 < \infty$. Let us consider brackets of size $2\varepsilon\|\dot{f}\|_2$ defined by $[f_\theta - \varepsilon\dot{f}, f_\theta + \varepsilon\dot{f}]$, where θ belongs to a suitably selected subset of Θ. Now, for any $\|\theta_1 - \theta_2\| \leq \varepsilon$, $\varepsilon > 0, \theta_1, \theta_2 \in \Theta$, by Lipschitz assumption (2.9) we have

$$f_{\theta_2} \in \left[f_{\theta_1} - \varepsilon\dot{f}, f_{\theta_1} + \varepsilon\dot{f} \right].$$

Thus, the number of brackets to cover $\mathcal{F}$ is the same as the number of balls of radius $\varepsilon/2$ to cover Θ. The size of bounded subset Θ is at most $\operatorname{diam}(\Theta)$, and we can cover it by $(\operatorname{diam}\Theta/\varepsilon)^p$ cubes of size ε. So, we have,

$$\mathcal{N}_{[]}(\varepsilon\|\dot{f}\|_2, \mathcal{F}, L_2(P)) \leq C \left(\frac{\operatorname{diam}\Theta}{\varepsilon} \right)^p, \quad 0 < \varepsilon < \mathit{diam}\,\Theta,$$

where C depends only on p and Θ. With the preceding display, the entropy is of smaller order than $\log \frac{1}{\varepsilon}$ up to a constant. Hence the $L_2(P)$-bracketing integral exists and is finite, the class $\mathcal{F}$ is P-Donsker, and therefore, P-Glivenko-Cantelli.

3. Multiple spectral kernel learning (multiple SKL)

This chapter gives a brief overview of the prediction problem, kernel methods, and the selection problem in kernel methods. Several treatments for the kernel selection problem, in particular, multiple kernel learning, are also studied. To improve the scalability and efficiency in multiple kernel learning, here, a novel approach is proposed, which uses rank one approximation of the given kernel matrices. Parts of this chapter are presented mainly in (Reyhani, in print), and partially in (Hino et al., 2010; Ogawa et al., 2011; Hino et al., 2012).

3.1 Empirical risk minimization

In prediction problems such as classification or regression, we assume that a set of independent and identically distributed observations are available for training a model, which is called the training set. We further assume that the observations are in form of input and output pairs and the output variable smoothly depends on the multivariate input variable. A new input sample is called a test sample, and the set of test samples is called the test set. We assume that test samples are independent to the training samples but drawn from the same distribution. The trained (fitted) model is used for predicting test samples outputs. This setup is commonly known as supervised learning (Vapnik, 1999; Schölkopf and Smola, 2002).

Let $\mathcal{S}_n = \{(\boldsymbol{x}_i, y_i), i = 1, \dots, n\}$ be a set of independent copies of random vector $(\boldsymbol{x}, y)$ with values in $\mathcal{X} \times \mathcal{Y}$. We assume that $\mathcal{X}$ is a compact subset of finite-dimensional Euclidean space $\mathbb{R}^p$, for some finite non-zero integer p, and $\mathcal{Y}$ is a subset of $\mathbb{R}$. We assume that y smoothly depends on $\boldsymbol{x}$. Here, $\mathcal{S}_n$ denotes the training set.

Let us define the risk of a function $f : \mathcal{X} \to \mathcal{Y}$ with respect to joint density of $(\boldsymbol{x}, y)$ by $\mathbb{E}\ell(y, f(\boldsymbol{x}))$. The function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is called the loss function and it measures the discrepancy between $f(\boldsymbol{x})$ and y. We assume that the loss function is a convex function. Hinge-loss, i.e. $\ell(z, y) = \max\{0, 1 - yz\}$, and least squares loss, i.e. $\ell(z, y) = (z - y)^2$ are common loss functions for classification and regression.

The goal in statistical learning is to estimate a function $\widehat{f} : \mathcal{X} \to \mathcal{Y}$ using the set $\mathcal{S}_n$ so that f minimizes the risk, i.e. $\widehat{f}$ is the solution of

$$\inf_{f:\mathcal{X}\to\mathbb{R}} \mathbb{E}\ell(y, f(\boldsymbol{x})), \tag{3.1}$$

where $f(\boldsymbol{x})$ predicts the output value of y for $\boldsymbol{x} \in \mathcal{X}$, and the expectation is with

respect to joint distribution of $(\boldsymbol{x}, y)$. We assume that the minimization (3.1) can be attained over all smooth functions. A function $\widehat{f}$ which attains the infimum in (3.1) is called prediction or decision rule. From now on, we restrict the set of candidate functions f (hypothesis set) to linear functions, i.e. $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$, for some $\boldsymbol{w} \in \mathbb{R}^p$.

The joint distribution of $(\boldsymbol{x}, y)$ is usually not available, therefore, a natural way to estimate f via (3.1) is through replacing the unknown measure with the empirical measure, which results in the empirical risk minimization:

$$\hat{f} = \underset{\boldsymbol{w} \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle).$$

The solution of the above minimization often leads to overfitting, i.e. almost zero prediction error on the training samples and large error on the test samples. (Cortes and Vapnik, 1995; Vapnik, 1999) among others, suggest restricting the solution $\boldsymbol{w}$ to have minimum norm by adding a regularization/penalization term. The penalized *empirical risk minimization* (ERM), is defined by

$$\min_{\boldsymbol{w} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i, \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle\right) + \frac{\lambda}{2} \|\boldsymbol{w}\|^2, \tag{3.2}$$

where the parameter $\lambda > 0$ is called the penalization parameter and is defined by the user. A similar way to restrict the solution space is to rewrite (3.2) as

$$\begin{aligned} \min_{\boldsymbol{w} \in \mathbb{R}^p} \quad & \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle) \\ & \|\boldsymbol{w}\|^2 \leq \gamma, \end{aligned}$$

for a user defined parameter $\gamma > 0$.

There are other models that end up with the same formulation as (3.2). For example, by using the hinge-loss we obtain support vector machines (SVM) (Cortes and Vapnik, 1995; Vapnik, 1999; Schölkopf and Smola, 2002; Steinwart and Christmann, 2008). Finding an estimator via optimization has a long history in statistics, which is known as M-estimation (Van der Vaart and Wellner, 1996; Lehmann and Casella, 1998; Koltchinskii, 2011). A typical example of this category is the maximum likelihood estimator, which is widely used in statistics and machine learning.

The solution of (3.2) is of the form

$$\langle \boldsymbol{w}, \boldsymbol{x} \rangle = -\frac{1}{n\lambda} \sum_{i=1}^{n} \partial \ell(y_i, \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle) \langle \boldsymbol{x}_i, \boldsymbol{x} \rangle, \tag{3.3}$$

where, $\partial \ell$ denotes subgradient of the convex function ℓ (Ruszczyński, 2006; Steinwart and Christmann, 2008). The above display is known as the Representer's theorem, for example (Zhang, 2001; Steinwart and Christmann, 2008). (3.3) can be simply derived by checking the first order optimality condition, that is

$$\sum_{i=1}^{n} \partial \ell(y_i, \langle \boldsymbol{w}_*, \boldsymbol{x}_i \rangle) \boldsymbol{x}_i + n\lambda \boldsymbol{w}_* = 0.$$

The penalized ERM, (3.2), produces interpretable linear decision functions, however, it may perform poorly in non-linear and complex classification and regression problems.

A major step in risk minimization problem in general, and SVM in particular was the introduction of a non-linear mapping $\psi : \mathcal{X} \to \mathcal{H}$, where the input space $\mathcal{X}$ is mapped to an inner product feature space $\mathcal{H}$ with higher dimension. The samples might be linearly separable in the new higher dimensional space (Schölkopf and Smola, 2002). The decision rule that is produced in this way becomes a nonlinear function in the original space (Cortes and Vapnik, 1995; Vapnik, 1999).

The inner product space $\mathcal{H}$ usually has higher dimension than the original set $\mathcal{X}$, which is a barrier in computations, particularly in evaluating the inner product in (3.3) or the norm computation in (3.2). This issue can be addressed using kernel functions (Aronszajn, 1950; Schölkopf and Smola, 2002), which are defined below.

Definition 3.1.1. *Let $\mathcal{X}$ be a non-empty set. Then, a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a kernel function on $\mathcal{X}$ if there exists a Hilbert space $\mathcal{H}$ over $\mathbb{R}$ and a map $\psi : \mathcal{X} \to \mathcal{H}$ such that for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, we have*

$$k(\boldsymbol{x}, \boldsymbol{x}') = \langle \psi(\boldsymbol{x}), \psi(\boldsymbol{x}') \rangle_{\mathcal{H}},$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is inner product in $\mathcal{H}$. The function ψ is called feature map and the space $\mathcal{H}$ is called feature space.

Given a kernel function k and a sample set $\mathcal{S}_n$, we define the kernel matrix K by

$$[K]_{i,j} := k(\boldsymbol{x}_i, \boldsymbol{x}_j), 1 \leq i, j \leq n. \tag{3.4}$$

In the rest of this chapter, we use the kernel function k and the kernel matrix K interchangeably.

Example 3.1.2. *Two kernel functions that are used in many applications are of the form $k(\boldsymbol{x}, \boldsymbol{x}') = f(\|\boldsymbol{x} - \boldsymbol{x}'\|)$ or $f(\boldsymbol{x}^\top \boldsymbol{x}')$, which are called distance kernel and inner product kernel, respectively. For example a widely used kernel function is radial basis function (RBF), which is defined by*

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{\sigma}\right),$$

where $\sigma > 0$ is the kernel parameter.

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel only if it is symmetric and positive semidefinite (Schölkopf and Smola, 2002; Steinwart and Christmann, 2008). Given a kernel function k and a point $\boldsymbol{x} \in \mathcal{X}$, we can define a feature map $\psi(\boldsymbol{x}) = k(\cdot, \boldsymbol{x}) := \big(k(\boldsymbol{x}', \boldsymbol{x})_{\boldsymbol{x}' \in \mathcal{X}}\big)$. In other words, $k(\cdot, \boldsymbol{x})$ is a vector with entries $k(\boldsymbol{x}', \boldsymbol{x})$ for all $\boldsymbol{x}' \in \mathcal{X}$. It is defined for all $\boldsymbol{x} \in \mathcal{X}$ pointwise, and it might have infinite dimension. Few additional properties of the kernel functions are as follows:

Definition 3.1.3 (Reproducing kernel Hilbert spaces (Aronszajn, 1950; Schölkopf and Smola, 2002; Steinwart and Christmann, 2008))**.** *Let us assume the Hilbert space $\mathcal{H}$ consists of functions mapping from some non-empty set $\mathcal{X}$ to $\mathbb{R}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel of $\mathcal{H}$ if we have $k(\cdot, \boldsymbol{x}) \in \mathcal{H}$, for all $\boldsymbol{x} \in \mathcal{X}$, and the reproducing property*

$$f(\boldsymbol{x}) = \langle f, k(\cdot, \boldsymbol{x}) \rangle_{\mathcal{H}}$$

holds for all $f \in \mathcal{H}$ and all $\boldsymbol{x} \in \mathcal{X}$. Also, $\mathcal{H}$ is called a reproducing kernel Hilbert space (RKHS) over $\mathcal{X}$ if for all $\boldsymbol{x} \in \mathcal{X}$, the Dirac functions $\delta_{\boldsymbol{x}} : \mathcal{H} \to R$, defined by $\delta_{\boldsymbol{x}}(f) := f(\boldsymbol{x})\ \forall \boldsymbol{x} \in \mathcal{X}$ and $f \in \mathcal{H}$, are continuous.

Every RKHS has a unique reproducing kernel function (Steinwart and Christmann, 2008). In addition, every kernel k defines a unique RKHS, which we denote it by $\mathcal{H}_k$. We assume that there exists at least one feature map ψ_k that corresponds to a kernel function k. For RKHS $\mathcal{H}_k$ with kernel function k the feature map $\psi(\boldsymbol{x}) := k(\cdot, \boldsymbol{x}), \boldsymbol{x} \in \mathcal{X}$ is called the *canonical feature map*.

Let us assume that the feature map $\psi_k(\boldsymbol{x})$ and the RKHS kernel function k are available. By replacing $\boldsymbol{x}$ by $\psi_k(\boldsymbol{x})$ in (3.2), we obtain an extension of the penalized ERM in the feature space:

$$m_\ell(k, \lambda) := \min_{\boldsymbol{w} \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i, \langle \boldsymbol{w}, \psi_k(\boldsymbol{x}_i) \rangle_{\mathcal{H}_k}\right) + \frac{\lambda}{2} \|\boldsymbol{w}\|_{\mathcal{H}_k}^2 \tag{3.5}$$

In a similar way as for (3.3), by the general Representer theorem (Zhang, 2001; Steinwart and Christmann, 2008), the prediction rule that is generated by (3.5) is of the form

$$\hat{f}(\boldsymbol{x}) = -\frac{1}{n\lambda} \sum_{i=1}^{n} \partial \ell(y_i, f(\boldsymbol{x}_i)) k(\boldsymbol{x}, \boldsymbol{x}_i). \tag{3.6}$$

Indeed, we can rewrite (3.5) as

$$m_\ell(k, \lambda) := \min_{\boldsymbol{w} \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i, z_i\right) + \frac{\lambda}{2} \|\boldsymbol{w}\|_{\mathcal{H}_k}^2 : \quad z_i = \langle \boldsymbol{w}, \psi_k(\boldsymbol{x}_i) \rangle_{\mathcal{H}_k} \forall i \leq n. \tag{3.7}$$

Let us denote the empirical risk by $\mathcal{L}(\boldsymbol{z}) := \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, z_i)$ with $\boldsymbol{z} = (z_1, \ldots, z_n)^\top$. The Lagrangian (Ruszczyński, 2006) associated with optimization problem (3.7) is

$$L(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\alpha}) := \mathcal{L}(\boldsymbol{z}) + \frac{\lambda}{2} \|\boldsymbol{w}\|_{\mathcal{H}_k}^2 - \sum_{i=1}^{n} \alpha_i (z_i - \langle \boldsymbol{w}, \psi_k(\boldsymbol{x}_i) \rangle_{\mathcal{H}_k}), \tag{3.8}$$

where $\alpha_1, \ldots, \alpha_n$ are Lagrange multipliers and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^\top$. By checking KKT conditions (Ruszczyński, 2006), we obtain

$$\boldsymbol{w} = -\frac{1}{\lambda} \sum_{i=1}^{n} \alpha_i \psi_k(\boldsymbol{x}_i) = -\frac{1}{\lambda} \sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}, \cdot). \tag{3.9}$$

The decision function is of the form $\hat{f}(\boldsymbol{x}) = \langle \boldsymbol{w}, \psi_k(\boldsymbol{x}) \rangle_{\mathcal{H}_k}$, which is equal to

$$\hat{f}(\boldsymbol{x}) = -\frac{1}{\lambda} \sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}, \boldsymbol{x}_i),$$

given that $\langle \psi_k(\boldsymbol{x}_i), \psi_k(\boldsymbol{x}) \rangle_{\mathcal{H}_k} = k(\boldsymbol{x}, \boldsymbol{x}_i)$. By multiplying (3.7) by n, we obtain $\frac{1}{n\lambda}$ instead of $\frac{1}{n}$ in the preceding display.

An important result in the theory of RKHSs is the Mercer's theorem, which provides a series representation for continuous kernels on compact domains. Let us define the operator $T_k : L_2(\mathcal{X}) \to L_2(\mathcal{X})$ by

$$(T_k f)(\boldsymbol{x}) = \int_{\mathcal{X}} k(\boldsymbol{x}, \boldsymbol{x}') f(\boldsymbol{x}') dP(\boldsymbol{x}').$$

The operator T_k is compact, positive, and self-adjoint (Lax, 2002). By the Spectral Theorem (Lax, 2002), the eigenvalues are at most countable. Then, we have the following representation result about the kernel functions:

Theorem 3.1.4 (Mercer's theorem, (Mercer, 1909; Schölkopf and Smola, 2002; Steinwart and Christmann, 2008)). *Suppose $k \in L_\infty(\mathcal{X} \times \mathcal{X})$ is a symmetric real-valued function such that the integral operator T_k is positive definite, i.e.*

$$\int_{\mathcal{X}^2} k(\boldsymbol{x}, \boldsymbol{y}) f(\boldsymbol{x}) f(\boldsymbol{y}) dP(\boldsymbol{x}) dP(\boldsymbol{y}) \geq 0,$$

for all $f \in L_2(\mathcal{X})$. Let $\phi_j \in L_2(\mathcal{X})$ be normalized orthogonal eigenfunctions of T_k associated with the eigenvalues $\lambda_j > 0$, indexed such that $\lambda_1 \geq \lambda_2 \geq \ldots$. Then, the following results hold.

- $\{(\lambda_j)_{j \geq 1}\} \in \ell_1$.
- $k(\boldsymbol{x}, \boldsymbol{x}') = \sum_{j \in I} \lambda_j \phi_j(\boldsymbol{x}) \phi_j(\boldsymbol{x}')$ *for almost all $(\boldsymbol{x}, \boldsymbol{x}') \in \mathcal{X} \times \mathcal{X}$. The index set I is either a finite subset of $\mathbb{N}$ or total $\mathbb{N}$. In the latter case, the series converges absolutely and uniformly for almost all $(\boldsymbol{x}, \boldsymbol{x}')$.*

Any kernel function that satisfies the Mercer's theorem conditions is called Mercer's kernel. Let us denote the eigenvalues and eigenfunctions of the operator T_k by $\lambda_i(k)$ and $u_i(k)$, respectively, and also the eigenvalues and eigenvectors of the kernel matrix K by $\lambda_i(K)$ and $\boldsymbol{u}_i(K)$, $i = 1, \ldots, n$. (Koltchinskii and Giné, 2000) shows that under certain rate of decay of eigenvalues of T_k, the eigenvalues and eigenvectors of K converge to those of T_k as $n \to \infty$, i.e. eigenvalues and eigenvectors of K are consistent estimators of eigenvalues and eigenvectors of T_k.

3.2 Kernel selection and multiple kernel learning: a brief survey

The representation (3.6) shows that the decision rule is a linear combination of similarities between previous observations and a new sample point. The similarity, here, is measured by an inner product or equivalently a kernel function. Therefore, the performance of the prediction rule depends also on the choice of the kernel function. This issue has been investigated empirically in literature, for example (Chapelle et al., 2002; Vapnik and Chapelle, 2000; Schölkopf and Smola, 2002). However, there is no clear guide on how to efficiently select a suitable kernel function for any given classification or regression task.

A common approach for finding a suitable kernel parameter is cross-validation (Schölkopf and Smola, 2002; Steinwart and Christmann, 2008). Let us assume that a finite set of kernel functions $\{k_1, \ldots, k_L\}$ is available, which are called *kernel bases*. Cross validation procedure randomly divides the training samples $\mathcal{S}_n$ into two subsets, training and validation sets. The sizes of these two sets are predefined by the user. The training set is used to find a solution for $m_\ell(k_l, \lambda)$ for some $l = 1, \ldots, L$, and the accuracy of the learned decision rule will be evaluated on the validation set.

After a few iterations of training and validation, a kernel $k_\circ \in \{k_1, \ldots, k_L\}$ with minimum average prediction error on the validation sets will be selected. This averaged error is called the validation error and it can be used to estimate the risk of the decision rule that is generated by the kernel function $k_\circ$. The cross-validation with only one sample in the validation set is called leave-one-out. It is known that leave-one-out error estimation is almost an unbiased estimator of

the risk (Schölkopf and Smola, 2002; Chapelle et al., 2002).

The numerical computation involved in cross-validation depends on the sample size and the number of kernels, L, which naturally confine the application of cross-validation when a large set of kernel candidates are available and/or the training set is large (Amari and Wu, 1999; Bengio, 2000; Chapelle et al., 2002; Lanckriet et al., 2004b). The computations involved in leave-one-out can be reduced by using the fact that the solution of $m_\ell(k, \lambda)$, for a kernel function k, does not change if a non-support vector, a sample $\boldsymbol{x}_i$ with $\partial \ell(y_i, f(\boldsymbol{x}_i)) = 0$ in (3.6) or $\alpha_i = 0$ in (3.9), is removed from the training set (Jaakkola and Haussler, 1999; Vapnik and Chapelle, 2000). Nevertheless, the computational load is not affordable in many applications.

(Amari and Wu, 1999; Bengio, 2000; Chapelle et al., 2002) further suggest parameterizing the given set of kernels, and using the gradient descent method (Ruszczyński, 2006) to find the optimal parameters; instead of using cross validation or other resampling techniques. The optimality is measured by the empirical risk. This framework has been revisited and further developed to kernel learning. For example, (Lanckriet et al., 2004b) proposes to find a positive definite matrix in the span of the kernel bases by minimizing the penalized empirical risk. This framework is called linear *Multiple Kernel Learning* (MKL). A large number of studies has been devoted to analyze different theoretical and numerical aspects of the MKL, such as the complexity of the hypothesis set and efficient algorithms. For instance, see (Bousquet and Herrmann, 2003; Micchelli and Pontil, 2005; Sonnenburg et al., 2006; Zien and Ong, 2007; Rakotomamonjy et al., 2008; Bach, 2008; Cortes et al., 2009; Xu et al., 2009; Koltchinskii and Yuan, 2010; Cortes et al., 2010; Suzuki and Sugiyama, 2011; Hino et al., 2012). The rest of this section introduces the MKL and similar approaches.

The kernel matrix learning, or kernel learning (in brief), can be defined by

$$\min_{K \in \widetilde{\mathcal{K}}} m_\ell(K, \lambda),$$

where $\widetilde{\mathcal{K}}$ is a bounded set of positive definite kernel matrices or functions and $m_\ell(K, \lambda)$ is defined in (3.5), e.g. (Lanckriet et al., 2004b; Micchelli and Pontil, 2005). For example, given a set of kernel bases $\{K_1, \ldots, K_L\}$, we can define the set $\widetilde{\mathcal{K}}$ by Linear Matrix Inequality (LMI) (Boyd et al., 1994):

$$\widetilde{\mathcal{K}} := \Big\{ K = \sum_{i=1}^{L} p_l K_l : \mu I_n \preceq K \preceq \gamma I_n, p_l \geq 0, \mu < \gamma \Big\},$$

where $\mu > 0$ and $\gamma > 0$ are determined by the user. The notation $A \preceq B$ denotes $A - B$ is negative semidefinite. The LMI in the definition of $\widetilde{\mathcal{K}}$ ensures that the condition number of the matrices in $\widetilde{\mathcal{K}}$ are bounded.

Here, we are interested in studying and developing the *linear MKL*, denoted by $m_\ell(\widetilde{\mathcal{K}}, \lambda)$, and is defined by

$$m_\ell(\widetilde{\mathcal{K}}, \lambda) := \min_{K \in \mathcal{K}} \min_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(\boldsymbol{x}_i)) + \frac{\lambda}{2} \|f\|^2_{\mathcal{H}_K}, \tag{3.10}$$

where the kernel set is

$$\widetilde{\mathcal{K}} := \Big\{ K \in \mathbb{R}^{n \times n} : K = \sum_{l=1}^{L} p_l K_l, \boldsymbol{p} \in \Delta_1 \Big\}, \tag{3.11}$$

and

$$\Delta_1 := \Big\{ \boldsymbol{p} \in \mathbb{R}_+^L : \boldsymbol{p} = (p_1, \ldots, p_L)^\top, \sum_{l=1}^{L} p_l \leq 1 \Big\}.$$

The decision rule generated by (3.10) is a linear combination of decision rules generated separately by each kernel function/matrix. This is an implication of lemma below.

Lemma 3.2.1. (Micchelli and Pontil, 2005)**.** *Let us assume that the kernel functions $k_1, \ldots, k_L$ are given, and denote the RKHS generated by these kernels by $\mathcal{H}_{k_1}, \ldots, \mathcal{H}_{k_L}$. Then, the following holds:*

$$\inf_{k \in \widetilde{\mathcal{K}}} \|f\|_{\mathcal{H}_k} = \min_{\substack{f = \sum_{l=1}^{L} f_l \\ f_l \in \mathcal{H}_{k_l},\, l=1,\ldots,L}} \left(\sum_{l=1}^{L} \|f_l\|_{\mathcal{H}_{k_l}}^2 \right)^{\frac{1}{2}},$$

where f belongs to the RKHS that is constructed by the direct sum of $\mathcal{H}_{k_1}, \ldots, \mathcal{H}_{k_L}$ and $\widetilde{\mathcal{K}}$ is the convex hull of $k_1, \ldots, k_L$.

Indeed, using the above lemma, the optimization problem (3.10) reads

$$m_\ell(\widetilde{\mathcal{K}}, \lambda) = \min_{\substack{f_l \in \mathcal{H}_{k_l} \\ l=1,\ldots,L}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_1(\boldsymbol{x}_i) + \cdots + f_L(\boldsymbol{x}_i)) + \frac{\lambda}{2} \sum_{l=1}^{L} \|f_l\|_{\mathcal{H}_{k_l}}^2, \qquad (3.12)$$

which implies that the solution of linear MKL, i.e. (3.10), has additive representation

$$f = f_1 + \cdots + f_L, f_l \in \mathcal{H}_{k_l}.$$

3.2.1 Previous works on linear MKL

(Lanckriet et al., 2004b) phrases the linear MKL in terms of semi-definite programming (SDP) (Ruszczyński, 2006), for the class of fixed trace positive definite linear combination of kernel bases, and for different loss functions. The main goal, there, is to find both bases coefficients (mixing coefficients) and the parameters of penalized ERM simultaneously within a single optimization round. However, the SDP-solvers in general are not scalable, which limits the applicability of this method (Sonnenburg et al., 2006; Rakotomamonjy et al., 2008).

The computational limitation of SDP-MKL has stimulated further researches to resolve numerical difficulties involved in MKL. Many have tackled this difficulty by using the differentiability of the dual of the penalized ERM at the maximal point. The idea originally goes back to, e.g. (Chapelle et al., 2002) and usually leads to an alternating-optimization, which means that the inner optimization, i.e. the dual of the penalized ERM, is solved first, while the outer optimization variables, i.e. mixing coefficients, are fixed and vice versa; until some convergence is achieved. Semi-infinite linear programming (SILP) (Sonnenburg et al., 2006), extended level method (Rakotomamonjy et al., 2008; Xu et al., 2009), and SimpleMKL (Rakotomamonjy et al., 2008) are various implementations of this framework. The main difference between these algorithms is the optimization techniques that are used for finding the mixing coefficients. SimpleMKL and SILP can be summarized in Algorithm 1 below (Xu et al., 2009).

Algorithm 1 SILP and SimpleMKL

Input: $\epsilon > 0, \lambda > 0$, and the kernel basis $K_1, \ldots, K_L$. **Output:** $\boldsymbol{p}$, the coefficients for the kernel combinations.

$i \leftarrow 0$.
repeat
solve

$$\max_{0 \le \boldsymbol{\alpha} \le \frac{1}{n\lambda}, \boldsymbol{\alpha}^\top \boldsymbol{y} = 0} h(\boldsymbol{p}^i, \boldsymbol{\alpha}),$$

and obtain dual variables $\boldsymbol{\alpha}^i$ for a fixed λ.
update $\boldsymbol{p}$ for SILP by the update rule

$$\begin{aligned} \boldsymbol{p}^i \leftarrow & \operatorname*{argmin}_{\boldsymbol{p}} \quad \nu \\ & \text{s.t.} \qquad h(\boldsymbol{p}, \boldsymbol{\alpha}^j) \le \nu, \ j = 0, \ldots, i, \end{aligned}$$

and for SimpleMKL by

$$\boldsymbol{p}^i \leftarrow \operatorname*{argmin}_{\boldsymbol{p}} \frac{1}{2} \|\boldsymbol{p} - \boldsymbol{p}^i\|^2 + \gamma_i (\boldsymbol{p} - \boldsymbol{p}^i) \nabla_{\boldsymbol{p}} h|_{\boldsymbol{p}^i, \boldsymbol{\alpha}^i}.$$

$i \leftarrow i + 1$
until $\|\boldsymbol{p}^i - \boldsymbol{p}^{i-1}\| \le \epsilon$

In this algorithm, the dual of $m_\ell(K, \lambda)$ for the hinge-loss with $K \in \widetilde{\mathcal{K}}$ is denoted by h:

$$h(\boldsymbol{p}, \boldsymbol{\alpha}) = \boldsymbol{\alpha}^\top \mathbf{1} - \frac{1}{2} (\boldsymbol{\alpha} \circ \boldsymbol{y})^\top \left(\sum_{l=1}^{L} p_l K_l \right) (\boldsymbol{\alpha} \circ \boldsymbol{y}),$$

where $\boldsymbol{\alpha}$ is the dual variable and $\boldsymbol{p}$ contains the kernel mixing coefficients. The gradient vector of the dual is

$$\nabla_{\boldsymbol{p}} h|_{\boldsymbol{p}^i, \boldsymbol{\alpha}^i} = -\frac{1}{2} \Big((\boldsymbol{\alpha}^i \circ \boldsymbol{y})^\top K_1 (\boldsymbol{\alpha}^i \circ \boldsymbol{y}), \ldots, (\boldsymbol{\alpha}^i \circ \boldsymbol{y})^\top K_L (\boldsymbol{\alpha}^i \circ \boldsymbol{y}) \Big)^\top.$$

In above, $\nabla_{\boldsymbol{p}} h|_{\boldsymbol{p}^i, \boldsymbol{\alpha}^i}$ denotes $\nabla_{\boldsymbol{p}} h$ evaluated at $\boldsymbol{p}^i, \boldsymbol{\alpha}^i$. Note that, the dual of (3.5) for the hinge loss, SVM, is

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \boldsymbol{\alpha}^\top \mathbf{1} - (\boldsymbol{\alpha} \circ \boldsymbol{y})^\top K (\boldsymbol{\alpha} \circ \boldsymbol{y}) \\ \text{s.t.} \quad & \mathbf{0} \le \boldsymbol{\alpha} \le \frac{1}{n\lambda}, \end{aligned} \tag{3.13}$$

where $\boldsymbol{\alpha} \circ \boldsymbol{y} = (\alpha_1 y_1, \ldots, \alpha_n y_n)^\top$. Additional constraint $\boldsymbol{\alpha}^\top \boldsymbol{y} = 0$ is required if we add bias term to the decision function, i.e. $(\langle \boldsymbol{w}, \psi(\boldsymbol{x}) \rangle + b)$, where $b \in \mathbb{R}$. The derivations can be also carried on for other loss functions, for details see (Sonnenburg et al., 2006; Rakotomamonjy et al., 2008).

In a slightly different setup, (Cortes et al., 2009) simplifies the MKL by limiting the loss function to ℓ_2 loss function, and the kernel set to

$$\widetilde{\mathcal{K}} = \left\{ \sum_{l=1}^{L} p_l K_l, \boldsymbol{p} \in \mathcal{M} \right\},$$

where $\mathcal{M} = \left\{ \boldsymbol{p} \in \mathbb{R}_+^L : \|\boldsymbol{p} - \boldsymbol{p}_0\|^2 \leq \Lambda^2 \right\}$ for a user defined vector $\boldsymbol{p}_0$ and threshold $\Lambda > 0$. Note that the dual of (3.5) for ℓ_2 loss, $\ell(z, y) = (z - y)^2$, is of the form

$$\max_{\boldsymbol{\alpha}} \quad -\lambda \boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \boldsymbol{\alpha}^\top K \boldsymbol{\alpha} + 2\boldsymbol{\alpha}^\top \boldsymbol{y},$$

where $\boldsymbol{\alpha} \in \mathbb{R}_+^n$ is the dual vector variable. Then, the MKL over $\widetilde{\mathcal{K}}$ is

$$\min_{\boldsymbol{p} \in \mathcal{M}} \max_{\boldsymbol{\alpha}} \quad -\lambda \boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \sum_{l=1}^{L} p_l \boldsymbol{\alpha}^\top K_l \boldsymbol{\alpha} + 2\boldsymbol{\alpha}^\top \boldsymbol{y}. \tag{3.14}$$

(Cortes et al., 2009) shows that the above MKL problem can be solved by alternating between $\boldsymbol{p}$ and $\boldsymbol{\alpha}$, where

$$\boldsymbol{p} = \boldsymbol{p}_0 + \Lambda \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} \quad \text{and} \quad \boldsymbol{\alpha} = (K + \lambda I_n)^{-1} \boldsymbol{y},$$

$\boldsymbol{v} = (v_1, \ldots, v_L)^\top$, $v_l = \boldsymbol{\alpha}^\top K_l \boldsymbol{\alpha}$, and $K = \sum_{l=1}^{L} p_l K_l$. $\boldsymbol{p}$ and $\boldsymbol{\alpha}$ are closed form solutions of (3.14). The MKL solution, $\boldsymbol{p}_*$, in above may have lots of small values close to zero due to solution space of $\boldsymbol{p}$. This slightly violates the original idea of kernel selection using MKL. Compared to the other MKL methods, here, we get an extra parameter, Λ, yet to be determined.

3.2.2 Previous works on kernel approximation in MKL

Kernel matrices with large dimensions are usually expected to have low efficient rank (Donoho, 2000; Drineas and Mahoney, 2005), i.e. a small subset of eigenvalues are large and distant, whereas the rest are small and similar. These matrices can be well approximated by low rank approximation, that is keeping the top eigenvalue/eigenvector pairs of the given matrix. The low efficient rank property of large dimensional kernel matrices has been utilized to improve the scalability of many kernel methods (Drineas and Mahoney, 2005; Talwalkar and Rostamizadeh, 2010; Jin et al., 2011). Therefore, it is also reasonable to use a low rank approximation of the given kernel matrices to improve the numerical efficiency in the MKL framework. For example, (Bousquet and Herrmann, 2003; Lanckriet et al., 2004b; Bach, 2008) propose to use rank one kernels as the kernel bases. Each rank one kernel is the self-outer product of a single eigenvector of the given kernel matrix.

(Lanckriet et al., 2004b) proposes a quadratic programming with quadratic constraint for MKL over rank one kernels. For the hinge loss SVM, they suggest the following optimization with quadratic constraints:

$$\begin{aligned} \max_{\boldsymbol{\alpha}, t} \quad & 2\boldsymbol{\alpha}^\top \mathbf{1} - ct \\ \text{s.t.} \quad & (\boldsymbol{v}_l^\top \boldsymbol{\alpha})^2 \leq t, \quad l = 1, \ldots, L \\ & \boldsymbol{\alpha}^\top \boldsymbol{y} = 0, \quad \mathbf{0} \leq \boldsymbol{\alpha} \leq \frac{1}{\lambda} \mathbf{1}, \end{aligned}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, $\boldsymbol{v}_l = (\text{diag } \boldsymbol{y}) \boldsymbol{b}_l$, $\boldsymbol{b}_l$s are orthonormal vectors (eigenvectors of the original kernel bases), and $c > 0$ is a predefined constant. The vector $\boldsymbol{\alpha}$ is the dual vector in the hinge-loss SVM.

Similarly, (Bousquet and Herrmann, 2003) studies the case where rank one bases are constructed from the whole set of eigenvectors of the given kernel matrix. It can be considered as adjusting the eigenvalues of a given kernel matrix.

This kernel learning algorithm consists of repeating a gradient descent step for finding a linear combination of kernels and a usual SVM learning step. However, the experimental results in (Bousquet and Herrmann, 2003) do not encourage this approach in kernel learning.

Both of these two approaches to rank one MKL are designed for transductive setup. This setup assumes that both training and test sets are available at the time of learning. The kernel learning for the inductive setting has not been addressed.

(Bach, 2008) studies a particular setup where the closed form of the spectral decomposition of the Mercer's kernel functions are all available. This method is called hierarchical MKL (HMKL). The closed forms of eigenfunctions are used for building rank one kernels that are functions of all possible selections of input space features, which results in a kernel bases of enormous size. The learning algorithm is a combination of a greedy search algorithm, to select a subset of kernel bases, and the general MKL optimization, e.g. SimpleMKL or SILP, in a loop. The greedy search algorithm confines the set of kernel bases that feeds MKL. The general MKL optimization finds the best mixing vectors and solves the penalized ERM/SVM. The advantage of HMKL over other rank one kernel learning or general MKL is the ability to perform feature selection in the input space. This cannot be easily handled by the previous rank one MKL methods due to the size of the produced kernel bases.

3.2.3 No-loss optimization approaches to MKL

There are other scenarios for kernel learning, which do not search for the optimal kernels combination through minimizing the empirical risk. For instance, (Shawe-Taylor and Kandola, 2002) proposes adjusting the spectrum of the given kernel matrix K, so that the inner product between K and the labels' Gram matrix, i.e. $\langle K, (y_1 \ \dots \ y_n)^\top (y_1 \ \dots \ y_n) \rangle_{\text{tr}}$, becomes maximal. (Shawe-Taylor and Kandola, 2002) shows that the kernel adjustment improves the classification accuracy in several datasets. This approach is called *kernel target-alignment*, see (Shawe-Taylor and Kandola, 2002) for details.

The kernel matrix generated by the kernel target-alignment may lead to overfitting (Bousquet and Herrmann, 2003). Because adjusting the spectrum of the kernel matrix could swap the order of the eigenvectors, which leads to a highly non-smooth decision function. (Bousquet and Herrmann, 2003) suggests adding an extra constraint in the kernel target-alignment to ensure that the order of eigenvectors remains fixed after adjusting the eigenvalues. The empirical prediction results did not show much improvement.

In a recent study (Hino et al., 2010), we proposed finding the mixing coefficient vector $\boldsymbol{p}$ by solving:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}_+^p, \boldsymbol{p} \in \Delta_1} \quad H\left(\sum_{i=1}^n \alpha_i k(\boldsymbol{x}, \boldsymbol{x}_i) \Big| y_1, \dots, y_n \right) - \gamma H\left(\sum_{i=1}^n \alpha_i k(\boldsymbol{x}, \boldsymbol{x}_i) \right), \tag{3.15}$$

where $k(\boldsymbol{x}, \boldsymbol{x}_i) = \sum_{l=1}^L p_l k_l(\boldsymbol{x}, \boldsymbol{x}_j), \forall i = 1, \dots, n$, and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)^\top, \boldsymbol{\alpha} \in \mathbb{R}_+^n$. $\gamma > 0$ is a penalization parameter determined by the user. The function $H(x)$ denotes the Shannon entropy (Cover and Thomas, 2006) of a random variable x and is defined by

$$H(x) = -\int p(z) \ln p(z) dz.$$

The conditional entropy (Cover and Thomas, 2006) of random variable x given z is defined by

$$H(x|z) = \int\int p(x_i, z_i) \ln \frac{p(z_j)}{p(x_i, z_j)} dx dz.$$

The conditional entropy satisfies $H(x|z) = H(z, x) - H(z)$.

The empirical results in (Hino et al., 2010) and (Ogawa et al., 2011) are competitive to other MKL algorithms. In particular, (Ogawa et al., 2011) shows that for the speaker identification problem, the information theory based MKL improves the previous MKL results significantly. Nevertheless, the objective function in (3.15) is highly nonlinear and solving this optimization problem is numerically hard.

In (Hino et al., 2012), we suggest finding a linear combination of kernels, such that the samples in the feature space are as Gaussian as possible. Then, the linear discriminant analysis (LDA) is a Bayes optimal classifier (Izenman, 2008) in the corresponding feature space, implying high classification accuracy. The feature space in this work is the space generated by the canonical feature map of the linear combination of kernels.

The objective function of MKL in (Hino et al., 2012) consists of a cost and a penalization term. The cost term measures the Gaussianity of both the positive and negative classes in the feature space. The penalization term is to guarantee that the positive and negative classes share same covariance matrix in the feature space. Formally, we suggest finding the coefficients by the following minimization:

$$\begin{aligned} \min_{\boldsymbol{p} \in \Delta_1} \quad & M_G\left(\psi_{\boldsymbol{p}}(\mathcal{S}_n^+)\right) + M_G\left(\psi_{\boldsymbol{p}}(\mathcal{S}_n^-)\right) \\ \text{s.t.} \quad & M_V\left(\psi_{\boldsymbol{p}}(\mathcal{S}_n^+), \psi_{\boldsymbol{p}}(\mathcal{S}_n^-)\right) \leq \gamma, \end{aligned} \tag{3.16}$$

where, γ is a penalization parameter,

$$\mathcal{S}_n^+ := \{\boldsymbol{x}_i : (\boldsymbol{x}_i, y_i) \in \mathcal{S}_n, y_i = +1\},$$

and

$$\mathcal{S}_n^- := \{\boldsymbol{x}_i : (\boldsymbol{x}_i, y_i) \in \mathcal{S}_n, y_i = -1\}.$$

The mapping $\psi_{\boldsymbol{p}}$ is the canonical feature map, which corresponds to the kernel function/matrix $k = \sum_{l=1}^{L} p_l k_l$.

The function M_G in (3.16) measures the distance between a Gaussian distribution/characteristic function and an empirical distribution/characteristic function of the observed data. The function M_G is defined using the fact that a random vector $\boldsymbol{x}$ with values in $\mathbb{R}^p$ and characteristic function $c(\boldsymbol{t}), t \in \mathbb{R}^p$ is Gaussian if and only if $-\ln|c(\boldsymbol{t})|^2 = \boldsymbol{t}^\top \Sigma \boldsymbol{t}$, where $\Sigma \in \mathbb{R}^{p \times p}$ is a positive definite matrix. For $\mathcal{D}_n$ is either $\mathcal{S}_n^+$ or $\mathcal{S}_n^-$, they defined the function

$$M_G(\psi_{\boldsymbol{p}}(\mathcal{D}_n)) = \frac{|\mathcal{D}_n|}{n} \left(g_1(\boldsymbol{p}, \mathcal{D}_n) + \ln g_2(\boldsymbol{p}, \mathcal{D}_n)\right)^2,$$

where $g_1(\boldsymbol{p}, \mathcal{D}_n) = \boldsymbol{p}^\top \Sigma_{\mathcal{D}_n} \boldsymbol{p}$,

$$\Sigma_{\mathcal{D}_n} = \frac{1}{|\mathcal{D}_n|} \sum_{\boldsymbol{x}_i \in \mathcal{D}_n} \boldsymbol{v}(\boldsymbol{t}, \boldsymbol{x}_i) \boldsymbol{v}^\top(\boldsymbol{t}, \boldsymbol{x}_i),$$

and

$$\boldsymbol{v}(\boldsymbol{t}, \boldsymbol{x}_j) = \left(k_1(\boldsymbol{t}, \boldsymbol{x}_j) - \frac{1}{|\mathcal{D}_n|} \sum_{\boldsymbol{x}_i \in \mathcal{D}_n} k_1(\boldsymbol{t}, \boldsymbol{x}_i), \ldots, k_L(\boldsymbol{t}, \boldsymbol{x}_j) - \frac{1}{|\mathcal{D}_n|} \sum_{\boldsymbol{x}_i \in \mathcal{D}_n} k_L(\boldsymbol{t}, \boldsymbol{x}_i)\right)^\top.$$

Also,

$$g_2(\boldsymbol{t}, \boldsymbol{p}, \mathcal{D}_n) = \left[\frac{1}{n}\sum_{\boldsymbol{x}_j \in \mathcal{D}_n} \cos\left(\sum_{l=1}^{L} p_l k_l(\boldsymbol{t}, \boldsymbol{x}_j)\right)\right]^2 + \left[\frac{1}{n}\sum_{\boldsymbol{x}_j \in \mathcal{D}_n} \sin\left(\sum_{l=1}^{L} p_l k_l(\boldsymbol{t}, \boldsymbol{x}_j)\right)\right]^2.$$

The expressions above are obtained by computing the characteristic functions of the relevant random variables, for further details and derivations see (Hino et al., 2012). Here, $\boldsymbol{t}$ is an arbitrary sample point that belongs to the training set. The function M_V in (3.16) measures the distance between empirical covariance matrices of two sets of data in the feature space and is defined as follows:

$$M_V(\psi_{\boldsymbol{p}}(S_n^+), \psi_{\boldsymbol{p}}(S_n^-)) = \left(\boldsymbol{p}^\top(\Sigma_{S_n^+} - \Sigma_{S_n^-})\boldsymbol{p}\right)^2.$$

The proposed algorithms for maximum Gaussianity MKL are computationally heavy. This computational issue makes it difficult to use this approach in large scale applications. On the other hand, the empirical results provided in (Hino et al., 2012) show improvement in the classification accuracy, which suggests that this type of MKL requires more work and deeper analysis.

3.3 Multiple spectral kernel learning (Multiple SKL): A novel efficient rank one MKL

3.3.1 Spectral kernel class

Our general idea is to improve the rank one MKL (see Section 3.2.2) by collecting a set of eigenvectors from each given kernel matrix and combine them so that the empirical risk is minimized. We also intend to keep the number of eigenvectors or basis as small as possible. The main goal is to improve scalability of MKL to have lighter memory demand and computational load, while the accuracy is kept quite similar to general MKL.

Let us define *multiple spectral kernel class* by

$$\mathcal{K} := \left\{K \in \mathbb{R}^{n\times n} : K = \sum_{l=1}^{L} p_l \boldsymbol{b}_l \boldsymbol{b}_l^\top, \boldsymbol{p} \in \Delta_1\right\}, \tag{3.17}$$

where the set $\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_L : \boldsymbol{b}_l \in \mathbb{R}^n\}$ is called the *spectral dictionary*, or *dictionary* for short. The vectors $\boldsymbol{b}_l$ are some eigenvectors of the given kernel matrices $K_l, l = 1, \ldots, L$. To build a dictionary, we can take a subset of *eigenvectors* from each given kernel matrix. Note that, the elements of the dictionary are not necessarily orthogonal to each other. One may also construct vectors $\boldsymbol{b}_l$ from a set of orthonormal basis functions in ℓ_2, such as Hermite polynomials, and construct basis vectors by evaluating those functions on the sample points.

We call the kernel learning over the spectral kernel class, *multiple spectral kernel learning* (multiple SKL). In the rest of this section, we first analyze the ERM in terms of eigenvectors of the kernel matrix, and later, present multiple SKL for different loss functions in Section 3.3.2 and 3.3.3.

It is natural to ask about the effect of low rank approximation or actually the role of eigenvectors of the kernel matrix K in the minimization (3.5). Note that,

with the choice of $\boldsymbol{w}$ in (3.9), we can write an equivalent form of (3.5) as

$$\begin{aligned} m_\ell(K,\lambda) = \min_{\boldsymbol{z}} \max_{\boldsymbol{\alpha}} \quad & -\boldsymbol{\alpha}^\top \boldsymbol{z} + \mathcal{L}(\boldsymbol{z}) - \frac{1}{2\lambda}\boldsymbol{\alpha}^\top K \boldsymbol{\alpha} \\ \max_{\boldsymbol{\alpha}} \quad & -\mathcal{L}^*(\boldsymbol{\alpha}) - \frac{1}{2\lambda}\boldsymbol{\alpha}^\top K \boldsymbol{\alpha}, \end{aligned} \tag{3.18}$$

where we used the fact that $\langle \psi_k(\boldsymbol{x}_i), \psi_k(\boldsymbol{x}_j)\rangle_{\mathcal{H}_k} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. The superscript "*" denotes the convex conjugate (Ruszczyński, 2006), which is defined by

$$g^*(\boldsymbol{z}) = \sup_{\theta} \ \theta^\top \boldsymbol{z} - g(\boldsymbol{\theta}),$$

for given function $g : \mathcal{X} \to \mathbb{R}$.

The maximum in (3.18) with respect to $\boldsymbol{\alpha}$ is attained at $\boldsymbol{\alpha}_* = -\lambda K^{-1}\boldsymbol{z}$. By plugging $\boldsymbol{\alpha}_*$ back into (3.18), we get

$$m_\ell(K,\lambda) = \min_{\boldsymbol{z}} \ \mathcal{L}(\boldsymbol{z}) + \frac{\lambda}{2}\boldsymbol{z}^\top K^{-1}\boldsymbol{z}. \tag{3.19}$$

In (3.19) we assume that K^{-1} exists. We can further expand K in terms of its eigenvectors and eigenvalues:

$$m_\ell(K,\lambda) = \min_{\boldsymbol{z}} \ \mathcal{L}(\boldsymbol{z}) + \frac{\lambda}{2}\sum_{i=1}^{n} \frac{1}{\lambda_i(K)} \langle \boldsymbol{z}, \boldsymbol{u}_i(K)\rangle^2, \tag{3.20}$$

where $\boldsymbol{u}_i(K)$ denotes the eigenvector of K corresponding to i-th largest eigenvalue of K. The expansion in (3.20) implies that the penalized ERM, $m_\ell(K,\lambda)$, searches for a vector that simultaneously minimizes the empirical risk $\mathcal{L}$ and is as dissimilar as possible to the eigenvectors of the kernel matrix K—depending on their eigenvalues. The penalized ERM tends to find a vector $\boldsymbol{z}$ such that the summands corresponding to small eigenvalues are close to zero.

In summary, the expansion above suggests that using low rank approximation of the kernel matrices may not change the decision rule significantly, provided that a sufficient number of eigenvector are present. The low rank approximation has the same effect as removing some of summands in the second term of (3.20). However, it requires special treatment due to rank deficiencies. In addition, to compensate the role of small eigenvectors, further eigenvalues adjustment may become necessary.

3.3.2 Multiple SKL with ℓ_2-loss

In this section, we present multiple SKL for ℓ_2 loss function. Later, we generalize it for more general loss functions.

Note that for ℓ_2 loss function $\ell(x,y) = |x-y|^2$, (3.5) reads

$$\begin{aligned} m^2_{\ell_2}(K,\lambda) &:= \min_{\boldsymbol{w}} \ \|\boldsymbol{y} - \boldsymbol{\psi}_K^\top \boldsymbol{w}\|^2 + \lambda\|\boldsymbol{w}\|^2 \\ &= \min_{\boldsymbol{w},\boldsymbol{z}} \ \|\boldsymbol{z}\|^2 + \lambda\|\boldsymbol{w}\|^2 : \boldsymbol{z} = \boldsymbol{y} - \boldsymbol{\psi}_K^\top \boldsymbol{w}, \end{aligned}$$

where $\boldsymbol{\psi}_K = (\psi_K(\boldsymbol{x}_1) \mid \ldots \mid \psi_K(\boldsymbol{x}_n))$ and $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$. By taking the Lagrangian with dual variables $\boldsymbol{\alpha}$, we can equivalently rewrite $m^2_{\ell_2}(K,\lambda)$ as

$$m^2_{\ell_2}(K,\lambda) = \min_{\boldsymbol{w},\boldsymbol{z}} \max_{\boldsymbol{\alpha}} \|\boldsymbol{z}\|^2 + \lambda\|\boldsymbol{w}\|^2 + 2\boldsymbol{\alpha}^\top(\boldsymbol{y} - \boldsymbol{\psi}_K^\top \boldsymbol{w} - \boldsymbol{z}).$$

We can exchange the min and max by Theorem 3.5.2. In addition, by removing the parameters $\boldsymbol{w}$ and $\boldsymbol{z}$, we obtain

$$m^2_{\ell_2}(K,\lambda) = \max_{\boldsymbol{\alpha}} \; 2\boldsymbol{y}^\top\boldsymbol{\alpha} - \frac{1}{\lambda}\boldsymbol{\alpha}^\top(K+\lambda I_n)\boldsymbol{\alpha}, \tag{3.21}$$

Note, that $m^2_{\ell_2} = \lambda \boldsymbol{y}^\top(K+\lambda I_n)^{-1}\boldsymbol{y}$ and $\boldsymbol{\alpha}_* = \lambda(K+\lambda I_n)^{-1}\boldsymbol{y}$. m_{ℓ_2} denotes the value of $m_{\ell_2}(\mathcal{K},\lambda)$ evaluated at the optimum point.

The multiple SKL over the kernel class $\mathcal{K}$, that is defined in (3.17), reads

$$\begin{aligned} m^2_{\ell_2}(\mathcal{K},\lambda) := & \min_{\boldsymbol{p}\in\Delta_1} m^2_{\ell_2}\Big(\sum_{l=1}^{L} p_l \boldsymbol{b}_l\boldsymbol{b}_l^\top, \lambda\Big) \\ = & \min_{\boldsymbol{p}\in\Delta_1}\max_{\boldsymbol{\alpha}} \; 2\boldsymbol{y}^\top\boldsymbol{\alpha} - \frac{1}{\lambda}\boldsymbol{\alpha}^\top\Big(\sum_{l=1}^{L} p_l \boldsymbol{b}_l\boldsymbol{b}_l^\top\Big)\boldsymbol{\alpha} \\ = & \max_{\boldsymbol{\alpha}} \; 2\boldsymbol{y}^\top\boldsymbol{\alpha} - \frac{1}{\lambda}\max_{1\le l\le L}(\boldsymbol{b}_l^\top\boldsymbol{\alpha})^2. \end{aligned}$$

The third line comes from the fact that the ℓ_1-norm achieves its maximum at the coordinates. At the optimum we have

$$m^2_{\ell_2} = \lambda \boldsymbol{y}^\top\left(\sum_{l=1}^{L} p_l \boldsymbol{b}_l\boldsymbol{b}_l^\top\right)^{-1}\boldsymbol{y},$$

and also,

$$\boldsymbol{\alpha}_* = \lambda\left(\sum_{l=1}^{L} p_l \boldsymbol{b}_l\boldsymbol{b}_l^\top\right)^{-1}\boldsymbol{y}. \tag{3.22}$$

Note that, in the above derivations we assume that the spectral dictionary is sufficiently rich so that the inverse in (3.22) exists. Applying the *scaling trick*, $\boldsymbol{\alpha} = \gamma\boldsymbol{\beta}$, such that $\gamma > 0$ and $\max_{1\le l\le L}\left|\boldsymbol{b}_l^\top\boldsymbol{\beta}\right| = 1$, to $m^2_{\ell_2}(\mathcal{K},\lambda)$, results in

$$\begin{aligned} m^2_{\ell_2}(\mathcal{K},\lambda) = \max_{\gamma,\boldsymbol{\beta}} \quad & 2\gamma\boldsymbol{y}^\top\boldsymbol{\beta} - \frac{1}{\lambda}\gamma^2 \\ \text{s.t.} \quad & \max_{1\le l\le L}|\boldsymbol{b}_l^\top\boldsymbol{\beta}| = 1. \end{aligned}$$

By solving the preceding maximization over variable γ, i.e $\gamma_* = \lambda(\boldsymbol{y}^\top\boldsymbol{\beta}_*)$, we obtain

$$\begin{aligned} m^2_{\ell_2}(\mathcal{K},\lambda) = \max_{\beta} \quad & \lambda\left(\boldsymbol{y}^\top\boldsymbol{\beta}\right)^2 \\ \text{s.t.} \; & \max_{1\le l\le L}|\boldsymbol{b}_l^\top\boldsymbol{\beta}| = 1. \end{aligned}$$

If we further replace the constraint $\max_{1\le l\le L}\left|\boldsymbol{b}_l^\top\boldsymbol{\beta}\right| = 1$ by $\left|\boldsymbol{b}_l^\top\boldsymbol{\beta}\right| \le 1, \forall l = 1,\ldots,L$, we get

$$\begin{aligned} m_{\ell_2}(\mathcal{K},\lambda) = \max_{\boldsymbol{\beta}} \quad & \sqrt{\lambda}\left(\boldsymbol{y}^\top\boldsymbol{\beta}\right) \\ \text{s.t.} \quad & \left|\boldsymbol{b}_l^\top\boldsymbol{\beta}\right| \le 1, \; \forall l = 1,\ldots,L, \end{aligned} \tag{3.23}$$

where again at the optimal solutions we have

$$\boldsymbol{\alpha}_* = \gamma_*\boldsymbol{\beta}_* = \sqrt{\lambda} m_{\ell_2}\boldsymbol{\beta}_*. \tag{3.24}$$

The Lagrangian of (3.23) is

$$L(\boldsymbol{q}, \boldsymbol{\beta}) := \sqrt{\lambda}\left(\boldsymbol{y}^\top \boldsymbol{\beta}\right) + \sum_{l=1}^{L}\left(|q_l| - q_l\left(\boldsymbol{b}_l^\top \boldsymbol{\beta}\right)\right),$$

where $\boldsymbol{q}$ is the dual vector. By checking the optimality conditions, we can remove the variable $\boldsymbol{\beta}$ and, therefore, m_{ℓ_2} reads

$$\begin{aligned} m_{\ell_2}(\mathcal{K}, \lambda) = &\min_{\boldsymbol{q}} \ \|\boldsymbol{q}\|_{\ell_1} \\ &\text{s.t.} \quad \boldsymbol{y} = \frac{1}{\sqrt{\lambda}} \sum_{l=1}^{L} q_l \boldsymbol{b}_l, \end{aligned} \tag{3.25}$$

which is a linear and impressively simple and efficient program. Instead of a least-squares type closed-from solution appearing in (Cortes et al., 2009), we end up with the Basis pursuit (Chen et al., 2001) formulations, where an efficient algorithm already exists and this optimization form guarantees to generate sparse solutions (Chen et al., 2001; Bickel et, al., 2009).

The primal variable $\boldsymbol{p}$ can be obtained by checking the optimality conditions. By (3.24) and (3.22) we have

$$\begin{aligned} \boldsymbol{y} =& \frac{1}{\lambda}\left(\sum_{l=1}^{L} p_{l*} \boldsymbol{b}_l \boldsymbol{b}_l^\top\right) \boldsymbol{\alpha}_* = \frac{1}{\lambda}\left(\sum_{l=1}^{L} p_{l*} \boldsymbol{b}_l \boldsymbol{b}_l^\top\right) \gamma_* \boldsymbol{\beta}_* \\ =& \left(\boldsymbol{y}^\top \boldsymbol{\beta}_*\right) \sum_{l=1}^{L} p_{l*} (\boldsymbol{b}_l^\top \boldsymbol{\beta}_*) \boldsymbol{b}_l. \end{aligned}$$

On the other hand, the optimality condition of (3.23) suggests that

$$q_{l*}(\boldsymbol{b}_l^\top \boldsymbol{\beta}_*) = |q_{l*}|,$$

which is equal to

$$|q_{l*}|(\boldsymbol{b}_l^\top \boldsymbol{\beta}) = q_{l*}.$$

Now we choose

$$p_l = \frac{|q_{l*}|}{m_{\ell_2}}, \forall l = 1, \ldots, L, \tag{3.26}$$

where m_{ℓ_2} is the value of (3.25). We obtain

$$\boldsymbol{y} = \frac{(\boldsymbol{y}^\top \boldsymbol{\beta})}{m_{\ell_2}} \sum_{l=1}^{L} |q_{l*}| (\boldsymbol{b}_l^\top \boldsymbol{\beta}_*) \boldsymbol{b}_l = \sum_{l=1}^{L} q_l \boldsymbol{b}_l = \boldsymbol{y},$$

where the last equality follows from the fact that at the optimal point the constraint (3.25) is satisfied. As conclusion, the optimal value of the mixing vector $\boldsymbol{p}$ can be recovered by (3.26).

The constraint in (3.25) is well defined if the vector $\boldsymbol{y}$ is in the range of matrix $(\boldsymbol{b}_1 \mid \ldots \mid \boldsymbol{b}_L)$. Otherwise, we can extend the dictionary by adding coordinate vectors $\boldsymbol{e}_j, j = 1, \ldots, n$ to the dictionary. Based on (3.25) we can write a similar kernel learning on the kernel class $\mathcal{K}^e := \mathcal{K} \cup \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$ as follows:

$$\begin{aligned} m_{\ell_2}(\mathcal{K}^e, \tilde{\lambda}) = &\min_{\boldsymbol{u}, \boldsymbol{v}} \quad \|\boldsymbol{u}\|_{\ell_1} + \|\boldsymbol{v}\|_{\ell_1} \\ &\text{s.t.} \quad \boldsymbol{y} = \sum_{l=1}^{L} u_l \boldsymbol{b}_l + \tilde{\lambda} \boldsymbol{v}. \end{aligned}$$

The preceding minimization is equivalent to

$$\tilde{\lambda} m_{\ell_2}(\mathcal{K}^e, \tilde{\lambda}) = \min_{\boldsymbol{u}} \quad \|\boldsymbol{y} - \sum_{l=1}^{L} u_l \boldsymbol{b}_l\|_{\ell_1} + \tilde{\lambda} \|\boldsymbol{u}\|_{\ell_1}. \tag{3.27}$$

The primal values can be recovered through

$$b_l = \frac{|v_{l*}|}{\tilde{\lambda} m_{\ell_2}} \quad \text{and} \quad p_l = \frac{|u_{l*}|}{m_{\ell_2}},$$

where $\boldsymbol{v} = \boldsymbol{y} - \sum_{l=1}^{L} u_l \boldsymbol{b}_l$ and m_{ℓ_2} is the value of (3.27). Then, the solution becomes

$$K = \sum_{l=1}^{L} p_l \boldsymbol{b}_l \boldsymbol{b}_l^\top + \sum_{i=1}^{n} b_i \boldsymbol{e}_i \boldsymbol{e}_i^\top.$$

3.3.3 Multiple SKL for a general loss function

Let us consider the dual of (3.5) for some kernel matrix K and bounded convex loss function ℓ as appeared in (3.19):

$$m_\ell(K, \lambda) = \min_{\boldsymbol{z}} \ \mathcal{L}(\boldsymbol{z}) + \frac{\lambda}{2} \boldsymbol{z}^\top K^{-1} \boldsymbol{z}.$$

The kernel learning over the kernel class $\widetilde{\mathcal{K}}$, as defined in (3.11), reads,

$$m_\ell(\widetilde{\mathcal{K}}, \lambda) = \min_{\boldsymbol{z}} \ \mathcal{L}(\boldsymbol{z}) + \frac{\lambda}{2} h^2(\boldsymbol{z}), \tag{3.28}$$

where

$$h(\boldsymbol{z}) := \min_{K \in \widetilde{\mathcal{K}}} \|K^{-\frac{1}{2}} \boldsymbol{z}\|.$$

By taking the convex conjugate of $h^2(\boldsymbol{z})$ we obtain

$$\frac{1}{2} h^2(\boldsymbol{z}) = \max_{\boldsymbol{\xi}} \ \boldsymbol{\xi}^\top \boldsymbol{z} - \frac{1}{2} \max_{K \in \widetilde{\mathcal{K}}} \ \boldsymbol{\xi}^\top K \boldsymbol{\xi}.$$

Again, we play the scaling trick: by replacing the variable $\boldsymbol{\xi}$ with $\gamma \boldsymbol{\eta}$, such that

$$\max_{K \in \widetilde{\mathcal{K}}} \boldsymbol{\eta}^\top K \boldsymbol{\eta} = 1. \tag{3.29}$$

The penalization term $h^2(\boldsymbol{z})$ can then be simplified to

$$\frac{1}{2} h^2(\boldsymbol{z}) = \max_{\gamma, \boldsymbol{\eta}} \ \gamma \boldsymbol{\eta}^\top \boldsymbol{z} - \frac{1}{2} \gamma^2 : \ \max_{K \in \widetilde{\mathcal{K}}} \boldsymbol{\eta}^\top K \boldsymbol{\eta} = 1.$$

By solving the above maximization for γ, we obtain,

$$\begin{aligned} \frac{1}{2} h^2(\boldsymbol{z}) &= \max_{\boldsymbol{\eta}} \ \frac{1}{2} (\boldsymbol{\eta}^\top \boldsymbol{z})^2 \quad : \max_{K \in \widetilde{\mathcal{K}}} \boldsymbol{\eta}^\top K \boldsymbol{\eta} = 1 \\ &= \max_{\boldsymbol{\eta}} \ \frac{1}{2} (\boldsymbol{\eta}^\top \boldsymbol{z})^2 \quad : \boldsymbol{\eta}^\top K_l \boldsymbol{\eta} \leq 1, \ l = 1, \ldots, L, \end{aligned}$$

where we used the fact that every kernel $K \in \widetilde{\mathcal{K}}$ has the form $K = \sum_{l=1}^{L} p_l K_l$. Similar to ℓ_2-loss derivation in the previous section, for the optimal value of $\boldsymbol{\eta}_*$, we have

$$\gamma_* = \boldsymbol{\eta}_*^\top \boldsymbol{z}.$$

Thus, we obtain,

$$h(z) = \max_{\eta} \quad \eta^\top z : \ \eta^\top K_l \eta \leq 1, \ l = 1, \ldots, L.$$

The above holds for the general kernel class $\widetilde{\mathcal{K}}$. For the *multiple spectral kernel class* $\mathcal{K}$, defined in (3.17), we can simplify $h(z)$ to

$$\begin{aligned} h(z) &= \max_{\eta} \quad \eta^\top z \ : (\eta^\top b_l)^2 \leq 1, \forall \ l = 1, \ldots, L \\ &= \max_{\eta} \quad \eta^\top z \ : \left| \eta^\top b_l \right| \leq 1, \forall \ l = 1, \ldots, L, \end{aligned} \tag{3.30}$$

which is a linear program. In a similar way as for ℓ_2-loss, we can construct the Lagrangian function

$$L(\eta, q) = \eta^\top z + \sum_{l=1}^{L} (|q_l| - q_l(\eta^\top b_l))$$

where $q = (q_1, \ldots, q_L)^\top$ is dual variable. Therefore, we can simplify $h(z)$ to

$$h(z) = \min_{q} \ \|q\|_{\ell_1} \ : \ z = \sum_{l=1}^{L} q_l b_l.$$

If we now plug $h(z)$ back in (3.28), we obtain

$$m_\ell(\mathcal{K}, \lambda) = \min_{z,q} \quad \mathcal{L}(z) + \frac{\lambda}{2} \|q\|_{\ell_1}^2 : \ z = \sum_{l=1}^{L} q_l b_l$$

or equivalently,

$$m_\ell(\mathcal{K}, \lambda) = \min_{q} \quad \mathcal{L}\Big(\sum_{l=1}^{L} q_l b_l \Big) + \frac{\lambda}{2} \|q\|_{\ell_1}^2. \tag{3.31}$$

The ℓ_1 penalization is due to the simplex constraint in definition of the kernel class $\mathcal{K}$. Therefore, instead of quadratic constraints, as in (Lanckriet et al., 2004b), we obtain linear constraints.

In the same line as in (3.25), the original variable p can be recovered by checking the optimality conditions. Indeed, the constraint used for making the scaling trick (3.29) leads us to

$$\max_{K \in \mathcal{K}} \ \eta_*^\top K \eta_* = \max_{p \in \Delta_1} \ \sum_{l=1}^{L} p_l (\eta_*^\top b_l)^2 = 1.$$

Using the optimality condition of (3.30), we see that by choosing

$$p_l = \frac{|q_{l*}|}{h}$$

we have

$$\begin{aligned} \sum_{l=1}^{L} p_l (\eta_*^\top b_l)^2 &= \frac{1}{h} \sum_{l=1}^{L} |q_l| (\eta_*^\top b_l)^2 \\ &= \frac{1}{h} \sum_{l=1}^{L} q_l (\eta_*^\top b_l) \\ &= \frac{1}{h} \sum_{l=1}^{L} |q_l| = \frac{h}{h} = 1, \end{aligned}$$

which implies that our choice for p_l is optimal. The second and third line are direct implication of the optimality condition of (3.30). The last line follows by the value of $h(z)$ at the optimal solution.

Thus, for low rank kernel matrices, instead of using the complete kernel matrix, we can just take a small subset of eigenvectors from each matrix and apply the optimization problems introduced in (3.25), (3.27), and (3.31). Using this framework we can achieve a considerable amount of decrease in the amount of memory and computation demands compared to general MKL algorithms such as SimpleMKL (Rakotomamonjy et al., 2008) and (Sonnenburg et al., 2006).

The spectral decomposition of kernel matrices is an expensive computational task. However, we only need the top eigenvectors of the kernel matrices once, which can be efficiently computed using power methods, such as Lanczos method; see (Golub and Van Loan, 1996) for more details.

3.4 Nyström-extension for inductive multiple SKL

The formulations presented in Section 3.3 apply to the transductive setting where both training and test sets are available at the time of learning. To use the results of multiple SKL for classifying test samples, access to eigenvectors at the test samples is crucial. A naive approach to solve this issue is to add the test sample to each of the original kernel matrices, and then take the eigenvalue decomposition of the new matrices. However, this solution becomes infeasible for large kernel bases.

Here, we suggest using the Nyström extension, which provides an approximation of eigenvectors of any (bounded) kernel function by discretizing the integral in the eigenvalue equation for a linear operator (Baker, 1977). The Nyström extension is used extensively for improving the scalability of machine learning methods, for example (Drineas and Mahoney, 2005; Talwalkar and Rostamizadeh, 2010).

Let us define a new kernel matrix K_{n+1} using the available kernel matrix K by

$$K_{n+1} = \begin{pmatrix} K & \boldsymbol{k} \\ \boldsymbol{k}^\top & k(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{n+1}) \end{pmatrix},$$

where $\boldsymbol{k} = (k(\boldsymbol{x}_{n+1}, \boldsymbol{x}_1), \ldots, k(\boldsymbol{x}_{n+1}, \boldsymbol{x}_n))^\top$ and $(\boldsymbol{x}_{n+1}, y_{n+1})$ is distributed identically and independently as $(\boldsymbol{x}_1, y_1) \in \mathcal{S}_n$. The Nyström extension approximates the eigenvector $\boldsymbol{u}_i(K_{n+1}), i = 1, \ldots, n+1$ by extending $\boldsymbol{u}_i(K)$ to

$$\boldsymbol{u}_i(K_{n+1}) \approx (\boldsymbol{u}_i(K)^\top \;\; c_i(\boldsymbol{x}_{n+1}))^\top,$$

where

$$\begin{aligned} c_i(\boldsymbol{x}_{n+1}) =& \frac{1}{n\lambda_i(K)} \sum_{j=1}^{n} k(\boldsymbol{x}_j, \boldsymbol{x}_{n+1}) \boldsymbol{e}_j^\top \boldsymbol{u}_i(K) \\ =& \frac{\boldsymbol{k}^\top \boldsymbol{u}_i(K)}{n\lambda_i(K)}. \end{aligned}$$

Going back to our multiple SKL, let us assume that the dictionary is built using eigenvectors of several kernel matrices. For any test sample, we first extend the elements $\boldsymbol{b}_l$ of the spectral dictionary with $l \in \mathcal{I} = \{l : 1 \leq l \leq L, q_l \neq 0\}$ to $\widetilde{\boldsymbol{b}}_l$ using the Nyström extension. We generate a new extended matrix $\sum_{l \in \mathcal{I}} p_l \widetilde{\boldsymbol{b}}_l \widetilde{\boldsymbol{b}}_l^\top$, and then evaluate the *learned classifier* for the kernel matrix with new entries.

Similarly, the Nyström extension can be used for extending other rank one MKL methods to the inductive setting, for example (Lanckriet et al., 2004b).

3.5 Empirical Results

Here, we present the empirical results of the multiple SKL on several classification datasets including a selection of UCI dataset, protein subcellular localization (Zien and Ong, 2007), and flower recognition (Nilsback and Zisserman, 2008).

Our empirical results show that the multiple SKL provides a good classification rate compared to other MKL methods, and to some extent, it improves the scalability of the kernel learning.

3.5.1 Empirical results on UCI data sets

Here, we present empirical comparisons of the classification accuracy between the multiple SKL and state-of-the-art MKL approaches including SimpleMKL (Rakotomamonjy et al., 2008), extended level method (Level method) (Xu et al., 2009), and SILP (Sonnenburg et al., 2006). We perform this study on a selection of UCI-classification datasets, available for download from `http://archive.ics.uci.edu/ml/`. By classification accuracy, we mean the performance of a soft-margin SVM with a kernel matrix obtained via any MKL algorithm.

The classification accuracies of different kernel learning methods are summarized in Table 3.1. For the multiple SKL, we employ the linear programming toolbox of Mosek[1] in Matlab[2] environment. For other methods we used the code provided in (Rakotomamonjy et al., 2008; Sonnenburg et al., 2006; Xu et al., 2009).

Each cell in the Table 3.1 contains two numbers. The number in top is the average classification accuracy (together with standard deviation information) and the number in bottom shows the computation time in seconds for computing the kernel coefficients. The classification accuracy is computed by averaging the classification rate over 50 trials. In each trial 50% of available samples are randomly selected as training and the rest as testing set. In each row n shows the size of the training set.

We used RBF kernels for each data set with kernel widths of $10^{-6}, 10^{-5}, \ldots, 10^{2}$. The singular values of the RBF kernel matrices, with the above width ranges are in line with the low rank assumption, thus we can apply the multiple SKL. We took 20 eigenvectors, corresponding to the 20 largest eigenvalues of the RBF kernel matrices and then built the spectral dictionary proposed in (3.27). The learned kernel is then plugged to SVM with ℓ_2-loss and the hinge-loss functions. Both showed almost identical performances. For simplicity, the results for the hinge-loss are presented in the Table 3.1. For all sample sets, the penalization parameter of SVM is selected using 5-fold cross validation for all MKL methods as well as multiple SKL separately.

The results show that the multiple SKL significantly improves the classification accuracy in 4 datasets and follows the state-of-art in the other two. How-

[1] `www.mosek.com`

[2] `www.mathworks.com`

ever, the computational time is remarkably reduced in all different cases.

Dataset Name	SimpleMKL	SLIP	Level Method	multiple SKL
Iono $n = 175$	92.10± 2.0	92.10± 1.9	92.10 ±1.3	**95.61± 2.51**
Time	33.5±11.6	1161.0±344.2	7.1±4.3	**1.55±.15**
Pima $n = 384$	76.5±1.9	76.9±2.8	76.9±2.1	**98.19±1.45**
Time	39.4±8.8	62.0±15.2	9.1±1.6	**2.12±0.75**
Sonar $n = 104$	79.1±4.5	79.3±4.2	79.0±4.7	**90.96±2.97**
Time	60.1±29.6	1964.3±68.4	24.9±10.6	**0.10±0.08**
Heart $n = 135$	**82.2±2.2**	**82.2±2.2**	**82.2±2.2**	**82.87±2.81**
Time	4.7±2.8	79.2±38.1	2.1±0.4	**1.51±0.45**
Wpbc $n = 198$	**77.0±2.9**	**77.0±2.8**	**76.9±2.9**	72.92 ±2.13
Time	7.8±2.4	142.0±122.3	5.3±1.3	**1.45±.34**
Wdbc $n = 285$	95.7±0.8	**96.4±0.9**	**96.4±0.8**	88.41±0.35
Time	122.9±38.2	146.3±48.3	15.5±7.5	**1.33±.62**
Vote $n = 218$	**96.0±1.1**	95.7±1.0	95.7±1.0	95.03±0.11
Time	23.7±9.7	26.3±12.4	4.1±1.3	**1.11±.43**

Table 3.1. UCI dataset-Numbers, the first line per row displays accuracy of each method, and the numbers in the second line show the learning time in seconds. Values that are in bold face have passed the two samples t-test with 95% confidence interval.

3.5.2 Empirical results for protein subcellular localization

The prediction of subcellular localization of proteins is an important subject in cell biology (Zien and Ong, 2007), where MKL has been successfully applied to this problem (Kloft et al., 2010; Zien and Ong, 2007). The data set in this experiment can be obtained from `http://www.fml.tuebingen.mpg.de/raetsch/suppl/protsubloc`, for details see (Kloft et al., 2010).

The data set contains 69 different kernels that are obtained for 4 different organisms: plants, non-plant eukaryotes, Gram-positive and Gram-negative bacteria. Similar to (Kloft et al., 2010), in each trial we divided the sample set into training and test sets using the sub-sampling index that is provided with the dataset. We applied the multiple SKL on the training set to obtain a kernel matrix. Using that kernel matrix we trained a SVM with hinge-loss and evaluated the classification result on the test set. The results of classification, in terms of averaged Mathew's correlation coefficient (MCC) (Zien and Ong, 2007), are shown in table 3.2. MCC is defined by

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}},$$

where TP and TN denote the number of true positive and true negative classified test samples. Similarly, FP and FN denote the number of false positive and false negative classified test samples.

In our experiments, the original kernels are first normalized with multiplicative normalization suggested in (Kloft et al., 2010; Zien and Ong, 2007). We then took the top 25 eigenvectors of each normalized kernel matrix to build the spectral dictionary. Note, that we executed the algorithm in transductive setting.

While the accuracy of classification is very similar to the result of SILP (as appeared in (Kloft et al., 2010)), the computational load and memory demand in the multiple SKL are greatly improved, as shown in Table 3.1.

organism	(Kloft et al., 2010) [SILP]	multiple SKL
plant	8.18±0.47	8.18±0.47
non plant	**8.97±0.26**	9.01±0.25
positive gram	9.99±0.35	**9.87±0.34**
negative gram	13.07±0.66	13.07±0.07

Table 3.2. The MCC scores computed for the multiple SKL and SILP. Values that are in bold face have passed the two samples t-test with 95% confidence interval.

3.5.3 Empirical results on flower recognition dataset

Flower recognition is an image based classification task with a large number of classes, for which the MKL is shown to improve the classification rate (Nilsback and Zisserman, 2008).

We applied the multiple SKL on the flower images dataset provided in `http://www.robots.ox.ac.uk/~vgg/data/flowers/index.html`. This webpage provides a set of distance matrices, which contain the distances between samples using different feature sets. Four sets of features are used to compute distance matrices over samples: the histogram of gradient orientations, HSV values of the pixels, and the scale invariant feature transforms that are sampled on both the foreground region and its boundary. In our experiments, we used these distance matrices to generate RBF kernel matrices, and the kernels widths are $10^{-2}, 10^{-1}, 1, 10, 10^2, \ldots, 10^7$.

The training set contains 1000 images with 17 classes and 361 samples as test set and the rest for training. We call this remaining set “total training" set. 340 samples out of the total training set are randomly selected for the validation. In each trial, the validation and training sets are resampled from the total training set.

In this experiment, we consider two settings: inductive and transductive. In the transductive setting, the classification accuracy (MSE) over all classes is 94 ± 0.20. The accuracy of inductive setting is 92 ± 0.28, which is lower than in the trandsuctive setting due to the error of Nyström approximation. The best results reported in (Varma and Ray, 2007) and (Nilsback and Zisserman, 2008) by MKL methods is 88 ± 0.3 (for the inductive setting), which is lower than results achieved by the multiple SKL.

Appendix

An error analysis of ERM

Let us consider a kernel matrix K and a perturbation of that, let us say $\widetilde{K}$. Here, we want to study the difference between the solutions of $m_\ell(K, \lambda)$ and $m_\ell(\widetilde{K}, \lambda)$. The following theorem provides an estimate of this difference, which appeared in (Bousquet and Elisseeff, 2002). For the sake of completeness, we also provide a proof for this theorem.

Theorem 3.5.1. *Let us consider kernel functions k and $\tilde{k}$, and denote their feature maps by ψ_k and $\psi_{\tilde{k}}$. For sample set $\mathcal{S}_n$, we denote the feature maps with ψ_K and $\psi_{\widetilde{K}}$, respec-*

tively. We assume that the feature maps are finite dimensional.

Let us consider $m_\ell(K,\lambda)$, where $\ell(\cdot,\cdot)$ is a convex function with respect to the second argument, with Lipschitz norm $|\ell|_L$. Let us further denote the solution of $m_\ell(K,\lambda)$ and $m_\ell(\widetilde{K},\lambda)$ by $\boldsymbol{w}_*$ and $\widetilde{\boldsymbol{w}}_*$. Then, the following holds.

$$\|\boldsymbol{w}_* - \widetilde{\boldsymbol{w}}_*\|^2 \leq \frac{|\ell|_L}{2n\lambda}(\|\boldsymbol{w}_*\| + \|\widetilde{\boldsymbol{w}}_*\|)\sum_{i=1}^{n}\left\|\psi_K(\boldsymbol{x}_i) - \psi_{\widetilde{K}}(\boldsymbol{x}_i)\right\|.$$

Proof (Theorem 3.5.1)**.** We follow similar derivation as appeared in (Bousquet and Elisseeff, 2002) and (Zhang, 2001). Let us define

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}) :=& \frac{1}{n}\sum_{i=1}^{n}\ell(y_i, \langle \boldsymbol{w}, \psi_K(\boldsymbol{x}_i)\rangle),\\ \widetilde{\mathcal{L}}(\widetilde{\boldsymbol{w}}) :=& \frac{1}{n}\sum_{i=1}^{n}\ell(y_i, \langle \widetilde{\boldsymbol{w}}, \psi_{\widetilde{K}}(\boldsymbol{x}_i)\rangle),\end{aligned}$$

and $\Delta\boldsymbol{w} = \widetilde{\boldsymbol{w}}_* - \boldsymbol{w}_*$. Also, assume that $t \in (0,1]$. Since $\boldsymbol{w}_*$ and $\widetilde{\boldsymbol{w}}_*$ attain the minimum of $m_\ell(K,\lambda)$ and $m_\ell(\widetilde{K},\lambda)$, respectively, we have

$$\mathcal{L}(\boldsymbol{w}_*) + \frac{\lambda}{2}\|\boldsymbol{w}_*\|^2 \leq \mathcal{L}(\boldsymbol{w}_* + t\Delta\boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}_* + t\Delta w\|^2,$$

and similarly

$$\widetilde{\mathcal{L}}(\widetilde{\boldsymbol{w}}_*) + \frac{\lambda}{2}\|\widetilde{\boldsymbol{w}}_*\|^2 \leq \widetilde{\mathcal{L}}(\widetilde{\boldsymbol{w}}_* - t\Delta\boldsymbol{w}) + \frac{\lambda}{2}\|\widetilde{\boldsymbol{w}}_* - t\Delta w\|^2.$$

Combining the above inequalities reads

$$\begin{aligned}\frac{\lambda}{2}(\|\boldsymbol{w}_*\|^2 - \|\boldsymbol{w}_* + t\Delta\boldsymbol{w}\|^2 + \|\widetilde{\boldsymbol{w}}_*\|^2 - \|\widetilde{\boldsymbol{w}}_* - t\Delta\boldsymbol{w}\|^2) \leq&\\ \mathcal{L}(\boldsymbol{w}_* + t\Delta\boldsymbol{w}) - \mathcal{L}(\boldsymbol{w}_*) + \widetilde{\mathcal{L}}(\widetilde{\boldsymbol{w}}_* - t\Delta\boldsymbol{w}) - \widetilde{\mathcal{L}}(\widetilde{\boldsymbol{w}}_*),&\end{aligned}$$

which is equal to

$$\lambda(1-t)t\|\Delta\boldsymbol{w}\|^2 \leq \mathcal{L}(\boldsymbol{w}_* + t\Delta\boldsymbol{w}) - \mathcal{L}(\boldsymbol{w}_*) + \widetilde{\mathcal{L}}(\widetilde{\boldsymbol{w}}_* - t\Delta\boldsymbol{w}) - \widetilde{\mathcal{L}}(\widetilde{\boldsymbol{w}}_*).$$

By assumption, the loss function is a convex function, and therefore, we can expand the right hand side of the above inequality:

$$\begin{aligned}\lambda(1-t)\|\Delta\boldsymbol{w}\|^2 \leq& \mathcal{L}(\widetilde{\boldsymbol{w}}_*) - \mathcal{L}(\boldsymbol{w}_*) + \widetilde{\mathcal{L}}(\boldsymbol{w}_*) - \widetilde{\mathcal{L}}(\widetilde{\boldsymbol{w}}_*)\\ =& \frac{1}{n}\sum_{i=1}^{n}\Big(\ell(y_i, \langle\widetilde{\boldsymbol{w}}_*, \psi_K(\boldsymbol{x}_i)\rangle) - \ell(y_i, \langle\boldsymbol{w}_*, \psi_K(\boldsymbol{x}_i)\rangle)\\ &+ \ell(y_i, \langle\boldsymbol{w}_*, \psi_{\widetilde{K}}(\boldsymbol{x}_i)\rangle) - \ell(y_i, \langle\widetilde{\boldsymbol{w}}_*, \psi_{\widetilde{K}}(\boldsymbol{x}_i)\rangle)\Big)\end{aligned}$$

Note that, in the preceding display, we divided both sides by t. By the Lipschitz assumption, we further have,

$$\begin{aligned}\lambda(1-t)\|\Delta\boldsymbol{w}\|^2 \leq& \frac{|\ell|_L}{n}\sum_{i=1}^{n}\Big(|\langle\widetilde{\boldsymbol{w}}_*, \psi_K(\boldsymbol{x}_i) - \psi_{\widetilde{K}}(\boldsymbol{x}_i)\rangle|\\ &+ |\langle\boldsymbol{w}_*, \psi_{\widetilde{K}}(\boldsymbol{x}_i) - \psi_K(\boldsymbol{x}_i)\rangle|\Big)\end{aligned}$$

Let us take the limit of the above expression when $t \to 0$. Then, by the Cauchy-Shwartz inequality, we obtain

$$\|\Delta\boldsymbol{w}\|^2 \leq \frac{|\ell|_L(\|\boldsymbol{w}_*\| + \|\widetilde{\boldsymbol{w}}_*\|)}{n\lambda}\sum_{i=1}^{n}\|\psi_K(\boldsymbol{x}_i) - \psi_{\widetilde{K}}(\boldsymbol{x}_i)\|$$

□

Let us assume that the kernel matrix K has eigenvalue decomposition

$$K = \sum_{i=1}^{n} \lambda_i(K) \boldsymbol{u}_i(K) \boldsymbol{u}_i^\top(K).$$

We define the feature map

$$\psi_K(\boldsymbol{x}_i) = (\sqrt{\lambda_1(K)} \boldsymbol{e}_i^\top \boldsymbol{u}_1(K), \ldots, \sqrt{\lambda_n(K)} \boldsymbol{e}_i^\top \boldsymbol{u}_n(K)).$$

Let us assume that the low rank approximation of $\widetilde{K}$ contains top R eigenvectors of the kernel matrix K. Then, it has similar feature map:

$$\psi_{\widetilde{K}}(\boldsymbol{x}_i) = (\sqrt{\lambda_1(K)} \boldsymbol{e}_i^\top \boldsymbol{u}_1(K), \ldots, \sqrt{\lambda_R(K)} \boldsymbol{e}_i^\top \boldsymbol{u}_R(K)).$$

Using Theorem 3.5.1, we have

$$\|\Delta \boldsymbol{w}\|^2 \leq \frac{|\ell|_L(\|\boldsymbol{w}_*\| + \|\widetilde{\boldsymbol{w}}_*\|)}{n\lambda} \sum_{i=1}^{n} \|(\boldsymbol{0}_R, \sqrt{\lambda_{R+1}(K)} \boldsymbol{e}_i^\top \boldsymbol{u}_{R+1}, \ldots, \sqrt{\lambda_n(K)} \boldsymbol{e}_i^\top \boldsymbol{u}_n(K))\|.$$

Then,

$$\begin{aligned}\|\Delta \boldsymbol{w}\|^2 \leq & \frac{|\ell|_L(\|\boldsymbol{w}_*\| + \|\widetilde{\boldsymbol{w}}_*\|)}{n\lambda} \sum_{i=1}^{n} \left(\lambda_{R+1}(K) \sum_{j=R+1}^{n} (\boldsymbol{e}_i^\top \boldsymbol{u}_j(K))^2 \right)^{\frac{1}{2}} \\ \leq & \frac{|\ell|_L(\|\boldsymbol{w}_*\| + \|\widetilde{\boldsymbol{w}}_*\|)}{\lambda} \sqrt{\lambda_{R+1}(K)(n-(R+1))}.\end{aligned}$$

On the other hand both $\boldsymbol{w}_*$ and $\widetilde{\boldsymbol{w}}_*$ have representation $\frac{1}{n\lambda} \sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}_i, \cdot)$ for different values of $\alpha_i \geq 0$ for each weight vector. If we further assume that $k(\boldsymbol{x}, \boldsymbol{x}) \leq T$, for $T > 0$, we obtain,

$$\begin{aligned}\|\boldsymbol{w}_*\| &= \left\| \frac{1}{n\lambda} \sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}_i, \cdot) \right\| \\ &\leq \frac{\|\boldsymbol{\alpha}\|_\infty}{n\lambda} \sum_{i=1}^{n} \|k(\boldsymbol{x}_i, \boldsymbol{x}_1), \ldots, k(\boldsymbol{x}_i, \boldsymbol{x}_n)\| \\ &\leq \frac{\|\boldsymbol{\alpha}\|_\infty \sqrt{n} T}{\lambda}.\end{aligned}$$

Similar holds for $\widetilde{\boldsymbol{w}}_*$. For the hinge loss, the dual variables in (3.13) satisfiy $\|\boldsymbol{\alpha}\|_\infty \leq \frac{1}{n\lambda}$. Therefore, for differences between weight vectors we obtain,

$$\|\Delta \boldsymbol{w}\|^2 \leq \frac{2T|\ell|_L}{\lambda^2} \sqrt{\frac{\lambda_{R+1}(K)(n-(R+1))}{n}}.$$

Saddle point theorem

For some function $\psi : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ where $\mathcal{X} \subseteq \mathbb{R}^p, \mathcal{Z} \subseteq \mathbb{R}^q$, let us define functions $t_z : \mathbb{R}^p \to (-\infty, \infty]$ and $r_{\boldsymbol{x}} : \mathbb{R}^q \to \mathbb{R}$ by

$$t_z(\boldsymbol{x}) = \begin{cases} \psi(\boldsymbol{x}, \boldsymbol{z}) & \boldsymbol{x} \in \mathcal{X} \\ \infty & \boldsymbol{x} \notin \mathcal{X} \end{cases},$$

and

$$r_x(z) = \begin{cases} -\psi(x,z) & z \in \mathcal{Z} \\ \infty & z \notin \mathcal{Z} \end{cases},$$

respectively. Let us further define functions $t(x) := \sup_{z\in\mathcal{Z}} t_z(x)$ and $r(z) := \sup_{x\in\mathcal{X}} r_x(z)$.

Theorem 3.5.2 (Saddle point theorem (Bertsekas et al., 2003)-Theorem 2.6.4)**.** *Let us assume that the function $t_z, \forall z \in \mathcal{Z}$ is closed and convex, and similarly, $r_x, \forall x \in \mathcal{X}$ is closed and convex,*

$$\inf_{x\in\mathcal{X}} \sup_{z\in\mathcal{Z}} \psi(x,z) < \infty,$$

and that the level sets $\{x|t(x) \leq \gamma\}, \gamma \in \mathbb{R}$, of the function t are compact. Then, the minimax equality

$$\sup_{z\in\mathcal{Z}} \inf_{x\in\mathcal{X}} \psi(x,z) = \inf_{x\in\mathcal{X}} \sup_{z\in\mathcal{Z}} \psi(x,z),$$

holds and the infimum over $\mathcal{X}$ in the right-hand side above is attained at a set of points that is nonempty and compact.

The saddle point theorem above is a variant of the minimax theorem (Bertsekas et al., 2003).

4. Error bound of multiple spectral kernel learning

This chapter present a new geometric bound for the Gaussian complexity of the hypothesis set of the multiple spectral class introduced in Chapter 3. A brief overview of the relation between the complexity and the generalization bound, and also previous upper bounds for the Rademacher complexity of the general MKL hypothesis set are provided. Parts of this chapter are presented in (Reyhani, in print).

4.1 Introduction

As previously mentioned in Section (3.1), in supervised learning, the goal is to find a decision rule $\hat{f}(\boldsymbol{x})$ by fitting a model to the training set $\mathcal{S}_n$. It is natural to ask about an upper bound for the error of prediction for a test sample. Indeed, we are often interested in having some estimate of

$$P\ell \circ \hat{f} = \mathbb{E}\ell(y, \hat{f}(\boldsymbol{x})),$$

where $\hat{f}$ is the solution of the empirical risk minimization and ℓ is a bounded and convex loss function. This error is called generalization error, or risk, and it has been extensively studied in statistics and machine learning, for example (Koltchinskii, 2011; Koltchinskii and Panchenko, 2005; Mendelson, 2003b; Bousquet and Elisseeff, 2002; Bartlett and Mendelson, 2003). Note that, to avoid overfitting we usually assume that f belongs to a small set of measurable functions $\mathcal{F}$. For example, in linear penalized ERM (3.2) the decision rule belongs to the set $\mathcal{F} = \{f_w | f_w(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}, \|\boldsymbol{w}\| \leq \gamma\}$, for a user defined parameter γ. The set $\mathcal{F}$ is called the hypothesis set.

A common characterization of generalization error relates the risk to the empirical risk and the complexity of the hypothesis set. Recall that the empirical risk is

$$P_n\ell \circ f = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(\boldsymbol{x}_i)).$$

For every function $f \in \mathcal{F}$, the risk is smaller than the empirical risk, $P_n\ell \circ f$, plus the supremum of the unweighted empirical process, i.e.

$$P\ell \circ f \leq P_n\ell \circ f + \sup_{h \in \ell \circ \mathcal{F}} (P - P_n)h. \tag{4.1}$$

We assume that $\sup_{f\in\mathcal{F}} P\ell \circ f < \infty$. The supremum term in the inequality (4.1) is bounded and therefore we can replace this random term by the expectation

term

$$\mathbb{E} \sup_{h \in \ell \circ \mathcal{F}} (P - P_n)h,$$

using concentration inequalities. Therefore, with probability $1 - \delta$ and $\forall f \in \mathcal{F}$, we have

$$P\ell \circ f \leq P_n \ell \circ f + \mathbb{E} \sup_{h \in \ell \circ \mathcal{F}} (P - P_n)h + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}.$$

Now, it remains to control the expectation term in the right hand side of the above inequality. As discussed earlier, this expectation depends on the size of $\ell \circ \mathcal{F}$. The most common approach to control this term is to use the bracketing number of $\ell \circ \mathcal{F}$, which is briefly explained in Section 2.3.2. However, by the symmetrization lemma (Theorem 2.3.1) and contraction lemma (Van der Vaart and Wellner, 1996), the expectation term can be replaced by the Rademacher complexity of $\mathcal{F}$ scaled by the Lipschitz norm of ℓ, see for example (Koltchinskii, 2011; Mendelson, 2012).

In the following we bring a result on the generalization bound, which relates the risk to the Gaussian complexity of the hypothesis set and the empirical risk.

Theorem 4.1.1 ((Bartlett and Mendelson, 2003) Corollary 15 and Theorem 8)**.** *Consider a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0,1]$ and a function $\phi : \mathcal{Y} \times \mathcal{Y} \to [0,1]$ that dominates the loss function ℓ, i.e. $\forall y \in \mathcal{Y}$ and $a \in \mathcal{Y}$, $\phi(y,a) \geq \ell(y,a)$. Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathcal{Y}$ and let $\mathcal{S}_n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $(\boldsymbol{x}_i, y_i)$ are independent copies of random vector $(\boldsymbol{x}, y)$. Then for any integer n and any $0 < \delta < 1$, with probability of at least $1 - \delta$, the following holds:*

$$P\ell \circ f \leq P_n \ell \circ f + c|\ell|_L \mathcal{G}_{\mathcal{F}}^n + \left(\frac{8}{n} \ln \frac{2}{\delta}\right)^{\frac{1}{2}} \quad \forall f \in \mathcal{F},$$

where $|\ell|_L$ is the Lipschitz norm of the loss function and $\mathcal{G}_{\mathcal{F}}^n$ denotes the Gaussian complexity of the hypothesis set $\mathcal{F}$.

(Bartlett and Mendelson, 2003) presents similar result where the Gaussian complexity is replaced by the Rademacher complexity, and shows that the following inequality holds under the conditions of the preceding theorem.

$$P\ell \circ f \leq P_n \ell \circ f + c|\ell|_L \mathcal{R}_{\mathcal{F}}^n + \left(\frac{8}{n} \ln \frac{2}{\delta}\right)^{\frac{1}{2}}, \tag{4.2}$$

where $\mathcal{R}_{\mathcal{F}}^n$ denotes the Rademacher complexity of $\mathcal{F}$. For more details on generalization bounds and proof techniques see (Koltchinskii and Panchenko, 2005; Mendelson, 2003a; Bartlett and Mendelson, 2003; Koltchinskii, 2011; Steinwart and Christmann, 2008). Note that, we can replace the complexity term in the above theorem by the empirical complexity at the cost of an additional term that depends on n and some constants.

In the following sections, we first briefly review some previous works on the empirical Rademacher complexity of MKL, and then provide our Gaussian complexity computations in Section 4.3.

4.2 Bounds for complexity of the general MKL

The multiple SKL framework is similar to the MKL, with the exception that all the kernel bases are of rank one. Therefore, we could still use the complexity

computations available for general MKL to estimate the risk of the multiple SKL. In this section, we bring some definitions/computations related to the complexity of the hypothesis set for the general MKL setup as well as previous results on the empirical Rademacher complexity of this set.

Let us denote the hypothesis set in the general multiple kernel learning setup by $\mathcal{F}_{\widetilde{\mathcal{K}}}$, where $\widetilde{\mathcal{K}}$ is a kernel set, e.g. (3.11) or (3.17). It consist of all linear functions of the form $\langle \boldsymbol{w}, \psi(\boldsymbol{x}) \rangle$, where the norms of vectors $\boldsymbol{w}$ are bounded. The inner product is evaluated by the corresponding kernel function that belongs to the kernel class $\widetilde{\mathcal{K}}$. In summary we have,

$$\mathcal{F}_{\widetilde{\mathcal{K}}} := \left\{ f_w \middle| f_w(\boldsymbol{x}) = \langle \boldsymbol{w}, \psi_K(\boldsymbol{x}) \rangle_{\mathcal{H}_K}, \|\boldsymbol{w}\|_{\mathcal{H}_K} \leq \frac{1}{\gamma}, K \in \widetilde{\mathcal{K}} \right\}.$$

Alternatively, the dual representation of the linear decision functions in the preceding display is of the form $\sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x})$ for $\alpha_i \in \mathbb{R}_+, i = 1, \ldots, n$, and k is a kernel function corresponding to $K \in \widetilde{\mathcal{K}}$. The norm constraint is equal to

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\alpha}^\top K \boldsymbol{\alpha} \leq \frac{1}{\gamma^2},$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^\top$ and $[K]_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j), 1 \leq i, j \leq n$. The set $\mathcal{F}_{\widetilde{\mathcal{K}}}$ can be written as

$$\mathcal{F}_{\widetilde{\mathcal{K}}} = \left\{ f_{\boldsymbol{\alpha}} \middle| f_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x}), \boldsymbol{\alpha}^\top K \boldsymbol{\alpha} \leq \frac{1}{\gamma^2}, K \in \widetilde{\mathcal{K}} \right\}.$$

Let $g_1, \ldots, g_n$ be independent standard Gaussian random variables and also define $Z = (g_1, \ldots, g_n)^\top$. We assume that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are independent and identically distributed random vectors, and that the random variable $\boldsymbol{x}_i$ is independent to g_j for all different $1 \leq i, j \leq n$. Conditioning on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, the following holds for the empirical Gaussian complexity:

$$\begin{aligned}
\widehat{\mathcal{G}}^n_{\mathcal{F}_{\widetilde{\mathcal{K}}}} =& \frac{1}{n} \mathbb{E} \left| \sup_{f \in \mathcal{F}_{\widetilde{\mathcal{K}}}} \sum_{i=1}^{n} g_i f(\boldsymbol{x}_i) \right| \qquad (4.3) \\
=& \frac{1}{n} \mathbb{E} \sup_{K \in \widetilde{\mathcal{K}}} \sup_{\|\boldsymbol{w}\|_{\mathcal{H}_K} \leq \frac{1}{\gamma}} \left| \left\langle \boldsymbol{w}, \sum_{i=1}^{n} g_i \psi_K(\boldsymbol{x}_i) \right\rangle_{\mathcal{H}_K} \right| \\
\leq& \frac{1}{n\gamma} \mathbb{E} \sup_{K \in \widetilde{\mathcal{K}}} \left\| \sum_{i=1}^{n} g_i \psi_K(\boldsymbol{x}_i) \right\|_{\mathcal{H}_K} \\
=& \frac{1}{n\gamma} \mathbb{E} \sup_{K \in \widetilde{\mathcal{K}}} \left(Z^\top K Z \right)^{\frac{1}{2}} \\
\leq& \frac{1}{n\gamma} \left(\mathbb{E} \sup_{K \in \widetilde{\mathcal{K}}} Z^\top K Z \right)^{\frac{1}{2}}. \qquad (4.4)
\end{aligned}$$

The third line follows from the definition of the norm. For the last line we used the fact that $\mathbb{E}\{x^{\frac{1}{2}}\} \leq \mathbb{E}\{x\}^{\frac{1}{2}}$ for any nonnegative random variable x.

Similar result holds for the Rademacher complexity, where the random variables g_i are replaced by Rademacher random variables $\epsilon_i, i = 1, \ldots, n$:

$$\widehat{\mathcal{R}}^n_{\mathcal{F}_{\widetilde{\mathcal{K}}}} \leq \frac{1}{n\gamma} \left(\mathbb{E} \sup_{K \in \widetilde{\mathcal{K}}} \boldsymbol{\epsilon}^\top K \boldsymbol{\epsilon} \right)^{\frac{1}{2}}, \qquad (4.5)$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$. Note, that the term $\boldsymbol{\epsilon}^\top K \boldsymbol{\epsilon}$ in (4.5) cannot be greater than $\|\boldsymbol{\epsilon}\|^2 \lambda_1(K)$. Therefore, conditioning on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, we have

$$\begin{aligned}
\mathbb{E} \sup_{K \in \widetilde{\mathcal{K}}} \boldsymbol{\epsilon}^\top K \boldsymbol{\epsilon} &= \mathbb{E} \max_{\boldsymbol{p} \in \Delta_1} \boldsymbol{\epsilon}^\top \left(\sum_{l=1}^{L} p_l K_l \right) \boldsymbol{\epsilon} \\
&\leq \mathbb{E} \max_{1 \leq l \leq L} \boldsymbol{\epsilon}^\top K_l \boldsymbol{\epsilon} \\
&\leq \|\boldsymbol{\epsilon}\|^2 \max_{1 \leq l \leq L} \lambda_1(K_l) \\
&= n \max_{1 \leq l \leq L} \lambda_1(K_l).
\end{aligned}$$

So, the empirical Rademacher complexity can be simply bounded as follows:

$$\widehat{\mathcal{R}}^n_{\mathcal{F}_{\widetilde{\mathcal{K}}}} \leq \frac{1}{\sqrt{n\gamma^2}} \max_{1 \leq l \leq L} \left(\lambda_1(K_l)\right)^{\frac{1}{2}}.$$

This is a simple bound that we can get directly from the definition. This bound is independent to the number of kernel bases and only depends on the spectral information of the kernel bases. In other words, we can insert new kernels to the bases set and the complexity remains the same as long as they do not change the maximum operator norm of the basis matrices in the set.

As it appeared, the main object to work with in complexity computations is either the Gaussian chaos term that is defined by $\mathbb{E} \sup Z^\top K Z$ or the Rademacher chaos which is defined by $\mathbb{E} \sup \boldsymbol{\epsilon}^\top K \boldsymbol{\epsilon}$. (Ledoux and Talagrand, 2011; Van der Vaart and Wellner, 1996; Talagrand, 2005; Mendelson and Paouris, 2012) provide general technologies to bound the Rademacher chaos, using geometric properties of the search space. However, we choose a simpler idea, which is called the decoupling technique, in Section 4.3 to compute a new bound for the empirical Gaussian complexity.

A comprehensive comparison of previous bounds for the Rademacher complexity of MKL can be found in (Cortes et al., 2010). Here, we briefly present results of few previous studies.

The first result is due to (Lanckriet et al., 2004b), which states that:

Theorem 4.2.1 ((Lanckriet et al., 2004b), Theorem 24)**.** *Let us assume that the kernel functions $k_1, \ldots, k_L$ are given. Then, for $c > 0$, the following holds.*

$$\widehat{\mathcal{R}}^n_{\mathcal{F}} \leq \left[\frac{c}{n\gamma^2} \min \left(L, n \max_{1 \leq l \leq L} \frac{\lambda_1(K_l)}{trK_l} \right) \right]^{\frac{1}{2}},$$

where $\lambda_1(K_l)$ denotes the largest eigenvalue of the matrix K_l and $\mathcal{F}$ denotes the hypothesis set of intersection of $\widetilde{\mathcal{K}}$ (defined in (3.11)) and matrices with trace $c > 0$.

Theorem 4.2.1 shows that the empirical Rademacher complexity is bounded by the minimum of $\sqrt{L}$ and the scaled supremum of the spectral ratio of the bases.

Recall that the spectral kernel class is defined by

$$\mathcal{K} := \left\{ K \in \mathbb{R}^{n \times n} : K = \sum_{l=1}^{L} p_l \boldsymbol{b}_l \boldsymbol{b}_l^\top, \boldsymbol{p} \in \Delta_1 \right\}, \tag{4.6}$$

where $\boldsymbol{b}_l$ belongs to the spectral dictionary

$$\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_L : \boldsymbol{b}_i \in \mathbb{R}^n, \boldsymbol{b}_i \neq \boldsymbol{b}_j, 1 \leq i \neq j \leq L\}.$$

In this chapter, we further assume that the norm of elements of dictionary are bounded from above. For the spectral kernel class, the spectral ratio $\lambda_1(K_l)/\mathrm{tr}K_l$ is 1. Thus, the empirical Rademacher complexity by Theorem 4.2.1 is at most

$$\sqrt{\frac{cL}{n\gamma^2}} \qquad \text{for } L < n,$$

otherwise, it is $\sqrt{c/\gamma^2}$. However, for $L < n$ the multiple SKL optimization, for example with ℓ_2 loss is not well defined and including the coordinate bases becomes necessary. So, for the spectral kernel class the above bound is not informative.

(Srebro and Ben-David, 2006) improves the dependency between the Rademacher complexity of the MKL hypothesis set and the number of kernels, and shows that for general loss functions $\widehat{\mathcal{R}}^n_{\mathcal{F}_{\widetilde{\mathcal{K}}}(\gamma)}$ is of order

$$\mathcal{O}\left(\sqrt{\frac{8}{n}\left[2 + L\ln\frac{128n^3eR^2}{\gamma^2 L} + 256\frac{R^2}{\gamma^2}\ln\frac{ne\gamma}{8R}\ln\frac{128nR^2}{\gamma^2}\right]^{\frac{1}{2}}}\right),$$

where $\sup_{\boldsymbol{x}\in\mathcal{X}} k(\boldsymbol{x},\boldsymbol{x}) \leq R^2$ and e is the natural logarithm base number. The above bound depends on the number of bases through the term $L\ln\frac{128n^3eR^2}{\gamma^2 L}$. (Cortes et al., 2010) shows that the bound may become greater than one. In addition, they present a tighter and simpler bound, as stated below.

Theorem 4.2.2. *[(Cortes et al., 2010)] The empirical Rademacher complexity of the hypothesis set $\mathcal{F}_{\widetilde{\mathcal{K}}}$ can be bounded as follows:*

$$\widehat{\mathcal{R}}^n_{\mathcal{F}_{\widetilde{\mathcal{K}}}} \leq \frac{\sqrt{\eta_\circ r\|\boldsymbol{u}\|_r}}{n\gamma}, \quad \forall r \in \mathbb{N}, r \geq 1, \tag{4.7}$$

where $\boldsymbol{u} = (trK_1, \ldots, trK_L)^\top$ and $\eta_\circ = \frac{23}{22}$.

Applying Theorem 4.2.2 to the spectral set requires computing the trace of elements of the spectral kernel class, which is

$$\mathrm{tr}K_l = \sum_{i=1}^n (\boldsymbol{e}_i^\top \boldsymbol{b}_l)^2 = \|\boldsymbol{b}_l\|^2, \quad \forall l = 1, \ldots, L.$$

Then, for any $r \geq 1$, (4.7) implies that

$$\begin{aligned}\widehat{\mathcal{R}}^n_{\mathcal{F}_{\mathcal{K}}} &\leq \frac{\sqrt{\eta_\circ r\|(1,\ldots,1)^\top\|_r \max_{1\leq l\leq L}\|\boldsymbol{b}_l\|^2}}{n\gamma} \\ &= \frac{\sqrt{\eta_\circ r L^{\frac{1}{r}}}}{n\gamma}\max_{1\leq l\leq L}\|\boldsymbol{b}_l\|.\end{aligned}$$

The above bound holds for any $r \geq 1$, so, we search for a value of r for which the squared root term is minimum. The function $r \mapsto rL^{\frac{1}{r}}$ attains its minimum at $r_\circ = \ln L$, thus, we obtain,

$$\widehat{\mathcal{R}}^n_{\mathcal{F}_{\mathcal{K}}} \leq \frac{\sqrt{\eta_\circ e\lceil\ln L\rceil}}{n\gamma}\max_{1\leq l\leq L}\|\boldsymbol{b}_l\|, \tag{4.8}$$

where $\lceil x\rceil$ denotes the smallest following integer to x and we assume that $L \geq 3$. (4.8) shows that the Rademacher complexity of set of rank one kernels can be controlled by the logarithm of the number of bases times the maximum Euclidean norm of bases. (Cortes et al., 2010) also shows similar inequality to (4.8) for general kernel matrices with bounded diagonal entries. There, the max term is replaced by $\max_{1\leq l\leq L}(\mathrm{tr}K_l)^{\frac{1}{2}}$.

4.3 A novel geometric bound for the Gaussian complexity

In this section, we compute a new bound for the empirical Gaussian complexity of the multiple spectral kernel class $\mathcal{K}$ defined in (4.6). Most of the previous results on the empirical complexity of the MKL hypothesis set are about computing the Rademacher complexity. However, a rich set of results about the Gaussian process and Gaussian chaos are available, which could make the complexity's computation easier and perhaps different. This is our main motivation to take a different approach and compute a bound for the empirical Gaussian complexity of the multiple SKL hypothesis set. The technique can be simply applied for computing bounds for the general MKL hypothesis set. Nevertheless, the Gaussian complexity is bounded from below and above by the Rademacher complexity. Let us state our results and leave the proof to Section 4.4.

Theorem 4.3.1. *Let us consider the spectral kernel class defined in (4.6). Then, for sufficiently large L, we have*

$$\widehat{\mathcal{G}}^n_{\mathcal{F}_\mathcal{K}} \leq \frac{1+\sqrt{18\ln L}}{n} \max_{1\leq l\leq L} \|\boldsymbol{b}_l\|\,, \tag{4.9}$$

where $\mathcal{F}_\mathcal{K} := \{f | f(\boldsymbol{x}) = \langle \psi_K(\boldsymbol{x}), \boldsymbol{w}\rangle, \|\boldsymbol{w}\|_2 \leq 1, K \in \mathcal{K}\}$ and $\widehat{\mathcal{G}}^n_{\mathcal{F}_\mathcal{K}}$ is the empirical Gaussian complexity for the hypothesis set $\mathcal{F}_\mathcal{K}$.

In addition, by assuming that $\|\boldsymbol{b}_l\| = 1$, we have,

$$\widehat{\mathcal{G}}^n_{\mathcal{F}_\mathcal{K}} \leq \frac{\sqrt{6\sqrt{2}C\ln L}}{n} \max_{1\leq l,l'\leq L} (1-\boldsymbol{b}_{l'}^\top \boldsymbol{b}_l)^{\frac{1}{4}} + \frac{1}{n}, \tag{4.10}$$

where $C > 0$ is a constant.

The bound presented in (4.9) shows the dependency between the complexity and logarithm of the number of kernel bases in a similar way to the empirical Rademacher complexity bound presented in (4.8). However, we got a larger constant, which is consistent with Theorem 2.3.3.

In addition, (4.10) relates the geometry of the dictionary to the empirical complexity via the term $\max_{l,l'}(1-\boldsymbol{b}_l^\top \boldsymbol{b}_{l'})$. The geometric term counts the similarity between the bases and achieves its maximum when at least two orthogonal vectors are present in the dictionary. This bound suggests that the complexity can also be increased by the angle between the bases, which is an additional information and we can not achieve this conclusion from any of previous bounds. The additional term $\frac{1}{n}$ in both bounds in (4.9) and (4.10) is due to the decoupling technique applied to the Gaussian chaos.

4.4 Proof

In this section we provide the proof of Theorem 4.3.1 that relies on decoupling of the Gaussian chaos, Slepian's lemma, and maximal inequalities, which we introduce shortly.

Lemma 4.4.1 ((Ledoux and Talagrand, 2011), pp. 79, Maximal Gaussian inequality)**.** *Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be a Gaussian random variable in $\mathbb{R}^n$. Then, we have*

$$\mathbb{E} \max_{1\leq i\leq n} x_i \leq 3\sqrt{\ln n} \max_{1\leq i\leq n} \sqrt{\mathbb{E}x_i^2}.$$

This result also holds for sub-Gaussian random variables where the coefficient 3 *is replaced by a constant* $C > 0$.

Lemma 4.4.2 (Slepian's lemma, (Ledoux and Talagrand, 2011) pp. 77 and 79, (Bartlett and Mendelson, 2003))**.** *Let* $x_1, \ldots, x_n$ *be random variables defined by*

$$x_j := \sum_{i=1}^{n} a_{ij} g_i,$$

for $g_i \sim \mathcal{N}(0,1), g_i \perp\!\!\!\perp g_j, \forall 1 \leq i \neq j \leq n$. *Then, there exists a constant* $C > 0$ *such that*

$$\mathbb{E} \max_{1 \leq i \leq n} x_i \leq C\sqrt{\ln n} \max_{1 \leq i,i' \leq n} \sqrt{\mathbb{E}(x_i - x_{i'})^2}.$$

Lemma 4.4.3 (Decoupling of Gaussian quadratic form, (Levina and Vershynin, 2010; De la Peña and Giné, 1999))**.** *Let* $\boldsymbol{z}$ *be a centered normal random vector in* $\mathbb{R}^p$, *and let* $\boldsymbol{z}'$ *be independent copy of random vector* $\boldsymbol{z}$. *Let* $\mathcal{A}$ *be a set of symmetric matrices. Then,*

$$\mathbb{E} \sup_{A \in \mathcal{A}} \left| \langle A\boldsymbol{z}, \boldsymbol{z} \rangle - \mathbb{E} \langle A\boldsymbol{z}, \boldsymbol{z} \rangle \right| \leq 2\mathbb{E} \sup_{A \in \mathcal{A}} \left| \langle A\boldsymbol{z}, \boldsymbol{z}' \rangle \right|.$$

Proof (Theorem 4.3.1)**.** Let $g_1, \ldots, g_n$ denote independent standard normal random variables, that are independent to the samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. We also define $Z := (g_1, \ldots, g_n)^\top$. In deriving the upper bound in (4.4) the set $\widetilde{\mathcal{K}}$ was selected arbitrarily. Thus, for the empirical Gaussian complexity of the spectral kernel class $\mathcal{K}$ with the hypothesis set $\mathcal{F}_{\mathcal{K}}$, we have

$$\begin{aligned}
\widehat{\mathcal{G}}^n_{\mathcal{F}_{\mathcal{K}}} \leq & \frac{1}{n} \left(\mathbb{E} \sup_{K \in \mathcal{K}} Z^\top K Z \right)^{\frac{1}{2}} \\
= & \frac{1}{n} \left(\mathbb{E} \max_{\|\boldsymbol{p}\| \leq 1} Z^\top \Big(\sum_{l=1}^{n} p_l \boldsymbol{b}_l \boldsymbol{b}_l^\top \Big) Z \right)^{\frac{1}{2}} \\
= & \frac{1}{n} \left(\mathbb{E} \max_{1 \leq l \leq L} Z^\top \boldsymbol{b}_l \boldsymbol{b}_l^\top Z \right)^{\frac{1}{2}}. \qquad (4.11)
\end{aligned}$$

Before applying the decoupling lemma we should first remove the mean of $Z^\top \boldsymbol{b}_l \boldsymbol{b}_l^\top Z$, which is equal to

$$\begin{aligned}
\mathbb{E} Z^\top \boldsymbol{b}_l \boldsymbol{b}_l^\top Z &= \sum_{i,j=1}^{n} \mathbb{E} g_i g_j \boldsymbol{e}_i^\top \boldsymbol{b}_l \boldsymbol{b}_l^\top \boldsymbol{e}_j \\
&= \sum_{i=1}^{n} (\boldsymbol{e}_i^\top \boldsymbol{b}_l)^2 = \|\boldsymbol{b}_l\|^2.
\end{aligned}$$

Therefore, from (4.11) we obtain

$$\begin{aligned}
\widehat{\mathcal{G}}^n_{\mathcal{F}_{\mathcal{K}}} \leq & \frac{1}{n} \left(\mathbb{E} \max_{1 \leq l \leq L} Z^\top \boldsymbol{b}_l \boldsymbol{b}_l^\top Z \right)^{\frac{1}{2}} \\
\leq & \frac{1}{n} \left(\mathbb{E} \max_{1 \leq l \leq L} (Z^\top \boldsymbol{b}_l \boldsymbol{b}_l^\top Z - \|\boldsymbol{b}_l\|^2) \right)^{\frac{1}{2}} + \frac{1}{n} \max_{1 \leq l \leq L} \|\boldsymbol{b}_l\| \\
\leq & \frac{1}{n} \left(2\mathbb{E}_Z \mathbb{E}_{Z'} \max_{1 \leq l \leq L} (Z^\top \boldsymbol{b}_l)(\boldsymbol{b}_l^\top Z') \right)^{\frac{1}{2}} + \frac{1}{n} \max_{1 \leq l \leq L} \|\boldsymbol{b}_l\|. \qquad (4.12)
\end{aligned}$$

In the third line we used Lemma 4.4.3. Both terms $Z^\top \boldsymbol{b}_l$ and $\boldsymbol{b}_l^\top Z'$ are Gaussian random variables and can be treated independently. In the following we compute the first term in the right hand side of (4.12) in two different ways:

1°:

$$\begin{aligned}\mathbb{E}_Z \mathbb{E}_{Z'} \max_{1\le l\le L} (Z^\top \boldsymbol{b}_l)(\boldsymbol{b}_l^\top Z') &\le 3\sqrt{\ln L}\mathbb{E}_Z \max_{1\le l\le L} \left((Z^\top \boldsymbol{b}_l)^2 \mathbb{E}_{Z'}(Z'^\top \boldsymbol{b}_l)^2\right)^{\frac{1}{2}} \\ &= 3\sqrt{\ln L}\mathbb{E}_Z \max_{1\le l\le L} \left|Z^\top \boldsymbol{b}_l\right| \|\boldsymbol{b}_l\| \\ &\le 9 \ln L \max_{1\le l\le L} \|\boldsymbol{b}_l\|^2 \end{aligned} \tag{4.13}$$

In the first line we used the Gaussian maximal inequality. We can further assume that $\|\boldsymbol{b}_l\| = 1, \forall l \le L$, and the bound will be reduced to $9 \ln L$.

2°:

$$\begin{aligned}\mathbb{E}_Z \mathbb{E}_{Z'} \max_{1\le l\le L} (Z^\top \boldsymbol{b}_l)(\boldsymbol{b}_l^\top Z') &\le 3\sqrt{\ln L}\mathbb{E}_Z \max_{1\le l\le L} \left((Z^\top \boldsymbol{b}_l)^2 \mathbb{E}_{Z'}(Z'^\top \boldsymbol{b}_l)^2\right)^{\frac{1}{2}} \\ &= 3\sqrt{\ln L}\mathbb{E}_Z \max_{1\le l\le L} \left|Z^\top \boldsymbol{b}_l\right| \|\boldsymbol{b}_l\| \\ &\le 3C \ln L \max_{1\le l,l'\le L} \left(\mathbb{E}(\|\boldsymbol{b}_l\|\boldsymbol{b}_l^\top Z - \|\boldsymbol{b}_{l'}\|\boldsymbol{b}_{l'}^\top Z)^2\right)^{\frac{1}{2}} \\ &= 3C \ln L \max_{1\le l,l'\le L} \left(\|\boldsymbol{b}_l\|^4 + \|\boldsymbol{b}_{l'}\|^4 - 2\|\boldsymbol{b}_l\|\|\boldsymbol{b}_{l'}\|\boldsymbol{b}_l^\top \boldsymbol{b}_{l'}\right)^{\frac{1}{2}}. \end{aligned}$$

In the first line we used the Gaussian maximal inequality and in the third line we used the Slepian's lemma. Using the assumption that norm of $\boldsymbol{b}_l$ is 1, we have

$$\mathbb{E}_Z \mathbb{E}_{Z'} \max_{1\le l\le L} (Z^\top \boldsymbol{b}_l)(\boldsymbol{b}_l^\top Z') \le 3\sqrt{2}C \ln L \max_{1\le l,l'\le L} (1 - \boldsymbol{b}_l^\top \boldsymbol{b}_{l'})^{\frac{1}{2}}. \tag{4.14}$$

By replacing the expectation terms in (4.12) by the upper bounds derived in (4.13) or (4.14), we obtain the claimed results (4.9) or (4.10), respectively. □

5. FastICA and bootstrap FastICA

This chapter contains a short summary about the ICA model, approaches to solve the ICA problem, FastICA, and bootstrap FastICA algorithm. Here, a sample convergence analysis of FastICA algorithm is provided. Parts of this chapter are presented in (Reyhani and Bickel, 2009; Reyhani and Oja, 2011; Reyhani et al., 2011).

5.1 ICA model

Let us assume that $s_1, \ldots, s_p$ are independent random variables with values in $\mathbb{R}$ and some continuous density functions, and that the matrix $\tilde{A} \in \mathbb{R}^{p\times p}$ is a deterministic matrix. The ICA model is defined by

$$\boldsymbol{x} = \tilde{A}\boldsymbol{s},$$

where $\boldsymbol{s} = (s_1, \ldots, s_p)^\top$. The random vector $\boldsymbol{x}$ is called the mixed signals and $s_1, \ldots, s_p$ are called the source signals.

Suppose that random vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are independent copies of $\boldsymbol{x}$ with covariance matrix

$$\Sigma_p = \mathbb{E}(\boldsymbol{x} - \mathbb{E}\boldsymbol{x})(\boldsymbol{x} - \mathbb{E}\boldsymbol{x})^\top.$$

The goal in ICA is to estimate $\tilde{A}$ given $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ without knowing the marginal distribution of $s_i, i \leq p$, (Comon, 1994; Hyvärinen and Oja, 1997; Hyvärinen et al., 2001) among others.

We can transfer the random vectors into isoperimetric position, using transformation $\boldsymbol{z} := \Sigma_p^{-\frac{1}{2}}\boldsymbol{x}$ for population covariance matrix Σ_p, and similarly for samples by defining,

$$\boldsymbol{z}_i = \widehat{\Sigma}_p^{-\frac{1}{2}}\boldsymbol{x}_i, \quad \forall i = 1, \ldots, n,$$

where

$$\widehat{\Sigma}_p := \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top,$$

and $\bar{\boldsymbol{x}} := \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$. We assume that the sample covariance matrix $\widehat{\Sigma}_p$ is invertible.

In the rest of this chapter, we work mainly with the following model

$$\boldsymbol{z} = A\boldsymbol{s}, \quad \text{provided that} \quad AA^\top = I_p,$$

with an orthogonal mixing matrix A. We define the set $\mathcal{S}_n = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\}$, which is called the sample set. It has been shown that ICA model is identifiable up

to rotation and scaling provided that at most one source signal is normally distributed (Comon, 1994).

A common approach for estimating the mixing matrix is to take a set of statistics, let us say $\{M_j\}_{j\geq 0}, M_j \in \mathbb{R}^{p\times p}$, such that all M_j admit decomposition

$$M_j = AD_jA^\top,$$

where $D_j, \forall j$, are diagonal matrices. For example, under the ICA model, fourth cumulant matrices of the random vector $\boldsymbol{z}$ can be decomposed into the mixing matrix and diagonal matrices. The diagonal matrices are cumulant matrices of the corresponding random vector $\boldsymbol{s}$. Also, the Hessian of logarithm of characteristic function, variant of Fisher information matrix, and covariance matrix (in case of time series), can be used to build M_js. All of these estimations generate diagonal matrices for independent source signals.

Within this framework, the mixing matrix can be estimated by finding a matrix A_*, which *simultaneously* factorizes all matrices $M_i, \forall i \geq 1$. The joint factorization can be casted as

$$\begin{aligned} \textit{min} \quad & \sum_{i=1}^{K} \left\| A^\top M_i A - \sum_{j=1}^{p} d_{i,j}\boldsymbol{e}_j\boldsymbol{e}_j^\top \right\|^2 \\ \text{w.r.t.} \quad & A \in \mathbb{R}^{p\times p}, d_{i,j} \in \mathbb{R}, 1 \leq i \leq K, 1 \leq j \leq p \\ \text{s.t.} \quad & AA^\top = I_p. \end{aligned}$$

A major difficulty is to design a fast algorithm for joint matrix factorization as well as finding a set of proper matrices M_i that are sufficiently different in norm. For details on joint matrix factorization approaches see (Pham and Cardoso, 2001; Hyvärinen et al., 2001; Samarov and Tsybakov, 2004; Reyhani and Bickel, 2009; Reyhani and Oja, 2011; Ylipaavalniemi et al., 2012).

5.2 FastICA algorithm

There are other approaches to estimate the mixing matrix that are based on different sets of statistical properties of independent or mixed signals. For instance, by the standard central limit theorem, the distribution of a linear combination of independent random variables is closer to Gaussian, compared to a single random variable. Therefore, we can estimate a separating direction through finding a suitable vector $\boldsymbol{w} \in \mathcal{S}^{p-1}$, which maximizes some measure of non-Gaussianity of the variable $\boldsymbol{w}^\top\boldsymbol{z}$. The constraint $\boldsymbol{w} \in \mathcal{S}^{p-1}$ reduces the solution space, and indicates that the algorithm is searching for a projection matrix. This is the main idea behind the FastICA algorithm (Hyvärinen and Oja, 1997; Hyvärinen and Oja, 2000).

In details, (Hyvärinen and Oja, 1997; Hyvärinen and Oja, 2000) proposes to find a demixing vector $\boldsymbol{w}$ through maximizing the non-Gaussianity of $\boldsymbol{w}^\top\boldsymbol{z}, \boldsymbol{w} \in \mathcal{S}^{p-1}$, where the non-Gaussianity of a random variable x is measured by $\mathbb{E}J(x)$. The function $J(x)$ is defined by

$$J(x) := \left(G(x) - G(\nu)\right)^2,$$

where $G : \mathbb{R} \to \mathbb{R}$ is an arbitrary nonlinear function and $\nu \sim \mathcal{N}(0,1)$. They also assume that all elements of $\boldsymbol{z}$ have unit variance. The demixing or separating

direction can be estimated by maximizing J over all orthogonal directions, with a fixed nonlinear function G, i.e.

$$\widehat{\boldsymbol{w}} = \underset{\boldsymbol{w}\in\mathcal{S}^{p-1}}{\operatorname{argmax}} \ \mathbb{E}J(\boldsymbol{w}^\top \boldsymbol{z}). \tag{5.1}$$

The solution space for preceding optimization problem is $\mathcal{S}^{p-1}$, which is compact. Therefore, the optimization problem has a solution in the quotient space of scaling and unitary rotations. We denote this solution by $\boldsymbol{w}_\circ$.

Furthermore, they proposed to replace the problem defined in (5.1) by

$$\widehat{\boldsymbol{w}} = \underset{\boldsymbol{w}\in\mathcal{S}^{p-1}}{\operatorname{argmax}} \mathbb{E}G\left(\boldsymbol{w}^\top \boldsymbol{z}\right), \tag{5.2}$$

as it has the same optimum, $\widehat{\boldsymbol{w}}$, for functional J, as long as $\boldsymbol{z}$ has unit variance (Hyvärinen et al., 2001). For finite sample $\mathcal{S}_n$, the demixing direction can be estimated as follows:

$$\widehat{\boldsymbol{w}}_n = \underset{\boldsymbol{w}\in\mathcal{S}^{p-1}}{\operatorname{argmax}} \frac{1}{n}\sum_{i=1}^{n} G\left(\boldsymbol{w}^\top \boldsymbol{z}_i\right).$$

Note that, finding the demixing projection by optimizing a Gaussianity-measure falls in Projection Pursuit framework (Huber, 1985), where we search for a desired low dimensional direction, or a projection of the given high dimensional data by optimizing a loss function.

Remark 5.2.1. *Here, the main assumption is that the quantity* $\mathbb{E}G\left(\boldsymbol{w}^\top \boldsymbol{z}\right)$ *exists and is bounded. This assumption may require bounds on moments of random vector* $\boldsymbol{z}$*. For example, for* $G(x) = x^4$*, the criterion function in (5.2) is well defined, provided that the fourth moment of* $\boldsymbol{z}$ *is bounded. In cases such as*

$$G(x) := \log\cosh(x),$$

we may assume samples are distributed over a compact manifold, or all their moments exists and are bounded.

The first order optimality condition of (5.2) reads

$$F := \mathbb{E}\boldsymbol{z}g\left(\boldsymbol{w}_\circ^\top \boldsymbol{z}\right) = 0, \ \boldsymbol{w}_\circ \in \mathcal{S}^{p-1}, \tag{5.3}$$

where $g : \mathbb{R} \to \mathbb{R}$ is the first derivative of function G. One can estimate the demixing matrix, by finding roots or zeros of F or approximately by its empirical estimate:

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i g(\boldsymbol{w}^\top \boldsymbol{z}_i).$$

Indeed, by Newton method we can estimate $\boldsymbol{w}$ iteratively by the fixed point iteration

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - F(\nabla F)^{-1},$$

where the gradient of F is equal to:

$$\begin{aligned}\nabla_{\boldsymbol{w}} F &= \mathbb{E}\boldsymbol{z}\boldsymbol{z}^\top g'\left(\boldsymbol{w}^\top \boldsymbol{z}\right) \\ &\approx \mathbb{E}g'\left(\boldsymbol{w}^\top \boldsymbol{z}\right).\end{aligned}$$

The approximation in second line is proposed in (Hyvärinen and Oja, 1997; Hyvärinen and Oja, 2000). Here, $g' : \mathbb{R} \to \mathbb{R}$ is the second derivative of the function G, which is assumed to be continuous. The approximation is due to the fact that the expectation in the above expression does not generally factorize. However, at solution $\boldsymbol{w}_\circ$, the error of approximation becomes negligible.

By simplifying $\frac{\partial}{\partial \boldsymbol{w}} F$, and plugging this quantity into the Newton iteration, we obtain the fixed point iterations:

$$\boldsymbol{w}(k+1) = \mathbb{E}\left(\boldsymbol{z} g(\boldsymbol{w}(k)^\top \boldsymbol{z}) - \boldsymbol{w}(k) g'(\boldsymbol{w}(k)^\top \boldsymbol{z})\right), \tag{5.4}$$

$$\boldsymbol{w}(k+1) = \frac{\boldsymbol{w}(k+1)}{\|\boldsymbol{w}(k+1)\|}. \tag{5.5}$$

The fixed-point iteration in (5.4), and the normalization in (5.5) constitute the core of FastICA algorithm. The sample estimator consists of similar iterations:

$$\widehat{\boldsymbol{w}}_n(k+1) = \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{z}_i g(\widehat{\boldsymbol{w}}_n(k)^\top \boldsymbol{z}_i) - \widehat{\boldsymbol{w}}_n(k) g'(\widehat{\boldsymbol{w}}_n(k)^\top \boldsymbol{z}_i)\right), \tag{5.6}$$

$$\widehat{\boldsymbol{w}}_n(k+1) = \frac{\widehat{\boldsymbol{w}}_n(k+1)}{\|\widehat{\boldsymbol{w}}_n(k+1)\|}, \tag{5.7}$$

where $\widehat{\boldsymbol{w}}_n(k)$ denotes the estimated value after k iterations using n samples.

From now on, the vector $\widehat{\boldsymbol{w}}_n$ refers to $\boldsymbol{w}_n(k+1)$ for some $k \geq 1$, such that $\|\boldsymbol{w}_n(k+1) - \boldsymbol{w}_n(k)\|_2 \leq \epsilon$, for a fixed small $\epsilon > 0$. This condition corresponds to a stopping criterion that may appear in implementations. Similar notation applies to the population case. In the rest of this chapter and also in the next chapter $s_\circ$ denotes the source corresponding to $\boldsymbol{w}_\circ^\top \boldsymbol{z}$.

Remark 5.2.2 (Stein's identity and FastICA iterations)**.** *The FastICA fixed point iteration can also be derived directly from Stein's identity (Stein, 1956).*

The Stein's identity shows that for a Gaussian vector $\boldsymbol{z}$ with identity covariance matrix and any smooth nonlinear function $h : \mathbb{R}^d \to \mathbb{R}$,

$$\mathbb{E}\boldsymbol{z}h(\boldsymbol{z}) - \nabla h(\boldsymbol{z}) = 0, \tag{5.8}$$

holds. However, for non-Gaussian random vectors the Stein's formula (5.8) is generally nonzero. By choosing

$$h(\boldsymbol{z}) = g(\langle \boldsymbol{w}, \boldsymbol{z} \rangle),$$

with smooth and nonlinear function $g : \mathbb{R} \to \mathbb{R}$, we arrive at the same expression as in (5.4) or (5.6).

It is not difficult to show that a single iteration of (5.8), with fixed function g and initial direction $\boldsymbol{w}$ would point to a non-Gaussian direction (Blanchard et al., 2006). Therefore, a recursive application of (5.8) would point to the most non-Gaussian direction in a few iterations. With the same argument as for FastICA, the most non-Gaussian direction would be the same as a separating direction.

5.2.1 The sample convergence of FastICA

In this section, we present a probabilistic convergence analysis of FastICA. To the best of our knowledge, this type of analysis has not been presented before.

We define the error of the sample FastICA after k iterations by

$$\|\boldsymbol{w}_\circ - \widehat{\boldsymbol{w}}_n(k)\|,$$

which can be bounded from above as follows.

$$\|\boldsymbol{w}_\circ - \widehat{\boldsymbol{w}}_n(k)\| \leq \|\widehat{\boldsymbol{w}}_n(k) - \widehat{\boldsymbol{w}}_n\| + \|\widehat{\boldsymbol{w}}_n - \boldsymbol{w}_\circ\|,$$

where $\widehat{\boldsymbol{w}}_n$ is the fixed point solution for the sample set $\mathcal{S}_n$. Below, we find bounds for each term on the right hand side separately.

Recall that the fixed point iteration in FastICA finds roots of the equation

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i g(\boldsymbol{w}^\top \boldsymbol{z}_i) = 0,$$

using the Newton method. Due to using second order information, i.e. Hessian, this algorithm is expected to converge fast, which agrees to some extent with the experimental results (Hyvärinen et al., 2001). Using Theorem 5.5.1 (see the Appendix), we can compute an upper bound for the distance between the value of the current iteration $\widehat{\boldsymbol{w}}_n(k)$ and the solution $\widehat{\boldsymbol{w}}_n$. Indeed, we have,

$$\|\widehat{\boldsymbol{w}}_n(k) - \widehat{\boldsymbol{w}}_n\| \leq \left(\frac{\|\widehat{\boldsymbol{w}}_n(0) - \widehat{\boldsymbol{w}}_n\|^2 |\widehat{H}_n|_L}{2\lambda}\right)^k, \tag{5.9}$$

where $|\widehat{H}_n|_L$ is the Lipschitz norm of the Hessian matrix at some $\boldsymbol{w} \in \mathcal{S}^{p-1}$, defined by

$$\widehat{H}_n := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i^\top g'\left(\boldsymbol{w}^\top \boldsymbol{z}_i\right),$$

and λ is the smallest eigenvalue of the matrix $\widehat{H}_n$.

In the derivation of FastICA, the random Hessian matrix $\widehat{H}_n$ is approximated by

$$\widehat{H}_n \approx \widehat{\Sigma}_p \frac{1}{n}\sum_{i=1}^{n} g'\left(\boldsymbol{w}^\top \boldsymbol{z}_i\right). \tag{5.10}$$

Using the above approximation we can easily compute Lipschitz norm of the sample Hessian, $|\widehat{H}_n|_L$. Let us denote $\widehat{H}_n$ evaluated at $\boldsymbol{w}_2$ and $\boldsymbol{w}_1$ by $\widehat{H}_n^1$ and $\widehat{H}_n^2$. Then, we have

$$\begin{aligned}
|\widehat{H}_n^2 - \widehat{H}_n^1| &= \left\|\widehat{\Sigma}_p\right\| \left|\frac{1}{n}\sum_{i=1}^{n}\left(g'(\boldsymbol{w}_2^\top \boldsymbol{z}_i) - g'(\boldsymbol{w}_1^\top \boldsymbol{z}_i)\right)\right| \\
&\leq \lambda_1\left(\widehat{\Sigma}_p\right) |g'|_L \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{z}_i\| \, \|\boldsymbol{w}_2 - \boldsymbol{w}_1\| \\
&= \left|\widehat{H}_n\right|_L \|\boldsymbol{w}_2 - \boldsymbol{w}_1\|,
\end{aligned}$$

where $\lambda_1(\widehat{\Sigma}_p)$ denotes the largest eigenvalue of $\widehat{\Sigma}_p \in \mathbb{R}^{p\times p}$. The vectors $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathcal{S}^{p-1}$, are arbitrary points in a small neighbourhood of $\boldsymbol{w}_\circ$. Now, the ratio between elements relating to the Hessian matrix in convergence bound (5.9) becomes

$$\frac{|\widehat{H}_n|_L}{\lambda} \approx \frac{\lambda_1\left(\widehat{\Sigma}_p\right)}{\lambda_p\left(\widehat{\Sigma}_p\right)} \cdot \frac{|g'|_L \sum_{i=1}^{n} \|\boldsymbol{z}_i\|}{n\,|\sum_{i=1}^{n} g'\left(\widehat{\boldsymbol{w}}_n^\top \boldsymbol{z}_i\right)|},$$

where $\lambda_p(\widehat{\Sigma})$ denotes the smallest eigenvalue of $\widehat{\Sigma}_p$. The effect of dimensionality should appear mostly in the approximation of $\widehat{\Sigma}_p$. However, in FastICA, it is

furthermore assumed that the samples are in isotropic position, i.e. $\widehat{\Sigma}_p = I_p$. The condition number term in the above expression becomes

$$\frac{\lambda_1\left(\widehat{\Sigma}_p\right)}{\lambda_p\left(\widehat{\Sigma}_p\right)} = 1.$$

Thus, the increment of dimension may not sharply affect the convergence of the sample FastICA. Note that, having white data is crucial for FastICA algorithm.

For sufficiently large number of samples, we can look at $\widehat{\boldsymbol{w}}_n$ as perturbation of $\boldsymbol{w}_\circ$. So, the initialization of (5.6) with $\boldsymbol{w}_\circ$ should result in $\widehat{\boldsymbol{w}}_n$, as it is argued in (Oja and Yuan, 2006). By initializing both population and sample FastICA iterations with $\boldsymbol{w}_\circ$, we obtain

$$\boldsymbol{w}_\circ = \mathbb{E}\left(\boldsymbol{z}g(\boldsymbol{w}_\circ^\top \boldsymbol{z}) - \boldsymbol{w}_\circ g'(\boldsymbol{w}_\circ^\top \boldsymbol{z})\right),$$

and

$$\widehat{\boldsymbol{w}}_n = \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{z}_i g(\boldsymbol{w}_\circ^\top \boldsymbol{z}_i) - \boldsymbol{w}_\circ g'(\boldsymbol{w}_\circ^\top \boldsymbol{z})\right),$$

assuming that n is sufficiently large. The above equalities imply that the norm difference $\|\boldsymbol{w}_\circ - \widehat{\boldsymbol{w}}_n\|$ depends on the size of the quantity $\|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\xi}_i - \mathbb{E}\boldsymbol{\xi}_i)\|$, where

$$\boldsymbol{\xi}_i := \left(\boldsymbol{z}_i g(\boldsymbol{w}_\circ^\top \boldsymbol{z}_i) - \boldsymbol{w}_\circ g'(\boldsymbol{w}_\circ^\top \boldsymbol{z}_i)\right).$$

By assuming that $|s_i| \leq C_1$, $|g| \leq C_2$, and $|g'| \leq C_3$, we obtain

$$\begin{aligned}\|\boldsymbol{\xi}_i\| &\leq \|A\boldsymbol{s}\||g(s_\circ)| + |g'(s_\circ)| \\ &\leq \sqrt{p}C_1C_2 + C_3.\end{aligned}$$

Therefore, by the Hoeffding's concentration inequality, the following probabilistic bound holds:

$$P\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\xi}_i - \mathbb{E}\boldsymbol{\xi}_i)\right\| \geq (\sqrt{p}C_1C_2 + C_3)\left(\sqrt{\frac{2\tau}{n}} + \sqrt{\frac{1}{n}} + \frac{4\tau}{3n}\right)\right\} \leq \exp(-\tau).$$

Therefore, with probability $1 - \exp(-\tau)$, $\tau > 0$ the following holds:

$$\|\boldsymbol{w}_\circ - \widehat{\boldsymbol{w}}_n\| \leq (\sqrt{p}C_1C_2 + C_3)\left(\sqrt{\frac{2\tau}{n}} + \sqrt{\frac{1}{n}} + \frac{4\tau}{3n}\right). \tag{5.11}$$

5.3 Bootstrap FastICA

5.3.1 Bootstrap

In most estimation problems, it is required to provide some estimation of the error or accuracy (Lehmann and Casella, 1998; Lehmann and Romano, 2005; Davison, 2003). Usually, the bias and variance are good indicators for this purpose. However, a more accurate indicator is the confidence interval. Let $\widehat{\theta}$ denotes the

estimator of the parameter θ for observations drawn according to P. The distribution of $\theta - \widehat{\theta}$ contains all the information we need for assessing the accuracy of θ. Indeed, conditioned on P, we have

$$P\left\{\widehat{\theta} - \xi_\beta \widehat{\sigma} \leq \theta \leq \widehat{\theta} - \xi_{1-\alpha}\widehat{\sigma}\right\} \geq 1 - \beta - \alpha,$$

where $\widehat{\sigma}$ typically is an empirical estimation of the standard deviation of $\widehat{\theta}$ and ξ_α is the upper α-quantile of the distribution of $\frac{\theta - \widehat{\theta}}{\widehat{\sigma}}$. The quantiles and the distribution of $\widehat{\theta} - \theta$ usually depend on P, and therefore, only observations alone are not enough to asses the accuracy of the estimator in terms of the confidence interval.

In the case that the distribution of $\frac{\widehat{\theta} - \theta}{\widehat{\sigma}}$ tends to a normal random variable, then, we can estimate the distribution of $\widehat{\theta} - \theta$ by a normal distribution with zero mean and variance $\widehat{\sigma}^2$. Therefore, an asymptotic confidence interval of level $1 - \alpha - \beta$ is

$$\left[\widehat{\theta} - z_\beta \widehat{\sigma}, \widehat{\theta} - z_{1-\alpha}\widehat{\sigma}\right],$$

where z_β is normal β-quantiles.

The bootstrap idea is to replace the distribution P in the above computations by $\widehat{P}$, which is estimated using the observations (Efron, 1979; El-Sherief and Sinha, 1979; Shao, 1990). For example, the empirical distribution can be used to estimate P. The distribution $\widehat{\theta} - \theta$ is a function of P, which can be then estimated using $\widehat{P}$. Let $\widehat{\theta}^*$ and $\widehat{\sigma}^*$ denote the estimations computed from observations that are drawn according to $\widehat{P}$ in the same way $\widehat{\theta}$ and $\widehat{\sigma}$ are computed from the original observations. Therefore, the bootstrap estimator for the distribution of $\frac{\widehat{\theta} - \theta}{\widehat{\sigma}}$ conditioned on P is the distribution of $\frac{\widehat{\theta}^* - \widehat{\theta}}{\widehat{\sigma}^*}$ conditioned on $\widehat{P}$. Here, we look at $\widehat{\theta}$ as a nonrandom variable. To compute the confidence interval we should compute the bootstrap quantiles.

A bootstrap estimator for a quantile ξ_α of $\frac{\widehat{\theta} - \theta}{\widehat{\sigma}}$ is a quantile of the distribution of $\frac{\widehat{\theta}^* - \widehat{\theta}}{\widehat{\sigma}^*}$ conditioned on $\widehat{P}$, which is the smallest value $x = \widehat{\xi}_\beta$ such that

$$P\left\{\frac{\widehat{\theta}^* - \widehat{\theta}}{\widehat{\sigma}^*} \leq x | \widehat{P}\right\} \geq 1 - \beta.$$

Then, the bootstrap confidence interval with asymptotic level $1 - \alpha - \beta$ is

$$\left[\widehat{\theta} - \widehat{\xi}_\beta \widehat{\sigma}, \widehat{\theta} - \widehat{\xi}_{1-\alpha}\widehat{\sigma}\right] = \left\{\theta : \widehat{\xi}_{1-\alpha} \leq \frac{\widehat{\theta} - \theta}{\widehat{\sigma}} \leq \widehat{\xi}_\beta\right\}.$$

In the above interval computation, we used the fact that, if $\widehat{P}$ is close to P, then the bootstrap quantiles should be close to the true quantiles, implying that

$$P\left\{\frac{\widehat{\theta} - \theta}{\widehat{\sigma}} \leq \widehat{\xi}_\beta | P\right\} \approx 1 - \beta.$$

For further details on the theory of the bootstrap, see (Van der Vaart and Wellner, 1996; Van der Vaart, 1998; Shao, 1990). In the rest of this section, we shortly explain simple bootstrap sampling.

For any set of samples $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, let us define bootstrap samples $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_n^*$, that are independently and uniformly drawn from the empirical distribution, with a given policy, e.g. with or without replacement. One can simulate the sampling policy by introducing a set of exchangeable random variables $D_{n1}, \ldots, D_{nn}$

that are called *random weights* into the empirical measure. The result is called *weighted bootstrap measure*, which is defined by

$$P_n^* := \frac{1}{n}\sum_{i=1}^{n} D_{ni}\delta_{\boldsymbol{x}_i}.$$

A general requirement is that the weights are exchangeable and have bounded variance. The exchangeability is to guarantee that there is no decision behind taking a particular sample. The requirements are listed below. For further details see (Van der Vaart and Wellner, 1996)-Condition 3.6.8.

B1. The vector $D = (D_{n1}, \ldots, D_{nn})^\top$ is exchangeable for all $n = 1, 2, \ldots$, i.e. the joint probability distribution of any permutation of $D_{n1}, \ldots, D_{nn}$ is the same as the original sequence,

B2. $D_{ni} \geq 0, \forall i$ and n, and $\sum_{i=1}^{n} D_{ni} = n$,

B3. For some positive constant $B < \infty$, $\int_0^\infty \sqrt{P_D\{D_{n1} > u\}}du \leq B$,

B4. $\lim_{\lambda\to\infty} \limsup_{n\to\infty} \sup_{t\geq\lambda} t^2 P_D\{D_{n1} > t\} = 0$

B5. $\frac{1}{n}\sum_{i=1}^{n}(D_{ni} - 1)^2 \xrightarrow{P_D} c^2 > 0$, for some $c > 0$.

P_D denotes the probability with respect to random vector D. For brevity, we drop the term *weighted* from the bootstrap. Two practical examples for the weight random variables are given as follows.

Example 5.3.1. One important example for weight vectors, D, satisfying conditions B.1-B.5 is the multinomial vector $(D_{n1}, \ldots, D_{nn})$, with parameters n and probabilities $\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$. This set of weights represents Efron's empirical bootstrap with replacement.

Example 5.3.2. Another example is the set of weight vectors $(D_{n1}, \ldots, D_{nn})$, built as a row of k times the number $n(n-k)^{-\frac{1}{2}}k^{-\frac{1}{2}}$ and $n-k$ times 0, randomly ordered and independent to samples. These weights correspond to bootstrap without replacement, or k-out-of-n bootstrap. For technical details, see (Van der Vaart and Wellner, 1996), Example 3.6.14.

5.3.2 Bootstrap FastICA

The sample FastICA is sensitive to the initial conditions and the sample set. The results may fluctuate either by subsampling or by changing the initialization. This phenomenon can be frequently observed in applications with sufficiently large number of dimensions. In practice, one might use different randomizations by changing the initial values or subsampling to have a robust estimation of the mixing matrix.

The bootstrap FastICA can be obtained by replacing the samples by the bootstrap samples:

$$\begin{aligned}
\widehat{\boldsymbol{w}}_n^*(k+1) &= \frac{1}{k}\sum_{i=1}^{n}\left(\boldsymbol{z}_i^* g(\widehat{\boldsymbol{w}}_n^{*\top}(k)\boldsymbol{z}_i^*) - \widehat{\boldsymbol{w}}_n^{*\top}(k) g'(\widehat{\boldsymbol{w}}_n^{*\top}(k)\boldsymbol{z}_i^*)\right), \\
\widehat{\boldsymbol{w}}_n^*(k+1) &= \frac{\widehat{\boldsymbol{w}}_n^*(k+1)}{\|\widehat{\boldsymbol{w}}_n^*(k+1)\|},
\end{aligned} \tag{5.12}$$

where we denote separating direction with bootstrapping at iteration k by $\widehat{\boldsymbol{w}}_n^*(k)$.

The bootstrap FastICA, at each bootstrapping trial, may provide slightly different results; therefore, an additional step is required to group the results from different bootstrap trials that are indicating to the same direction.

A bootstrap FastICA algorithm that is an extension of a similar algorithm in (Ylipaavalniemi and Soppela, 2009; Reyhani et al., 2011) is summarized in Algorithm 2. The method *FastICA* in the algorithm takes an optional initialization and a sample set, and it returns source separating directions by solving iterations (5.6, 5.7) or (5.12). The method *bootstrap* returns a bootstrapped sample set based on a weight vector as explained in, for instance, Example 5.3.2.

Algorithm 2 Bootstrap FastICA

Input: samples $\mathcal{S}_n$ and the number of bootstrap trials R
Output: Clusters
$\widehat{\Sigma} \leftarrow \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top$ and $\bar{\boldsymbol{x}} \leftarrow \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$
$\mathcal{S}'_n \leftarrow \{\widehat{\Sigma}^{-\frac{1}{2}}(\boldsymbol{x}_1 - \bar{\boldsymbol{x}}), \ldots, \widehat{\Sigma}^{-\frac{1}{2}}(\boldsymbol{x}_n - \bar{\boldsymbol{x}})\}$
$W(0) \leftarrow$FastICA$(\mathcal{S}'_n)$
for $i = 1, \ldots, R$ **do**
  $\mathcal{S}_n(i) \leftarrow$Bootstrap$(\mathcal{S}'_n)$
  $W(i) \leftarrow$FastICA$(\mathcal{S}_n(i); W(0))$
end for
$W \leftarrow (W(1), \ldots, W(R))$
$C \leftarrow$Corr(W)
Cluster$(W(1), \ldots, W(R); C)$

The method *cluster* in Algorithm 2 returns a grouping on all different directions stored in $W(1), \ldots, W(R)$, for a given correlation matrix. Any clustering algorithm can be used for the grouping here. The correlation matrix provides distance information between different directions. Empirical results show that it might be more suitable if we first produce a binary matrix out of the correlation matrix by a threshold. A threshold between 0.97 and 0.99 is suitable for many applications. Then, the clustering is performed using this binary matrix as distance information. By ranking the clusters based on the number of estimations appeared in each cluster we can obtain a suitable aggregation.

5.4 Extensions of the ICA model

In this section, we briefly explain some extensions of the ICA model. In the standard ICA model, the component or source signals are univariate. One possible extension is to assume that each component is multivariate, and that the multivariate components are independent. In other words, we have,

$$\boldsymbol{x} = A\boldsymbol{s}, \quad \boldsymbol{s} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_d), \boldsymbol{s}_1 \in \mathbb{R}^{p_1}, \ldots, \boldsymbol{s}_d \in R^{p_d},$$

and $\boldsymbol{s}_i \perp\!\!\!\perp \boldsymbol{s}_j, 1 \leq i \neq j \leq d$. $p_1, \ldots, p_d$ denote the dimension of each component. This model is called Independent Subspace Analysis (ISA).

Some of the ICA methods such as joint matrix factorization, which are explained in the beginning of section 5.1, can be adapted to handle multivariate situations with the exception that the diagonal matrix in the matrix decomposition part may be replaced by a block diagonal matrix.

Another extension of ICA is called non-Gaussian component analysis (NGCA), which is also a special case of ISA model. NGCA assumes that only two subspaces exist. One of them, usually the one with higher dimension, is assumed to be a pure Gaussian signal, and the other one is a multivariate component. The NGCA model is originally proposed in (Kawanabe et al., 2006), and is mainly used for noise reduction purposes.

In NGCA, we have

$$\boldsymbol{x} = A_N \tilde{\boldsymbol{s}}_N + \boldsymbol{n}, \quad \boldsymbol{n} \sim \mathcal{N}(0, \Sigma_p).$$

This model can be recasted as

$$\boldsymbol{x} = A_N \boldsymbol{s}_N + A_G \boldsymbol{s}_G,$$

where s_N contains both the non-Gaussian signal $\tilde{\boldsymbol{s}}_N$ and part of $\boldsymbol{n}$ that is on the range of A_N. Random vector $\boldsymbol{s}_G$ contains pure Gaussian that is on kernel space of A_N. This model is identifiable as we look only for the non-Gaussian subspace, i.e. a projection matrix.

The joint factorization methods are also applicable to this model. The main difference compared to the case of ICA model is that the matrix D_i is block diagonal, and consists of two blocks. The first one is an arbitrary matrix, whereas the second one is at most a diagonal matrix for whitened samples, which means that the diagonal elements can be zero, for example in cumulant matrices. The block with lower dimension corresponds to non-Gaussian signals. For details see (Kawanabe et al., 2006; Blanchard et al., 2006; Sugiyama et al., 2006; Kawanabe et al., 2007; Reyhani and Oja, 2011).

5.5 Appendix

Theorem 5.5.1 (Speed of convergence of the Newton method (Ruszczyński, 2006), Theorem 5.13)**.** *Assume function f is twice continuously differentiable, and its Hessian is positive definitive at all $\boldsymbol{\theta}$ in the set $\Theta_\circ = \{\boldsymbol{\theta} : f(\boldsymbol{\theta}) \leq f(\boldsymbol{\theta}_\circ)\}$. Assume $\Theta_\circ$ is bounded, and $\{\boldsymbol{\theta}_k\}_{k\geq 1}$ is generated by the Newton's iteration. Then, $\{\boldsymbol{\theta}_k\}$ is convergent to a minimum $\boldsymbol{\theta}_\circ$ of f. The rate of convergence is*

$$\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_\circ\| \leq \frac{L}{2\lambda} \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_\circ\|^2,$$

where λ is the smallest singular value of the Hessian matrix, and L is the Lipschitz norm of the Hessian.

Another computation for the sample convergence of FastICA

Without the approximation (5.10) in computing the Hessian matrix, we can still compute the elements required for computing the convergence rate of the FastICA iteration, i.e. the Lipschitz norm of the Hessian and the smallest eigenvalue of the sample Hessian.

Using similar notation as in 5.2.1, for the Lipschitz norm we have,

$$\begin{aligned}\left\|\widehat{H}_n^2 - \widehat{H}_n^1\right\| &= \left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^\top g'(\boldsymbol{w}_2^\top \boldsymbol{z}_i) - \frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^\top g'(\boldsymbol{w}_1^\top \boldsymbol{z}_i)\right\| \\ &\leq \left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^\top |g_L'|\|\boldsymbol{z}_i\|\right\| \|\boldsymbol{w}_2 - \boldsymbol{w}_1\| \\ &\leq |g'|_L C_1\sqrt{p}\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^\top\right\| \|\boldsymbol{w}_2 - \boldsymbol{w}_1\| \\ &= |\widehat{H}_n|_L\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|.\end{aligned}$$

Similar to previous computations, we assume that $\|\boldsymbol{s}\| \leq C_1\sqrt{p}$. Thus, we obtain,

$$|\widehat{H}_n|_L \leq p\sqrt{p}C_1\lambda_1(\widehat{\Sigma}_p).$$

At the solution $\boldsymbol{w}_\circ$ the population Hessian matrix is $H = \mathbb{E}\boldsymbol{z}\boldsymbol{z}^\top g'(\boldsymbol{w}_\circ^\top \boldsymbol{z})$, which can be decomposed into

$$H = AOA^\top,$$

where O is a diagonal matrix with $p-1$ entries equal to $\mathbb{E}g'(s_\circ)$ and one entry equal to $\mathbb{E}s_\circ^2 g'(s_\circ)$. Indeed, we have

$$\begin{aligned}H &= \mathbb{E}\boldsymbol{z}\boldsymbol{z}^\top g'(\boldsymbol{w}_\circ^\top \boldsymbol{z}) \\ &= A\mathbb{E}\boldsymbol{s}\boldsymbol{s}^\top g'(s_\circ)A^\top \\ &= AOA^\top,\end{aligned}$$

where $[O]_{i,j} = \mathbb{E}s_i s_j g'(s_\circ)$, which is

$$[O]_{i,j} = \begin{cases} 0 & i \neq j \\ \mathbb{E}g'(s_\circ) & i = j \neq \circ \\ \mathbb{E}s_\circ^2 g'(s_\circ) & i = j = \circ \end{cases}.$$

Here, "$\circ$" denotes the index of the source signal which corresponds to the vector $\boldsymbol{w}_\circ$. The smallest eigenvalue of H, i.e. $\lambda_p(H)$, is $\min\left\{\mathbb{E}g'(s_\circ), \mathbb{E}s_\circ^2 g'(s_\circ)\right\}$. Similarly, the sample Hessian matrix at $\boldsymbol{w}_\circ$ admits the decomposition

$$\widehat{H}_n = A\widehat{O}_n A^\top,$$

where, for all $1 \leq k, l \leq p$

$$[\widehat{O}_n]_{k,l} = \frac{1}{n}\sum_{i=1}^n s_{k,i}s_{l,i}g'(s_{\circ,i}).$$

In above $s_{k,i}$ denotes the i-th sample of random variable s_k where $i = 1, \ldots, n$ and $k = 1, \ldots, p$.

In sample case the matrix $\widehat{O}_n$ is not a diagonal matrix for n small. For bounded source signals and $|g'| \leq C_2$, the entries of $\widehat{O}_n$ are bounded and they are concentrated around their mean. By Hoeffding's inequality, with probability $1 - 2\exp(-\tau)$ for $\tau > 0$, we have

$$|[\widehat{O}_n]_{i,j} - [O]_{i,j}| \leq C_1^2 C_2\sqrt{\frac{\tau}{2n}}. \tag{5.13}$$

Additionally, by Weyl's matrix perturbation inequality (Bhatia, 1997), we have

$$|\lambda_i(\widehat{O}_n) - \lambda_i(O)| \leq \|E_n\|,$$

where the norm is the operator norm and $E_n = \widehat{O}_n - O$. In addition, the operator norm is bounded by the trace norm, and we have

$$\|E_n\| \leq \|E_n\|_{\text{fro}} \leq \left(\sum_{i,j=1}^{p} [E_n]_{i,j}^2 \right)^{\frac{1}{2}} \tag{5.14}$$

By combining (5.13) and (5.14), we have, with probability $1 - 2\exp(-\tau)$

$$|\lambda_i(\widehat{O}_n) - \lambda_i(O)| \leq pC_1^2C_2\sqrt{\frac{\tau}{2n}} \quad \forall 1 \leq i \leq p.$$

Therefore, the ratio L/λ in Theorem 5.5.1, with probability $1 - 2\exp(-\tau), \tau > 0$, is bounded by

$$\frac{p\sqrt{p}C_1\lambda_1(\widehat{\Sigma}_p)}{\min\left\{\left|\lambda_p(H) \pm pC_1^2C_2\sqrt{\frac{\tau}{2n}}\right|\right\}}.$$

6. Statistical analysis of FastICA and bootstrap FastICA

This chapter provides a consistency and asymptotic normality of FastICA estimation. Similar results for bootstrap FastICA are also provided. These results can be used, for example, to derive a statistical test of the convergence of these ICA algorithms. Parts of this chapter are presented in (Reyhani et al., 2011; Ylipaavalniemi et al., 2012).

6.1 Introduction

FastICA algorithm has been widely used in source separation applications due to its speed of convergence and the accuracy, for example (Hyvärinen and Oja, 1997; Hyvärinen et al., 2001; Comon and Jutten, 2010; Suzuki and Sugiyama, 2011). Some of the statistical properties of FastICA algorithm have been studied too. For example, (Hyvärinen, 1999) presents a population analysis showing that the fixed point algorithm converges quadratically with any continuous differentiable nonlinear function. In their proof it is required that the nonlinear function has up to the fourth-order bounded derivatives and source signals have bounded fourth moment. The proof relies on the assumption that the algorithm is initialized by the true solution. The result in (Hyvärinen, 1999) does not provide any sample analysis.

Moreover, (Oja and Yuan, 2006) shows that the fixed point algorithm is stable in the presence of small-norm perturbations of the true direction. (Tichavsky et al., 2006) shows that the FastICA fixed-point iteration is asymptotically normal, if the iteration converges in a single step. Single step convergence again requires that the algorithm is initialized by the true solution. The setup can be relaxed to some extent using the convergence under perturbation results in (Oja and Yuan, 2006).

In summary, the previous works mainly showed that, the FastICA population fixed-point iteration finds a separating direction if it is initialized within a small neighborhood of the true solution. The main contribution of this chapter is to establish consistency and asymptotic normality of FastICA and bootstrap FastICA. To show these results we borrow some techniques from M-/Z- estimation theory, which is popular in mathematical statistics.

In the rest of this chapter, we first introduce M-and Z-estimators, the related consistency and asymptotic convergence results. We then present our theoretical and numerical results.

6.2 M- and Z-estimator

A significant number of statistical estimators are defined through maximizing or minimizing a random criterion function over a finite dimensional Euclidean subset. This type of estimator is called M-estimator (Van der Vaart and Wellner, 1996; Van der Vaart, 1998). Empirical risk minimization (ERM) (see Section 3.1) and maximum likelihood estimator are the most popular examples of M-estimators (Shao, 1990; Lehmann and Casella, 1998; Davison, 2003; Koltchinskii, 2011).

Let us assume that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ with values in $\mathcal{X} \subseteq \mathbb{R}^p$ are independent and identically distributed samples. We consider the problem of estimating parameter $\theta_\circ \in \Theta \subset \mathbb{R}^p$, by maximizing (or minimizing) the functional

$$Pm_\theta = \int_{\mathcal{X}} m_\theta(\boldsymbol{x}) dP(\boldsymbol{x}).$$

We assume that the function $m_\theta : \mathcal{X} \times \Theta \to \mathbb{R}$ is known.

An M-estimator is defined by

$$\widehat{\theta} := \underset{\theta \in \Theta}{\operatorname{argmax}}\, M(\theta) := Pm_\theta, \tag{6.1}$$

and similarly the sample M-estimator is defined by

$$\widehat{\theta}_n := \underset{\theta \in \Theta}{\operatorname{argmax}}\, M_n := P_n m_\theta, \tag{6.2}$$

where $P_n m_\theta = \frac{1}{n} \sum_{i=1}^n m_\theta(\boldsymbol{x}_i)$. We may replace $\widehat{\theta}$ and $\widehat{\theta}_n$ by sets of solutions when the solutions are not unique.

There are situations where a statistical estimator is defined by the root(s) of a system of equations:

$$\widehat{\theta} := \left\{\theta : \left\|P\psi_\theta\right\| = 0, \theta \in \Theta\right\}, \tag{6.3}$$

where the norm in the above expression is a proper norm and $\psi_\theta : \Theta \times \mathcal{X} \to \mathbb{R}^p$, for a finite integer p. This estimator is called Z-estimator. The equations involved in defining a Z-estimator, i.e. (6.3), often can be seen as the optimality condition of some optimization problem, thus, M and Z-estimators are strongly related.

A sample Z-estimator is an estimator $\widehat{\theta}_n$, which makes the score function approximately zero, i.e.

$$\widehat{\theta}_n := \left\{\theta : \left\|P_n\psi_\theta\right\| = o_P\left(n^{-\frac{1}{2}}\right), \theta \in \Theta\right\}. \tag{6.4}$$

The notation $o_P(1)$ denotes the convergence to zero in P-probability.

Example 6.2.1 (FastICA is an M-estimator)**.** FastICA can be seen both as an M-estimator and a Z-estimator. Indeed, a demixing direction estimation is characterized by

$$\widehat{\boldsymbol{w}}_n \in \left\{\boldsymbol{w} : \|P_n f_{\boldsymbol{w}}\| = o_P(n^{-1/2}), \boldsymbol{w} \in \mathcal{S}^{p-1}\right\}, \tag{6.5}$$

where $f_{\boldsymbol{w}}(\boldsymbol{z}) = \boldsymbol{z}g(\boldsymbol{w}^\top \boldsymbol{z})$. Alternatively, a demixing direction can be defined as an optimization problem:

$$\widehat{\boldsymbol{w}}_n = \underset{\boldsymbol{w} \in \mathcal{S}^{p-1}}{\operatorname{argmax}}\, P_n G_{\boldsymbol{w}}, \tag{6.6}$$

where $G_{\boldsymbol{w}} = G(\boldsymbol{w}^\top \boldsymbol{z})$. In above, function g is the derivative of nonlinear function $G : \mathbb{R} \to \mathbb{R}$.

Statistical properties such as rate of convergence and asymptotic normality of the M- and Z-estimators are extensively studied in statistical literature, for example see (Van der Vaart and Wellner, 1996; Van der Vaart, 1998; Van de Geer, 2000).

One of the tools for showing the asymptotic distribution of an M-estimator is the *argmax mapping lemma* (Van der Vaart and Wellner, 1996). The argmax lemma states that the convergence in distribution of a random criterion function would imply the convergence in distribution of the point of maximum to the point of maximum in the limit. This property holds as long as the limit function has a well-separated maximum. The well separated condition means that the criterion function at the maximum should be strictly greater than any point of its neighborhood. The proof of argmax lemma relies on the continuous mapping theorem, for details see (Van der Vaart and Wellner, 1996).

The classical approach to prove the asymptotic normality of M-estimator is through Taylor expansion of the criterion function around the true solution (Van der Vaart, 1998; Van de Geer, 2000), which is sometimes called *linearization technique*. Similar techniques have been applied to prove convergences of Z-estimator.

In the following, we bring two theorems about the consistency, i.e. convergence in probability to the true solution, and convergence in distribution of M-estimator, which we will use later to establish statistical properties of FastICA. For a comprehensive treatment on this topic see, for example, (Van der Vaart and Wellner, 1996; Van der Vaart, 1998; Van de Geer, 2000).

Theorem 6.2.2 (Consistency of M-estimator, (Van der Vaart and Wellner, 1996), Theorem 3.3.7, or (Van der Vaart, 1998) Theorem 5.7)**.** *Let $\widehat{\theta}_n$, θ, M_n and M be as defined in (6.1) and (6.2). In addition, let us assume that the following conditions hold:*

$$\textbf{Condition I.} \quad \sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{P} 0,$$

$$\textbf{Condition II.} \quad \sup_{\theta: d(\theta,\theta_\circ)\geq\varepsilon} M(\theta) < M(\theta_\circ),$$

for some $\epsilon > 0$. Then, the sequence of estimators $\widehat{\theta}_n$ converges in probability to $\theta_\circ$ if $M_n(\widehat{\theta}_n) \geq M_n(\theta_\circ) - o_P(1)$.

The convergence in distribution requires some extra conditions in addition to the conditions I and II. These extra conditions are:

Condition III. Suppose the function m_θ is differentiable in quadratic mean at $\theta_\circ$, or m_θ has Frechet derivative, i.e. there exists a function $\dot{m}_\theta : \Theta \to \mathbb{R}^p$, with components in $L_2(P)$, such that

$$\left\| m_\theta - m_{\theta_\circ} - (\theta - \theta_\circ)^\top \dot{m}_\theta \right\| = o\left(\|\theta - \theta_\circ\|\right).$$

Condition IV. For $\theta \in \Theta$, $\|\theta - \theta_\circ\| \leq \delta$ with $\delta > 0$ small, the difference $P(m_\theta - m_{\theta_\circ})$ can be well approximated by a quadratic form, i.e. there exists a positive definite matrix $V_{\theta_\circ}$, such that

$$P(m_\theta - m_{\theta_\circ}) = \frac{1}{2}(\theta - \theta_\circ)^\top V_{\theta_\circ}(\theta - \theta_\circ) + o(\|\theta - \theta_\circ\|^2).$$

Condition V. The class $\mathcal{F}_\theta = \left\{ f_\theta : f_\theta = \frac{m_\theta - m_{\theta_\circ}}{\|\theta - \theta_\circ\|}, 0 < \|\theta - \theta_\circ\| \leq \delta \right\} \cup \{0\}$ is P-Donsker.

Lemma 6.2.3 (Asymptotic normality of the M-estimator, (Van de Geer, 2000)-Theorem 12.6)**.** *Let $\widehat{\theta}_n$, θ, M_n and M be as above and assume that conditions III, IV, and V hold. Moreover, assume that $\widehat{\theta}_n$ is a consistent estimator of $\theta_\circ$. Then,*

$$\sqrt{n}(\widehat{\theta}_n - \theta_\circ) \rightsquigarrow \mathcal{N}(0, \Sigma),$$

where $\Sigma := V_{\theta_\circ}^{-1} U_{\theta_\circ} V_{\theta_\circ}^{-1}$, where $U_{\theta_\circ} := P\dot{m}_{\theta_\circ}\dot{m}_{\theta_\circ}^\top$ is a non-singular and bounded matrix.

Remark 6.2.4. *With the above theorem, given any consistent estimator $\widehat{\Sigma}$ of Σ, we have $\sqrt{n}\widehat{\Sigma}^{-\frac{1}{2}}(\widehat{\theta}_n - \theta_\circ) \rightsquigarrow \mathcal{N}(0, I_d)$.*

6.2.1 Bootstrap Z-estimator

A bootstrap Z-estimator, $\widehat{\theta}_n^*$, is an estimator that makes the score functions approximately zero with respect to the product measure, i.e.

$$\widehat{\theta}_n^* := \left\{\theta : \|P_n^* \psi_\theta\| = o_{P_{XD}}\left(n^{-\frac{1}{2}}\right)\right\},$$

where

$$P_n^* := \frac{1}{n}\sum_{i=1}^n D_{ni}\delta_{x_i},$$

and ψ_θ is the estimating function (Wellner and Zhan, 1996; Cheng and Huang, 2010). Here, $P_{XD} = P_X \times P_D$. $D_{n1}, \ldots, D_{nn}$ is a set of exchangeable random variables defined in Section 5.3 that satisfy the technical conditions enlisted in the same section.

Similarly, the bootstrap M-estimator, $\widehat{\theta}_n^*$, is defined by maximizing the functional $P_n^* m_\theta$, i.e.

$$\widehat{\theta}_n^* =: \operatorname*{argmax}_{\theta\in\Theta} P_n^* m_\theta,$$

for some criterion function m_θ.

Example 6.2.5. *For a set of random weights $D_{n1}, \ldots, D_{nn}$ that satisfy conditions B.1-B.5 in pp. 48 the bootstrap FastICA can be obtained by replacing the samples by the bootstrap samples. In other words, the bootstrap FastICA is defined by*

$$\widehat{\boldsymbol{w}}_n^* = \operatorname*{argmax}_{\boldsymbol{w}\in\mathcal{S}^{p-1}} P_n^* G_{\boldsymbol{w}},$$

or equivalently,

$$\widehat{\boldsymbol{w}}_n^* = \left\{\boldsymbol{w} : \|P_n^* f_{\boldsymbol{w}}\| = o_P(n^{-\frac{1}{2}})\right\}, \tag{6.7}$$

where $G_{\boldsymbol{w}}$ and $f_{\boldsymbol{w}}$ are defined in Example 6.2.1.

Under certain technical conditions, the asymptotic normality of bootstrap Z-estimator can be established, which is summarized in lemma below.

Lemma 6.2.6 ((Wellner and Zhan, 1996)-Corollary 3.1, (Cheng and Huang, 2010)-Theorem 3.1)**.** *Let us assume that:*

Condition VI. *There exists a $\theta_\circ \in \Theta$, such that*

$$P\psi_{\theta_\circ} = 0,$$

and the function $P\psi_\theta$ is differentiable at $\theta_\circ$ with non-singular derivative matrix

$$V_{\theta_\circ} = P\frac{\partial \psi_\theta}{\partial \theta}\Big|_{\theta_\circ}.$$

Condition VII. *For any $\delta_n \to 0$, the following stochastic equicontinuity condition holds at the point $\theta_\circ$:*

$$\sup_{\|\theta-\theta_\circ\|\le\delta_n} \frac{\|\sqrt{n}(P_n - P)(\psi_\theta - \psi_{\theta_\circ})\|}{1+\sqrt{n}\|\theta-\theta_\circ\|} = o_P(1).$$

Condition VIII. *The function ψ_θ is square integrable at $\theta_\circ$, with the covariance matrix*

$$U = P\psi_{\theta_\circ}\psi_{\theta_\circ}^\top < \infty,$$

and for any $\delta_n \to 0$, the envelope function

$$F_n(x) := \sup_{\|\theta-\theta_\circ\|\le\delta_n} \frac{|e_i^\top(\psi_\theta - \psi_{\theta_\circ})|}{1+\sqrt{n}\|\theta-\theta_\circ\|}, i = 1,\dots,p, \tag{6.8}$$

satisfies the following condition

$$\lim_{\lambda\to\infty}\limsup_{n\to\infty}\sup_{t\ge\lambda} t^2 P(F_n(X_1) > t) = 0.$$

Condition IX. *Both $\widehat{\theta}_n$ and $\widehat{\theta}_n^*$ are consistent estimators, i.e. $\|\widehat{\theta}_n - \theta_\circ\| \xrightarrow{P_X} 0$, and $\|\widehat{\theta}_n^* - \theta_\circ\| \xrightarrow{P_{XD}} 0$.*

Condition X. *The bootstrap weights satisfy conditions B.1-B.5 (see pp. 50).*

Then, $\sqrt{n}(\widehat{\theta}_n - \theta_\circ)$ converges in distribution to a Gaussian distribution with zero mean and covariance matrix $Z_p := (V_{\theta_\circ})^{-1}U_{\theta_\circ}(V_{\theta_\circ})^{-1}$. Moreover, $\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n)$ converges in distribution to a normal distribution with the zero mean and the covariance matrix $c^2 Z_p$, in P-probability. The constant $c > 0$ depends on the weight random variables and is defined in B.5.

6.3 Consistency and asymptotic normality of FastICA and Bootstrap FastICA

Using the results introduced in the previous section, we can derive the consistency and asymptotic normality of FastICA by checking the necessary conditions of Theorem 6.2.2 and Theorem 6.2.3. We summarize the requirements and the result in the following theorem.

Theorem 6.3.1 (Consistency and Asymptotic Normality of FastICA)**.** *Let us assume that $\mathbb{E}\boldsymbol{z} = 0$, and $\boldsymbol{z}$ has all the moments up to the fourth, $\mathbb{E}\boldsymbol{z}\boldsymbol{z}^\top = I_d$, the function $G : \mathbb{R} \to \mathbb{R}$ has bounded and continuous derivatives, and that G and its first derivative are Lipschitz. Furthermore, we assume that the quantities $\mathbb{E}g'(s_i) \neq 0, \mathbb{E}s_i^2 g'(s_i) \neq 0, \mathbb{E}g^2(s_i) \neq 0$, and $\mathbb{E}s_i^2 g^2(s_i) \neq 0$, and $\mathbb{E}G(\boldsymbol{w}^\top\boldsymbol{z}), \forall \boldsymbol{w} \in \mathcal{S}^{p-1}, i = 1,\dots,p$, exist and are bounded. Then the sequence*

$$\widehat{\boldsymbol{w}}_n = \operatorname*{argmax}_{\boldsymbol{w}\in\mathcal{S}^{p-1}} P_n G_{\boldsymbol{w}},$$

that is produced by the FastICA iteration (5.6) is consistent and is asymptotically normal. In other words, we have

$$\widehat{\boldsymbol{w}}_n \xrightarrow{P} \boldsymbol{w}_\circ,$$

and

$$\sqrt{n}(\widehat{\boldsymbol{w}}_n - \boldsymbol{w}_\circ) \rightsquigarrow \mathcal{N}\left(0, V_{\boldsymbol{w}_\circ}^{-1} U_{\boldsymbol{w}_\circ} V_{\boldsymbol{w}_\circ}^{-1}\right),$$

where

$$V_{\boldsymbol{w}_\circ} = \mathbb{E}\boldsymbol{z}\boldsymbol{z}^\top g'(\boldsymbol{w}_\circ^\top \boldsymbol{z}),$$

and

$$U_{\boldsymbol{w}_\circ} = \mathbb{E}\boldsymbol{z}\boldsymbol{z}^\top g^2(\boldsymbol{w}_\circ^\top \boldsymbol{z}).$$

Note, that we can simplify $V_{\boldsymbol{w}_\circ}^{-1} U_{\boldsymbol{w}_\circ} V_{\boldsymbol{w}_\circ}^{-1}$ further to Σ, where

$$\Sigma = A \begin{pmatrix} \frac{\mathbb{E}g^2(s_\circ)}{\mathbb{E}g'(s_\circ)^2} & & & & \\ & \ddots & & & \\ & & \frac{\mathbb{E}s_\circ^2 g^2(s_\circ)}{(\mathbb{E}s_\circ^2 g'(s_\circ))^2} & & \\ & & & \ddots & \\ & & & & \frac{\mathbb{E}g^2(s_\circ)}{\mathbb{E}g'(s_\circ)^2} \end{pmatrix} A^\top.$$

For the proof of the above theorem see Section 6.6.1.

Remark 6.3.2. For any consistent estimator $\widehat{\Sigma}$ of Σ, we have

$$\sqrt{n}\widehat{\Sigma}^{-\frac{1}{2}}(\boldsymbol{w}_n - \boldsymbol{w}_\circ) \rightsquigarrow \mathcal{N}(0, I_p).$$

It is of both practical and theoretical interest to check if the bootstrap FastICA converges, or if it shows asymptotic normality. Note that due to randomization, the local analysis approach used in previous works, such as (Hyvärinen, 1999; Oja and Yuan, 2006; Tichavsky et al., 2006) is not applicable anymore. Using the setup introduced in previous section, we can check the asymptotic normality of the bootstrap FastICA by showing that the conditions (VI)-(X) are satisfied under certain conditions. The requirements and the result are provided in the following theorem.

Proposition 6.3.3. *In addition to the assumption of Theorem 6.3.1, let us assume that, for $i = 1, \ldots, p$, $\mathbb{E}s_i^2 g'^2(s_i)$, $\mathbb{E}s_i g'^2(s_i)$, $\mathbb{E}g'^2(s_i)$, $\mathbb{E}s_i^4 g'^2(s_i)$, and source signals are bounded. Then the bootstrap estimator is consistent and the following holds,*

$$\sqrt{n}(\widehat{\boldsymbol{w}}_n^* - \widehat{\boldsymbol{w}}_n) \rightsquigarrow \mathcal{N}\left(0, c^2 V_{\boldsymbol{w}_\circ}^{-1} U_{\boldsymbol{w}_\circ} V_{\boldsymbol{w}_\circ}^{-1}\right) \qquad (\text{in } P)$$

where

$$V_{\boldsymbol{w}_\circ} = \mathbb{E}\boldsymbol{z}\boldsymbol{z}^\top g'(\boldsymbol{w}_\circ^\top \boldsymbol{z}),$$

and

$$U_{\boldsymbol{w}_\circ} = \mathbb{E}\boldsymbol{z}\boldsymbol{z}^\top g^2(\boldsymbol{w}_\circ^\top \boldsymbol{z}).$$

The positive constant c depends on weight random variables $(D_{1n}, \ldots, D_{nn})$ through $\frac{1}{n}\sum_{i=1}^n (D_{in} - 1)^2 \xrightarrow{P_D} c^2 > 0$.

Similar statistical results can be established for other ICA methods, which are shortly introduced in the beginning of Section 5.1. However, a different set of statistical techniques, such as *U*-statistics (De la Peña and Giné, 1999; Van der Vaart, 1998), might be needed to study the rate of convergence or other statistical properties of matrices M_i, defined on pp. 43.

6.4 Empirical Results

To illustrate the theoretical implications in practice, a series of experiments were performed with both artificially generated and real-world data using Algorithm 2.

In both simulated and real cases, we first run the FastICA for 100 runs without bootstrapping but with different initializations, in order to determine representative sets of initial conditions. All other parameters are the same as used in the following experiments. This set of initial conditions are kept fixed during the bootstrap analysis so that all the randomness in the estimated solutions are due to the resampling. For the normality test, we used the Henze-Zirkler's Multivariate Normality Test (Henze and B. Zirkler, 1990). A matlab implementation of this test can be found in (Trujillo-Ortiz et al., 2007).

6.4.1 Simulated Data

Here, we used three simulated signals available in FastICA toolbox. These signals are a sinusoid, a sawtooth wave, and a periodic sigmoidal wave. They were mixed with a random 3×3 mixing matrix generated in such a way that it produces an already whitened data-matrix of size 3×500. 100 runs of FastICA were computed, searching for 3 independent components in each run and using a bootstrap sampling with 400 (80% out of 500) samples in each run. Figure 6.1 depicts an example of the mixed signals.

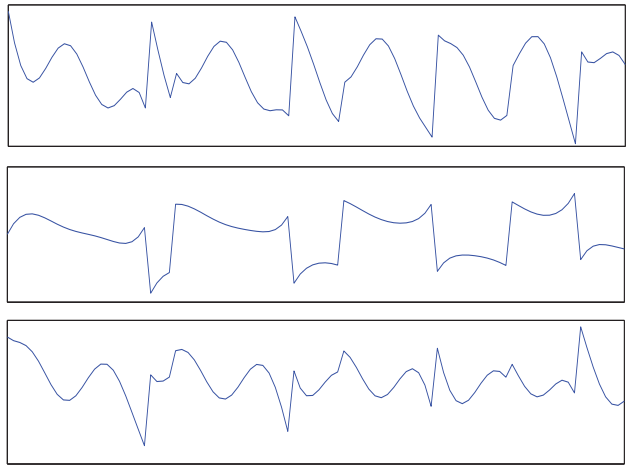


Figure 6.1. Example of mixture signals used in the simulated data.

Following Algorithm 2, the estimations were then clustered using a correlation threshold of 0.99 and taking into account only direct links between estimations. Figure 6.2 shows the resulting three independent components. Only a portion of the periodic signals is shown. For each component, light gray lines correspond to the 100 estimation ($\widehat{w}_n^*$) after the sign correction, and the solid line depicts the mean of the estimations. In each case, the estimation errors were found to be normally distributed, with p-values of 0.87, 0.93 and 0.41 respectively. Note that we always correct the signs.

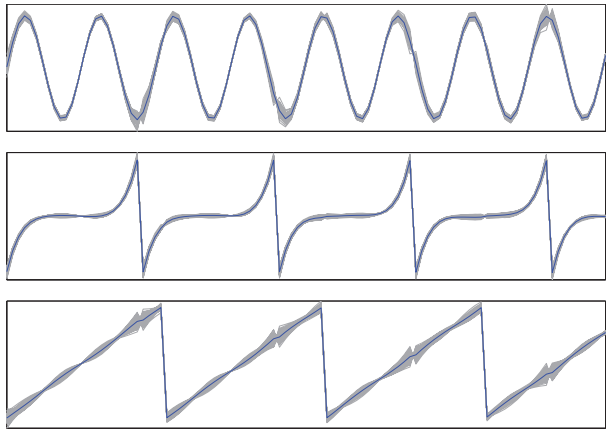


Figure 6.2. Results with simulated data. For each of the three independent components, the 100 bootstrap estimations in light gray overlaid with their average as a solid line. In each case, the estimation errors were found to be normally distributed, with p-values equal 0.87, 0.93 and 0.41 respectively.

6.4.2 fMRI data analysis

Further experiments were done with functional magnetic resonance imaging (fMRI) data from an auditory experiment. A series of whole-head recordings of a single subject were used. In the fMRI study, subjects listened to safety instructions in 30s intervals, interleaved with 30s resting periods. All the data were acquired at the Advanced Magnetic Imaging Centre of Aalto University, using a 3.0 Tesla MRI scanner (Signa EXCITE 3.0T; GE Healthcare, Chalfont St. Giles, UK) with a quadrature birdcage head coil, and using Gradient Echo (GRE) Echo Planar Imaging (EPI) (TR 3s, TE 32ms, 96x96 matrix, FOV 20cm, slice thickness 3mm, 37 axial slices, 80 time points (excl. 4 first ones), flip angle 90°). For further details on the data set, see (Ylipaavalniemi and Vigário, 2008).

We performed data preprocessing including realignment, normalization, smoothing and masking off areas outside the brain. The resulting data-matrix has a size of 80×263361. 500 runs of ICA were performed, searching for 15 components from a whitened space with 30 dimensions, and using a bootstrap sampling with 138944 (52.76% out of 263361) samples in each run. The estimations were clustered using correlations between the component time-courses, with a threshold of 0.97.

The fMRI data, similar to any real-world measurement, may violate at least some of the strict assumptions in FastICA, or in the derivations in this chapter. Therefore, the comparison of the bootstrap estimations is not as straightforward as in the simulated case. One difficulty is that the whitening step of FastICA can produce different results for each bootstrap set. For example, whitening step can flip the signs of individual dimensions among the bootstrap rounds. So, we re-clustered the estimations using cosine similarity with a suitably high threshold.

A subset of 39 independent components were identified from the bootstrap FastICA. Figure 6.3 depicts 5 found independent components. The average estimate and variations are also provided.

In Figure 6.3, the first component has small variability, whereas the other four components show significant variations. Only 43 estimations of the fifth compo-

nent were found during the 500 bootstrap rounds, implying that this component is hard to estimate with FastICA. The first component represents activation of the primary auditory cortices, whereas the other components split activity along the cingulate gyrus in four different parts. Note, that in some cases the temporal variability is higher around some time points than in others, and also the spatial variance is focused on certain regions. It is speculated in (Ylipaavalniemi and Vigário, 2008) that the last four components belong to a subspace.

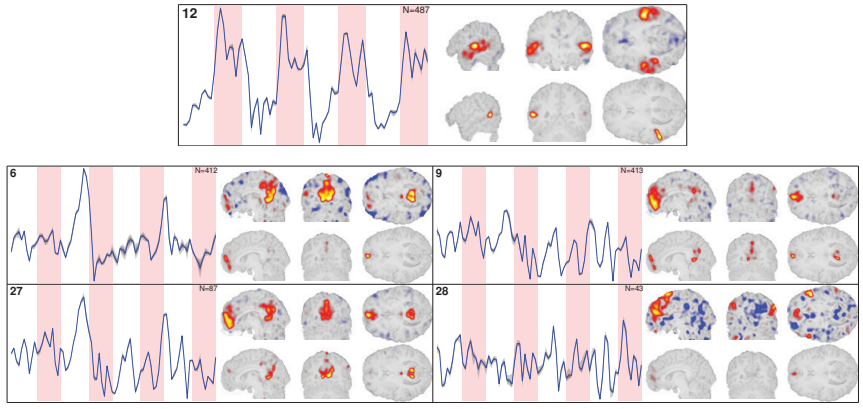


Figure 6.3. Examples of independent components estimated with bootstrapped FastICA. For each of the 5 components, on the left: the index of the component, the temporal average, and the temporal quantiles as light shades of gray, are overlaid on the stimulus block reference. On the right side of each panel: the three slices on top show the spatial mean overlaid on a structural reference brain; and similarly on the bottom, the spatial variance, overlaid on the same reference. The bootstrap was performed 500 times and N shows how many estimations of each component were found.

Figure 6.4 shows the estimated demixing vectors for component 12, 6, and 27, which are selected from Figure 6.3. In (a), apart from the obvious sign flips, the variations are small on average. Also, the covariance matrix of the variations is nearly the identity matrix. Five subgroups accounting for different configurations of sign flips were identified. Each of the subgroups was then tested against the best matching ground truth component. All except one passed the normality test with p-values of 0.1912, 0.0812, 0.1321 and 0.1615. The group that did not pass the test could contain outliers, even with a high clustering threshold.

Figure 6.4 (b) is part of a subspace of four components. The covariance matrix shows clear block structure, for example in neighborhood of coordinates 20 and 27. In this case, there are also five subgroups (due to the sign flip), but they all pass the normality test with p-values 0.2530, 0.2076, 0.0613, 0.1290 and 0.0512. For the last component, the number of estimations is too small to allow running a normality test, but otherwise the situation seems similar to the previous components. For the omitted components 9 and 28, the situation is very similar. All subgroups in component 9 pass the normality test and the covariance shows a weaker structure than in component 6. Component 28 has a similar covariance to component 27, and again too few estimations to allow for normality testing.

The normality tests show that the experimental results match the developed theory, even when the components are considered to belong to a subspace. This suggests that bootstrap FastICA is able to estimate reliable directions within the subspace. Although the estimated components passed the normality test, there should be some further evidence of the subspace covariation.

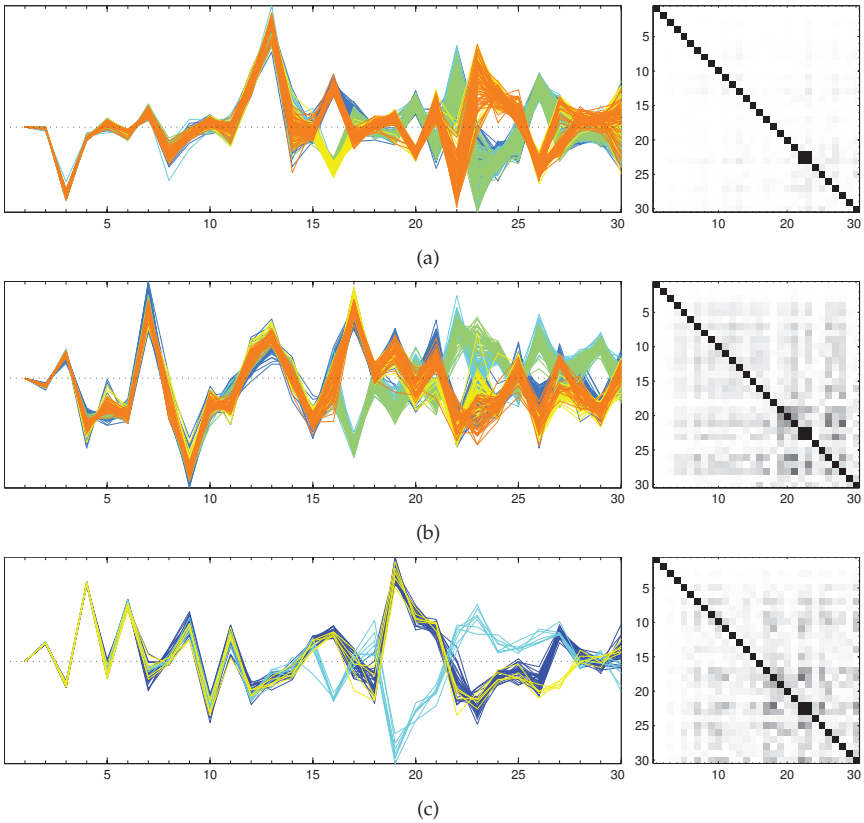


Figure 6.4. The estimated demixing vectors and their sample covariance matrices. The groups of estimated demixing vectors, with different sign configurations are depicted on the left, and the sample covariance matrix of the estimations on the right for (a) component 12, (b) component 6, and (c) component 27. The dashed line is the 0-vector.

Figure 6.5 shows coordinate-wise histograms of the largest subgroup of vectors from component 6. The histogram of most of the dimensions is close to normal density function. The histogram of estimations of coordinate 27 is similar to a bimodal, which might be due to different local minima in FastICA objective. The histogram Coordinate 27 is close to bimodal in some of the other subgroups and components belonging to the same subspace, as further experiments revealed.

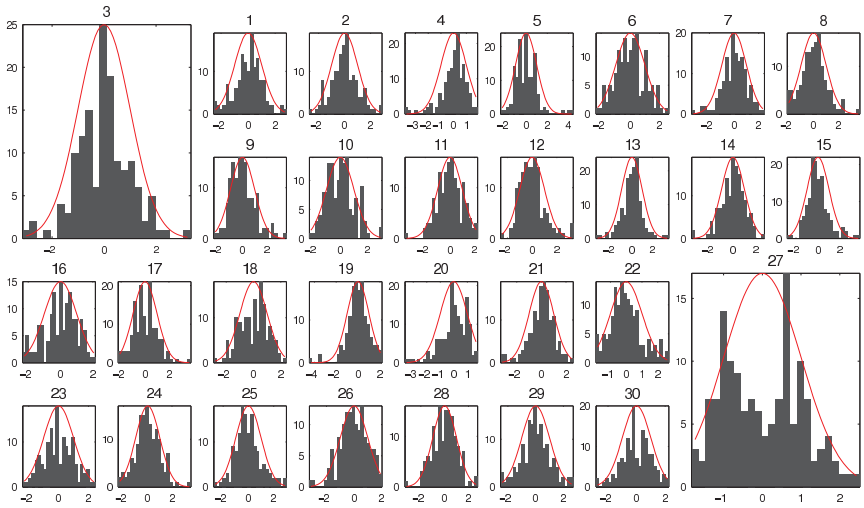


Figure 6.5. Coordinate-wise histograms of the estimated demixing vectors. The histograms are calculated from the first subgroup (with respect to sign flip) of vectors in component 6. For reference, a fitted Gaussian probability density function is shown with a solid curve overlaid on each histogram.

6.5 Discussion

ICA algorithms, and FastICA in particular, have been successfully utilized for source separation in many applications, e.g. biomedical, audio signal processing, and hyperspectral image analysis (Comon and Jutten, 2010). However, when applying the algorithm for several times with the same data, with different initializations or data subsamplings, one may encounter variations in the estimated sources. To address this issue one can run FastICA with a sufficient number of different sub-samplings and initializations, and select the results that appear more frequently. Empirical studies show that global optima can be found in such a randomized way (Ylipaavalniemi and Vigário, 2008). Multiple runs with bootstrap samples are able to efficiently explore some hidden structures of data (Ylipaavalniemi et al., 2009).

Intuitively, the randomization of both initial conditions and the subsampling improves the likelihood of finding the global optimum by slightly changing the objective function landscape and the direction of search therein. To the best of our knowledge, this is the first study on validity of multiple run approach to FastICA. In Chapter 5 and 6, we derived a probabilistic convergence rate of FastICA, which depends on the number of samples, the initialization and the concentration of sample distributions. Moreover, using empirical process theory,

we show that FastICA is statistically consistent and its convergence to a true solution is asymptotically normal. We also extend this result to bootstrap FastICA. These all together justify the use of FastICA in a bootstrapped and randomly initialized way.

Empirical results on both the synthetic data set and the real data set confirm the proposed theory. However, there might be difficulties in real data, as the requirements may not all be fulfilled. In particular, fMRI data may present non-stationarity, and independency between components is not always guaranteed. This may result in a lower rate of convergence and therefore the normality may not be achieved.

6.6 Proofs and further details

For the rest of this section, let us define $f_w(z) = zg(w^\top z)$, $G_w(z) = G(w^\top z)$ and $h_w(z) := zz^\top g'(w^\top z)$.

Proposition 6.6.1. *Let us assume that z has all the moments up to the fourth, and $G : \mathbb{R} \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$, and $g' : \mathbb{R} \to \mathbb{R}$ are Lipschitz and differentiable. We further assume z has zero mean with identity covariance. For fixed $p < \infty$, the function classes*

$$\mathcal{F}_G := \{G_w : w \in \mathcal{S}^{p-1}, z \in \mathbb{R}^p\},$$
$$\mathcal{F}_g := \{f_w : w \in \mathcal{S}^{p-1}, z \in \mathbb{R}^p\},$$

and

$$\mathcal{F}_g(\delta) := \{f_w - f_{w_\circ} : \|w - w_\circ\| \le \delta, w \in \mathcal{S}^{p-1}, z \in \mathbb{R}^p\},$$

are coordinate-wise P-Donsker and P-Glivenko-Cantelli.

Proof (Proposition 6.6.1)**.** Suppose $w_1, w_2 \in \mathcal{S}^{p-1}$ for finite integer p. For the class $\mathcal{F}_G$ we have

$$\left|G(w_2^\top z) - G(w_1^\top z)\right| \le (|G|_L \|z\|)\|w_2 - w_1\|.$$

By Remark 2.3.7, we should check that the term inside parenthesis in the right hand side of the above inequality has bounded second moment:

$$\mathbb{E}|G|_L^2 \|z\|^2 = |G|_L^2 \mathbb{E}\|z\|_2^2 = |G|_L^2 \sum_{j=1}^{p} \mathbb{E}z_j^2 = p|G|_L^2,$$

where we used the assumption that the second moment of $z_i, i = 1, \dots, p$ is equal to one. Therefore, $\mathbb{E}|G|_L^2\|z\|^2 < \infty$, and the class $\mathcal{F}_G$ has finite bracketing integral and is a P-Donsker class.

For the class $\mathcal{F}_g$ and $\forall i, 1 \le i \le p$, we have

$$\left|e_i^\top z \left(g(w_2^\top z) - g(w_1^\top z)\right)\right| \le \left(\left|e_i^\top z\right| |g|_L \|z\|\right) \|w_2 - w_1\|.$$

As before, we check that the term in parenthesis in the right hand side of preceding inequality has bounded second norm. Indeed, we have

$$|g|_L^2 \mathbb{E}\|z\|_2^2 \left|e_i^\top z\right|^2 = |g|_L^2 \Big(\mathbb{E} \sum_{\substack{j=1 \\ j \ne i}}^{p} z_j^2 + \mathbb{E}z_i^4\Big) = |g|_L^2 \left(p - 1 + \mathbb{E}z_1^4\right),$$

which is bounded by the assumptions on $|g|_L$ and the fourth moment of z. Thus, the bracketing integral is finite and class $\mathcal{F}_g$ is P-Donsker and P-Glivenko-Cantelli. Similarly, $\mathcal{F}_g(\delta)$ is P-Donsker and P-Glivenko-Cantelli.

□

6.6.1 Proof of Theorem 6.3.1

Here, we show that the requirements for the consistency and asymptotic normality of M-estimators hold for FastICA. Note that the consistency of the estimator is required for the proof of the asymptotic normality.

-Consistency

To check the consistency of FastICA, we need to check conditions (I) and (II) of Theorem 6.2.2, for the class $\mathcal{F}_G$:

(I) Lemma 6.6.1 implies that the function class

$$\mathcal{F}_G = \left\{ G_{\boldsymbol{w}} : \boldsymbol{w} \in \mathcal{S}^{p-1} \right\},$$

is P-Glivenko-Cantelli. Therefore, by definition we have

$$\sup_{\boldsymbol{w} \in \mathcal{S}^{p-1}} |(P_n - P) G_{\boldsymbol{w}}| \to 0, \quad P\text{-almost surely}$$

which implies that the condition I in Theorem 6.2.2 holds.

(II) This condition is also called *well-separated* condition in statistics literature. For $0 < \|\boldsymbol{w} - \boldsymbol{w}_\circ\| \leq \delta$, $\delta > 0$ small, and $\boldsymbol{w}, \boldsymbol{w}_\circ \in \mathcal{S}^{p-1}$, let us assume that

$$\boldsymbol{w}^\top \boldsymbol{z} = \sum_{i=1}^{p} \alpha_i s_i,$$

for some $\alpha_i \in \mathbb{R}$, such that at least two α_i are non-zero. Then, $\boldsymbol{w}^\top \boldsymbol{z}$ is at least a mixture of two different independent source signals and should not attain the maximum non-Gaussianity. This implies that $G_{\boldsymbol{w}_\circ}$ is well separated. Furthermore, we can expand $f_{\boldsymbol{w}}$ around $\boldsymbol{w}_\circ$, assuming that the second order derivative of the function G is bounded. Then,

$$\begin{aligned} P f_{\boldsymbol{w}} &= P f_{\boldsymbol{w}_\circ} + (\boldsymbol{w} - \boldsymbol{w}_\circ)^\top P h_{\boldsymbol{w}_\circ} \\ &= 0 + (\boldsymbol{w} - \boldsymbol{w}_\circ) A \mathbb{E} \boldsymbol{s} \boldsymbol{s}^\top g'(s_\circ) A^\top \\ &= (\boldsymbol{w} - \boldsymbol{w}_\circ) A K_p A^\top, \end{aligned}$$

where K_p is a diagonal matrix with entries $\mathbb{E} g'(s_\circ)$, $i \neq \circ$ and $\mathbb{E} s_\circ^2 g'(s_\circ)$, where "$\circ$" is the index of source signal corresponding to $\boldsymbol{w}_\circ$. The norm of $A K_p A^\top$ is non-zero by the assumption. Thus, $P f_{\boldsymbol{w}}$ is non-zero for $\|\boldsymbol{w} - \boldsymbol{w}_\circ\| \neq 0$, which implies that the condition II holds. Note that in the above computations terms with smaller orders are omitted.

Now, for the FastICA both condition I and II holds. Therefore by Theorem 6.2.2 sample FastICA is a consistent estimator of separating directions, i.e.

$$\widehat{\boldsymbol{w}}_n \xrightarrow{P} \boldsymbol{w}_\circ.$$

-Asymptotic Normality

The asymptotic normality of the solution $\widehat{\boldsymbol{w}}_n$ can be established by checking conditions (III) to (V) in Lemma 6.2.3:

(III) This condition requires $G_{\boldsymbol{w}}$ to be smooth with bounded derivatives, which is the case by the assumptions.

(IV) This condition holds if we show that the Taylor expansion around $\boldsymbol{w}_\circ$ is bounded, and that $P(G_{\boldsymbol{w}} - G_{\boldsymbol{w}_\circ})$ can be represented up to the second order, when $\|\boldsymbol{w} - \boldsymbol{w}_\circ\| \leq \delta$ for small $\delta > 0$. Note, that for $\|\boldsymbol{w} - \boldsymbol{w}_\circ\| \leq \delta$, we have

$$\begin{aligned} P\left(G_{\boldsymbol{w}} - G_{\boldsymbol{w}_\circ}\right) =& P\big(G_{\boldsymbol{w}_\circ} + (\boldsymbol{w} - \boldsymbol{w}_\circ)^\top f_{\boldsymbol{w}_\circ}\big) \\ &+ \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_\circ)^\top P h_{\boldsymbol{w}_\circ}(\boldsymbol{w} - \boldsymbol{w}_\circ) - P G_{\boldsymbol{w}_\circ} \\ &+ o(\|\boldsymbol{w} - \boldsymbol{w}_\circ\|^2) \\ =& \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_\circ)^\top P h_{\boldsymbol{w}_\circ}(\boldsymbol{w} - \boldsymbol{w}_\circ) + o(\|\boldsymbol{w} - \boldsymbol{w}_\circ\|^2). \end{aligned}$$

In the last line we used the fact that $P f_{\boldsymbol{w}_\circ} = 0$, which is due to optimality condition, see (5.3). Comparing the quadratic representation to the notation of Lemma 6.2.3, we have

$$\begin{aligned} V_{\boldsymbol{w}_\circ} &= P h_{\boldsymbol{w}_\circ} \\ &= \mathbb{E}\boldsymbol{z}\boldsymbol{z}^\top g'(\boldsymbol{w}_\circ^\top \boldsymbol{z}) \\ &= A \mathbb{E}\boldsymbol{s}\boldsymbol{s}^\top g'(s_\circ) A^\top \\ &= A \mathrm{diag}[\mathbb{E}g'(s_\circ), \ldots, \mathbb{E}s_\circ^2 g'(s_\circ), \ldots, \mathbb{E}g'(s_\circ)] A^\top. \end{aligned} \tag{6.9}$$

The argument of diag in the right hand side of (6.9) contains only one entry with $\mathbb{E}s_\circ^2 g'(s_\circ)$ and the rest are $\mathbb{E}g'(s_\circ)$. By assumption, g' is nonsingular and both $\mathbb{E}g'(s_\circ)$ and $\mathbb{E}s_\circ^2 g'(s_\circ)$ exist and are nonzero and therefore, both $V_{\boldsymbol{w}_\circ}$ and $V_{\boldsymbol{w}_\circ}^{-1}$ are well defined.

(V) This condition is equivalent to P-Donsker condition for the class $\mathcal{F}_G$, which is shown in Lemma 6.6.1.

The matrix $U_{\boldsymbol{w}_\circ}$ in Lemma 6.2.3 can be computed as follows:

$$\begin{aligned} U_{\boldsymbol{w}_\circ} &= P\left(\nabla G_{\boldsymbol{w}}|_{\boldsymbol{w}_\circ}\right)\left(\nabla G_{\boldsymbol{w}}|_{\boldsymbol{w}_\circ}\right)^\top \\ &= \mathbb{E}\boldsymbol{z}\boldsymbol{z}^\top g^2(\boldsymbol{w}_\circ^\top \boldsymbol{z}) \\ &= A \mathbb{E}\boldsymbol{s}\boldsymbol{s}^\top g^2(s_\circ) A^\top \\ &= A \mathrm{diag}[\mathbb{E}g^2(s_\circ), \ldots, \mathbb{E}s_\circ^2 g^2(s_\circ), \ldots, \mathbb{E}g^2(s_\circ)] A^\top. \end{aligned} \tag{6.10}$$

By assumption, we have $\mathbb{E}g^2(s_\circ) \neq 0$, and $\mathbb{E}s_\circ^2 g^2(s_\circ) \neq 0$, and all are bounded, which implies $U_{\boldsymbol{w}_\circ}$ is bounded and non-singular.

Therefore, all necessary conditions in Lemma 6.2.3 hold for FastICA, implying that it is asymptotically normal, i.e.

$$\sqrt{n}(\widehat{\boldsymbol{w}}_n - \boldsymbol{w}_\circ) \rightsquigarrow \mathcal{N}\left(0, V_{\boldsymbol{w}_\circ}^{-1} U_{\boldsymbol{w}_\circ} \left(V_{\boldsymbol{w}_\circ}^{-1}\right)^\top\right) = \mathcal{N}(0, \Sigma). \tag{6.11}$$

The covariance above can be computed as

$$\Sigma = A \mathrm{diag}\left[\frac{\mathbb{E}g^2(s_\circ)}{(\mathbb{E}g'(s_\circ))^2}, \ldots, \frac{\mathbb{E}s_\circ^2 g^2(s_\circ)}{(\mathbb{E}s_\circ^2 g'(s_\circ))^2}, \ldots, \frac{\mathbb{E}g^2(s_\circ)}{(\mathbb{E}g'(s_\circ))^2}\right] A^\top.$$

□

6.6.2 Proof of the Proposition 6.3.3

We first bring two lemmas which are useful tools to show requirements in asymptotic normality of bootstrap Z-estimators. Also we bring a consistency lemma for Z-estimators.

Lemma 6.6.2 ((Van der Vaart and Wellner, 1996)-Lemma 3.3.5)**.** *Suppose the class of functions*

$$\{\psi_\theta - \psi_{\theta_0} : \|\theta - \theta_0\| \leq \delta\}$$

is P-Donsker for some $\delta > 0$, and that

$$P(\psi_\theta - \psi_{\theta_\circ})^2 \to 0, \text{ as } \theta \to \theta_\circ. \tag{6.12}$$

If $\widehat{\theta}_n$ converges in probability to $\theta_\circ$, we then have,

$$\|\sqrt{n}(P_n - P)(\psi_{\widehat{\theta}_n} - \psi_{\theta_\circ})\| = o_P(1 + \sqrt{n}\|\widehat{\theta}_n - \theta_\circ\|).$$

Lemma 6.6.3 (Multiplier Glivenko-Cantelli (Van der Vaart and Wellner, 1996), Lemma 3.6.16)**.** *Let $\mathcal{F}$ be a Glivenko-Cantelli class of measurable functions. For each n, let $(D_{n1}, \ldots, D_{nn})$ be a set of exchangeable nonnegative random variables that are independent to $\{\boldsymbol{x}_i\}_{i\geq 1}$. Furthermore, assume that*

$$\sum_{i=1}^{n} D_{ni} = 1, \quad \text{and} \quad \max_{1\leq i\leq n} D_{ni} \xrightarrow{P_D} 0.$$

Then, for every $\epsilon > 0$, as $n \to \infty$, we have

$$P_D\left\{\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n} D_i(\delta_{\boldsymbol{x}_i} - P)f\right| > \epsilon\right\} \to 0.$$

Theorem 6.6.4 (Consistency of Z-estimator (Van der Vaart, 1998), Theorem 5.9)**.** *Let us consider θ_n, that is defined in (6.4), be random vector-valued functions and θ, defined in (6.3) be a fixed vector-valued function such that for every $\epsilon > 0$*

$$\sup_{\theta\in\Theta} \|P_n\psi_\theta - P\psi_\theta\| \xrightarrow{P} 0,$$
$$\inf_{\|\theta-\theta_\circ\|\geq\epsilon} \|P\psi_\theta\| > 0 = P\psi_{\theta_\circ}.$$

Then any sequence of estimators $\widehat{\theta}_n$ such that $P_n\psi_{\theta_n} = o_P(1)$ converges in probability to $\theta_\circ$.

Now, we begin to prove Proposition 6.3.3 by checking the conditions VI—X required in Theorem 6.2.6. We denote the derivative of $g : \mathbb{R} \to \mathbb{R}$ by $g' : \mathbb{R} \to \mathbb{R}$.

Condition (VI) follows directly from the assumptions. Also we assume that the random weights in the bootstrapping satisfy condition (X).

(VII) To show that the condition (VII) holds for FastICA estimator, we use Lemma 6.6.2. This lemma requires that the function set

$$\mathcal{F} := \left\{f_{\boldsymbol{w}} - f_{\boldsymbol{w}_\circ} : \|\boldsymbol{w} - \boldsymbol{w}_\circ\| \leq \delta_n, \boldsymbol{w} \in \mathcal{S}^{p-1}\right\},$$

is P-Donsker, which is shown in Proposition 6.6.1. In addition, we should show that

$$P\left\|f_{\boldsymbol{w}} - f_{\boldsymbol{w}_\circ}\right\|^2 \to 0 \qquad \text{as} \quad \|\boldsymbol{w} - \boldsymbol{w}_\circ\| \to 0. \tag{6.13}$$

By Taylor expansion, we have

$$\begin{aligned} P\left\|f_{\boldsymbol{w}} - f_{\boldsymbol{w}_\circ}\right\|^2 =& P\left\|h_{\boldsymbol{w}_\circ}(\boldsymbol{w} - \boldsymbol{w}_\circ)\right\|^2 \\ =& (\boldsymbol{w} - \boldsymbol{w}_\circ)^\top P h_{\boldsymbol{w}_\circ} h_{\boldsymbol{w}_\circ}(\boldsymbol{w} - \boldsymbol{w}_\circ). \end{aligned}$$

The terms with smaller order are omitted in above. So, to show (6.13) we can check if the entries of the matrix $Ph^2_{\boldsymbol{w}_\circ}$ are bounded. Note, that

$$\begin{aligned} Ph^2_{\boldsymbol{w}_\circ} &= \mathbb{E}\boldsymbol{s}\boldsymbol{s}^\top \boldsymbol{s}\boldsymbol{s}^\top g'^2(s_\circ) \\ &= \mathbb{E}\Big(g'^2(s_\circ)\boldsymbol{s}\boldsymbol{s}^\top \sum_{i=1}^{p} s_i^2\Big). \end{aligned}$$

Thus, the entries of $Ph^2_{\boldsymbol{w}_\circ}$ are bounded if $\mathbb{E}s^2_\circ g'^2(s_\circ)$, $\mathbb{E}s_\circ g'^2(s_\circ)$, $\mathbb{E}s^4_\circ g'^2(s_\circ)$ are bounded. In addition, we need that all random variables $s_i, \forall 1 \le i \le p$ have moments up to the forth moment. Also $\mathbb{E}g'^2(s_\circ)$ should be bounded. These conditions hold by the assumption.

(VIII) There are two different approaches to show the requirement (VIII). We can either show that the envelope function $F_n(x)$, defined in (6.8), is uniformly bounded, i.e.

$$\limsup_{n\to\infty} F_n(x) \le M < \infty, \forall x \in \mathcal{X},$$

or check the moments condition, i.e. $\limsup_{n\to\infty} \mathbb{E}[(F_n(X_1))^{2+\delta}] < \infty$, for some $\delta > 0$. For the mapping $f_{\boldsymbol{w}}$, we have

$$\begin{aligned} F_n(\boldsymbol{z}) &= \sup_{\|\boldsymbol{w}-\boldsymbol{w}_\circ\|\le\delta_n} \frac{\left|\boldsymbol{e}_i^\top (f_{\boldsymbol{w}} - f_{\boldsymbol{w}_\circ})\right|}{1+\sqrt{n}\|\boldsymbol{w}-\boldsymbol{w}_\circ\|} \\ &\le \sup_{\|\boldsymbol{w}-\boldsymbol{w}_\circ\|\le\delta_n} \frac{\left|\boldsymbol{e}_i^\top h_{\boldsymbol{w}_\circ}(\boldsymbol{w}-\boldsymbol{w}_\circ)\right| + \left|\boldsymbol{e}_i^\top \boldsymbol{z}\right| o(\|\boldsymbol{w}-\boldsymbol{w}_\circ\|^2)}{1+\sqrt{n}\|\boldsymbol{w}-\boldsymbol{w}_\circ\|} \\ &\le \sup_{\|\boldsymbol{w}-\boldsymbol{w}_\circ\|\le\delta_n} \frac{\|h_{\boldsymbol{w}_\circ}\|\|\boldsymbol{w}-\boldsymbol{w}_\circ\| + \left|\boldsymbol{e}_i^\top \boldsymbol{z}\right| o(\|\boldsymbol{w}-\boldsymbol{w}_\circ\|^2)}{1+\sqrt{n}\|\boldsymbol{w}-\boldsymbol{w}_\circ\|} \\ &\le \sup_{\|\boldsymbol{w}-\boldsymbol{w}_\circ\|\le\delta_n} \frac{\|h_{\boldsymbol{w}_\circ}\|\delta_n}{1+\sqrt{n}\delta_n} + \frac{\|\boldsymbol{z}\| o(\delta_n^2)}{1+\sqrt{n}\delta_n}. \end{aligned}$$

In above the terms with smaller order are omitted. In our setup $\boldsymbol{z} = A\boldsymbol{s}$, where $\boldsymbol{s} \in \mathcal{X}$. Therefore,

$$\|h_{\boldsymbol{w}_\circ}\| = \|\boldsymbol{z}\boldsymbol{z}^\top g'(\boldsymbol{w}_\circ^\top \boldsymbol{z})\| \le |g'(s_\circ)|\|A\boldsymbol{s}\boldsymbol{s}^\top A^\top\|,$$

that are bounded by assumption for nonsingular function g'. Alternatively, we have

$$\begin{aligned} F_n(\boldsymbol{z}) &= \sup_{\|\boldsymbol{w}-\boldsymbol{w}_\circ\|\le\delta_n} \frac{\left|\boldsymbol{e}_i^\top (f_{\boldsymbol{w}} - f_{\boldsymbol{w}_\circ})\right|}{1+\sqrt{n}\|\boldsymbol{w}-\boldsymbol{w}_\circ\|} \\ &\le \sup_{\|\boldsymbol{w}-\boldsymbol{w}_\circ\|\le\delta_n} \frac{\left|\boldsymbol{e}_i^\top \boldsymbol{z}\right|\|\boldsymbol{z}\|\|g\|_L\|\boldsymbol{w}-\boldsymbol{w}_\circ\|}{1+\sqrt{n}\|\boldsymbol{w}-\boldsymbol{w}_\circ\|}, \end{aligned}$$

that is bounded from above by the assumption.

Other parts of (VIII) are shown in the proof of the asymptotic normality of FastICA.

(IX) The consistency of FastICA, i.e. $\|\widehat{\boldsymbol{w}}_n - \boldsymbol{w}_\circ\| \xrightarrow{P} 0$, is shown in Theorem 6.3.1. To show the consistency of the bootstrap FastICA, we need to show that the conditions of Theorem 6.6.4 are satisfied, i.e. the objective function is well-separated and the following holds:

$$\sup_{\boldsymbol{w}\in\mathcal{S}^{p-1}} \left|(P_n^* - P)f_{\boldsymbol{w}}\right| \xrightarrow{P_{XD}} 0. \tag{6.14}$$

Now, let us assume $A_n = o_{P_D}(1)$. For arbitrary $\epsilon, \delta > 0$, we have

$$\begin{aligned} P_{XD}\{|A_n| \geq \epsilon\} =& \mathbb{E}_X P_{D|X}\{|A_n| \geq \epsilon\} \\ =& \mathbb{E}_X \left[P_{D|X}\{|A_n| \geq \epsilon\} 1_{P_{D|X}\{|A_n| \geq \epsilon\} \geq \delta} \right] \\ &+ \mathbb{E}_X \left[P_{D|X}\{|A_n| \geq \epsilon\} 1_{P_{D|X}\{|A_n| \geq \epsilon\} < \delta} \right] \\ \leq& \mathbb{E}_X \left[1_{P_{D|X}\{|A_n| \geq \epsilon\} \geq \delta} \right] + \delta \\ \leq& P_X\{P_D\{|A_n| \geq \epsilon\} \geq \delta\} + \delta. \end{aligned}$$

By the assumption, the above goes to zero and δ is arbitrary, therefore we obtain $\lim_{n\to\infty} P_{XD}\{|A_n| \geq \epsilon\} = 0$, for any $\epsilon > 0$. Therefore,

$$A_n = o_{P_D}(1) \Rightarrow A_n = o_{P_{XD}}(1). \tag{6.15}$$

By Proposition 6.6.1, the class $\mathcal{F}_g$ is P-Glivenko-Cantelli, therefore, by Lemma 6.6.3, we have

$$P_D \left\{ \sup_{f \in \mathcal{F}_g} \left| \sum_{i=1}^{n} D_i(\delta_{x_i} - P) f \right| > \epsilon \right\} \to 0.$$

Together with conclusion (6.15) we obtain (6.14). Therefore, by Theorem 6.6.4 we obtain the consistency of the bootstrap FastICA. □

7. Concluding Remarks

In this thesis, we proposed an efficient approximation for kernel learning, able to cope with large sample sets and large number of kernel bases. Our method requires lighter memory and computational demand compared to the original kernel learning methods. The idea comes from realizing that the penalized empirical risk minimizer searches for a vector which has the minimum similarity to the eigenvectors with small eigenvalues of the kernel matrix. This implies that approximating the kernel matrix with its top eigenvectors, and with some additional adjustments, may not change significantly the accuracy of the prediction. On the other hand, most of the large kernel matrices are of low rank. Thus, for multiple kernel learning (MKL), we suggest constructing a dictionary, called spectral dictionary, by collecting a few eigenvectors from each kernel basis. We showed that the MKL over the Gram matrices of the spectral dictionary can be reduced to an efficient sparse optimization problem. For example, the MKL with least squares loss in this setting can be reduced to Basis Pursuit or Lasso regression.

Furthermore, we derived bounds for the Gaussian complexity of the hypothesis set generated by the spectral dictionary. Our bound shows that the complexity depends on the size of the dictionary and on its diameter. This implies that the complexity does not increase monotonically with the size of spectral class, whereas the opposite is suggested by previous bounds.

The kernel class which is used in our work is a linear combination of Gram matrices that are built by the spectral dictionary. We assumed that the coefficients belong to the ℓ_1 simplex. Extending the optimization for ℓ_p simplex and deriving bounds for the complexity of the corresponding hypothesis class is left for future work. In addition, from empirical and theoretical results we learned that an adjustment of the kernel matrix may increase the accuracy of the classification method. We can further extend this idea to other kernel methods such as kernel principal component analysis, kernel linear discriminant analysis, or the other types of MKL methods such as the method of maximizing the Gaussianity, which are left as well for future work.

In the latter part of the thesis, we have studied the statistical convergence of FastICA, as an efficient estimator for independent component analysis. This was done by applying statistical learning theory to the estimations involved in the ICA problem. We mainly focused on the statistical convergence, as well as the numerical convergence of the FastICA algorithm. In particular, we showed that this estimator converges asymptotically in distribution to a normal random variable, and that similar results hold for the bootstrap FastICA. This type of

convergence helps to design hypothesis testing, and to see how reliable the estimated directions are in terms of confidence intervals. The results may help the practitioner to decide whether to use the FastICA for a particular application. Extending such results to other ICA methods should be considered in future works.

Bibliography

Amari, S. I., & Wu, S. (1999). Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, 12(6), 783-789.

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc*, 68(3), 337-404.

Bach, F. (2008). Exploring large feature spaces with hierarchical multiple kernel learning. *arXiv preprint arXiv:0809.1493*.

Baker, C. T. H. (1977). *The numerical treatment of integral equations* (Vol. 13). Oxford: Clarendon press.

Bartlett, P. L., & Mendelson, S. (2003). Rademacher and Gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3, 463-482.

Bengio, Y. (2000). Gradient-based optimization of hyperparameters. *Neural Computation*, 12(8), 1889-1900.

Bertsekas, D. P., Nedi, A., & Ozdaglar, A. E. (2003). *Convex analysis and optimization*. Athena Scientific optimization and computation series. Athena Scientific.

Bhatia, R. (1997). *Matrix analysis* (Vol. 169). Springer Verlag.

Bickel, P. J., Ritov, Y. A., & Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4), 1705-1732.

Blanchard, G., Kawanabe, M., Sugiyama, M., Spokoiny, V., & Müller, K. R. (2006). In search of non-Gaussian components of a high-dimensional distribution. *The Journal of Machine Learning Research*, 7, 247-282.

Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. *The Journal of Machine Learning Research*, 2, 499-526.

Bousquet, O., & Herrmann, D. J. (2003). On the complexity of learning the kernel matrix. In *Advances in neural information processing systems (NIPS)*, 15, 415-422.

Boyd, S., el Ghaoui, L., Feron, E., & Balakrishnan, V. (1987). *Linear matrix inequalities in system and control theory* (Vol. 15). Society for Industrial Mathematics.

Burbidge, R., Trotter, M., Buxton, B., & Holden, S. (2001). Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Computers & Chemistry*, 26(1), 5-14.

Chapelle, O, Vapnik, V, Bousquet, O., & Mukherjee, S. (2002). Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1), 131-159.

Chen, S. S., Donoho, D. L., & Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1), 33-61.

Cheng, G., & Huang, J. Z. (2010). Bootstrap consistency for general semiparametric M-estimation. *The Annals of Statistics*, 38(5), 2884-2915.

Comon, P., & Jutten, C. (2010). *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press.

Comon, P. (1994). Independent component analysis, a new concept?. *Signal processing*, 36(3), 287-314.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.

Cortes, C., Mohri, M., & Rostamizadeh, A. (2009). L 2 regularization for learning kernels. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, AUAI Press, 109-116.

Cortes, C., Mohri, M., & Rostamizadeh, A. (2010). Generalization bounds for learning kernels. In *Proceedings of the 27th international conference on Machine learning (ICML)*.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory*. Wiley-interscience.

Davison, A. C. (2003). *Statistical models* (Vol. 11). Cambridge University Press.

De la Peña, V., & Giné, E. (1999). *Decoupling: from dependence to independence*. Springer Verlag.

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, 1-32.

Drineas, P., & Mahoney, M. W. (2005). On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *The Journal of Machine Learning Research*, 6, 2153-2175.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The annals of Statistics*, 7(1), 1-26.

El-Sherief, H., & Sinha, N. (1979). Bootstrap estimation of parameters and states of linear multivariable systems. *Automatic Control, IEEE Transactions on*, 24(2), 340-343.

Evgeniou, T., Micchelli, C. A., & Pontil, M. (2006). Learning multiple tasks with kernel methods. *The Journal of Machine Learning Research*, 6, 615-637.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10), 906-914.

Geer, S. A. (2000). *Applications of empirical process theory*. Cambridge University Press.

Golub, G. H., & Van Loan, C. F. (1996). *Matrix computations* (Vol. 3). Johns Hopkins University Press.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1), 389-422.

Henze, N., & Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Communications in Statistics-Theory and Methods*, 19(10), 3595-3617.

Hino, H., Reyhani, N., & Murata, N. (2010). Multiple kernel learning by conditional entropy minimization. In *Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on*, IEEE, 223-228.

Hino, H., Reyhani, N., & Murata, N. (2012). Multiple kernel learning with Gaussianity measures. *Neural Computation*, 24(7), 1853-1881.

Huber, P. J. (1985). Projection pursuit. *The Annals of Statistics*, 435-475.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *Neural Networks, IEEE Transactions on*, 10(3), 626-634.

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4), 411-430.

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis: algorithms and applications*. Wiley-Interscience.

Hyvärinen, A., & Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural computation*, 9(7), 1483-1492.

Izenman, A. J. (2008). *Modern multivariate statistical techniques: regression, classification, and manifold learning*. Springer.

Jaakkola, T., & Haussler, D. (1999). Probabilistic kernel regression models. In *Proceedings of the 1999 Conference on AI and Statistics*, 126, 00-04.

Jin, R., Yang, T., & Mahdavi, M. (2011). Improved Bound for the Nyström's Method and its Application to Kernel Classification. *arXiv preprint arXiv:1111.2262*.

Kawanabe, M., Blanchard, G., Sugiyama, M., Spokoiny, V., & Müller, K. R. (2006). A novel dimension reduction procedure for searching non-gaussian subspaces. In *6th International Conference on Independent Component Analysis and Blind Signal Separation*, 149-156.

Kawanabe, M., Sugiyama, M., Blanchard, G., & Müller, K. R. (2007). A new algorithm of non-Gaussian component analysis with radial kernel functions. *Annals of the Institute of Statistical Mathematics*, 59(1), 57-75.

Kim, K. J. (2003). Financial time series forecasting using support vector machines. *Neurocomputing*, 55(1), 307-319.

Kloft, M., Brefeld, U., Sonnenburg, S., & Zien, A. (2010). Non-sparse regularization and efficient training with multiple kernels. *arXiv preprint arXiv:1003.0079v3*.

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École d'Été de Probabilités de Saint-Flour XXXVIII-2008* (Vol. 2033). Springer.

Koltchinskii, V., & Giné, E. (2000). Random matrix approximation of spectra of integral operators, *Bernoulli*, 6(1):113–167.

Koltchinskii, V., & Panchenko, D. (2005). Complexities of convex combinations and bounding the generalization error in classification. *The Annals of Statistics*, 33(4), 1455-1496.

Koltchinskii, V., & Yuan, M. (2010). Sparsity in multiple kernel learning. *The Annals of Statistics*, 38(6), 3660-3695.

Lanckriet, G. R., De Bie, T., Cristianini, N., Jordan, M. I., & Noble, W. S. (2004,a). A statistical framework for genomic data fusion. *Bioinformatics*, 20(16), 2626-2635.

Lanckriet, G. R., Cristianini, N., Bartlett, P., Ghaoui, L. E., & Jordan, M. I. (2004,b). Learning the kernel matrix with semidefinite programming. *The Journal of Machine Learning Research*, 5, 27-72.

Lax, P.D. (2002). *Functional analysis*. Pure and applied mathematics. Wiley.

Ledoux, M. (2001). *The concentration of measure phenomenon* (Vol. 89). Amer. Mathematical Society.

Ledoux, M., & Talagrand, M. (2011). *Probability in Banach Spaces: isoperimetry and processes* (Vol. 23). Springer.

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (Vol. 31). Springer.

Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses*. Springer.

Leopold, E., & Kindermann, J. (2002). Text categorization with support vector machines. How to represent texts in input space?. *Machine Learning*, 46(1), 423-444.

Levina, E., & Vershynin, R. (2011). Partial estimation of covariance matrices. *Probability Theory and Related Fields*, 1-15.

Lu, C. J., Lee, T. S., & Chiu, C. C. (2009). Financial time series forecasting using independent component analysis and support vector regression. *Decision Support Systems*, 47(2), 115-125.

Meinshausen, N., & Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1), 246-270.

Mendelson, S. (2003,a). A few notes on statistical learning theory. *Advanced Lectures on Machine Learning*, 1-40.

Mendelson, S. (2003,b). On the performance of kernel classes. *The Journal of Machine Learning Research*, 4, 759-771.

Mendelson, S., & Paouris, G. (2012). On generic chaining and the smallest singular value of random matrices with heavy tails. *Journal of Functional Analysis*, 262 (9), 3775-3811.

Mendelson, S. (2012). Oracle inequalities and the isomorphic method, available at *http://maths-people.anu.edu.au/∼mendelso/papers/subgaussian-12-01-2012.pdf*.

Mercer, J. (1909). Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 415-446.

Micchelli, C. & Pontil, M. (2005). Learning the kernel function via regularization. *The Journal of Machine Learning Research*, 6, 1099–1125.

Nilsback, M. E., & Zisserman, A. (2008). Automated flower classification over a large number of classes. In *Computer Vision, Graphics & Image Processing, 2008. ICVGIP'08. Sixth Indian Conference* on, IEEE, 722-729.

Ogawa, T., Hino, H., Reyhani, N., Murata, N., & Kobayashi, T. (2011). Speaker recognition using multiple kernel learning based on conditional entropy minimization. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference* on, 2204-2207.

Oja, E., & Yuan, Z. (2006). The FastICA algorithm revisited: Convergence analysis. *Neural Networks, IEEE Transactions on*, 17(6), 1370-1381.

Pavlo, A., Paulson, E., Rasin, A., Abadi, D. J., DeWitt, D. J., Madden, S., & Stonebraker, M. (2009). A comparison of approaches to large-scale data analysis. In *Proceedings of the 35th SIGMOD international conference on Management of data*, ACM, 165-178.

Pham, D. T., & Cardoso, J. F. (2001). Blind separation of instantaneous mixtures of nonstationary sources. *Signal Processing, IEEE Transactions on*, 49(9), 1837-1848.

Rakotomamonjy, A., Bach, F., Canu, S., & Grandvalet, Y. (2008). SimpleMKL. *The Journal of Machine Learning Research*, 9, 2491-2521.

Reyhani, N. & Bickel, P. (2009). Nonparametric ICA for nonstationary instantaneous mixtures. In Amini, Massih-Reza, Habrard, Amaury, Ralaivola, Liva, and Usunier, Nicolas (eds.), *Workshop on Learning from non IID Data: Theory, Algorithms and Practice , ECML PKDD 2009*.

Reyhani, N., & Oja, E. (2011). Non-Gaussian component analysis using Density Gradient Covariance matrix. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, IEEE, 966-972.

Reyhani, N., Ylipaavalniemi, J., Vigário, R., & Oja, E. (2011). Consistency and asymptotic normality of FastICA and bootstrap FastICA. *Signal Processing* 92(8), 1767–1778.

Reyhani, N. (2013) Multiple spectral kernel learning and a gaussian complexity computation. *Neural Computation*, in print.

Ruszczyński, A. (2006). *Nonlinear optimization* (Vol. 13). Princeton university press.

Samarov, A., & Tsybakov, A. (2004). Nonparametric independent component analysis. *Bernoulli*, 10(4), 565-582.

Schölkopf, B., & Smola, A. J. (2001). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1), 1-47.

Shao, J. (1990). Bootstrap estimation of the asymptotic variances of statistical functionals. *Annals of the Institute of Statistical Mathematics*, 42(4), 737-752.

Shawe-Taylor, N., & Kandola, A. (2002). On kernel target alignment. In *Advances in Neural Information Processing Systems (NIPS)*, 14.

Sonnenburg, S., Rätsch, G., Schäfer, C., & Schölkopf, B. (2006). Large scale multiple kernel learning. *The Journal of Machine Learning Research*, 7, 1531-1565.

Srebro, N., & Ben-David, S. (2006). Learning bounds for support vector machines with learned kernels. In *Annual Conference On Learning Theory (COLT)*, 169-183.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, 1(399), 197-206.

Steinwart, I., & Christmann, A. (2008). *Support vector machines*. Springer.

Sugiyama, M., Kawanabe, M., Blanchard, G., Spokoiny, V., & Muller, K. R. (2006). Obtaining the best linear unbiased estimator of noisy signals by non-Gaussian component analysis. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*.

Suzuki, T., & Sugiyama, M. (2011). Least-squares independent component analysis. *Neural Computation*, 23(1), 284-301.

Talagrand, M. (2005). *The generic chaining: upper and lower bounds of stochastic processes*. Springer.

Talwalkar, A., & Rostamizadeh, A. (2010). Matrix coherence and the Nyström method. *arXiv preprint arXiv:1004.2008*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Tichavsky, P., Koldovsky, Z., & Oja, E. (2006). Performance analysis of the FastICA algorithm and Cramér-Rao bounds for linear independent component analysis. *Signal Processing, IEEE Transactions on*, 54(4), 1189-1203.

Tomczak-Jaegermann, N. (1989). Banach-Mazur distances and finite-dimensional operator ideals. *Pitman monographs and surveys in pure and applied mathematics*, Longman Scientific & Technical.

Tripathi, S., Srinivas, V. V., & Nanjundiah, R. S. (2006). Downscaling of precipitation for climate change scenarios: A support vector machine approach. *Journal of Hydrology*, 330(3), 621-640.

Trujillo-Ortiz, A., Hernandez-Walls, R., Barba-Rojo, K., & Cupul-Magana., L. (2007). URL `http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=17931`.

Van der Vaart, A. W. (2000). *Asymptotic statistics* (Vol. 3). Cambridge university press.

Van der Vaart, A., & Wellner, J. (1996). *Weak convergence and empirical processes: with applications to statistics*. Springer.

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer.

Vapnik, V., & Chapelle, O. (2000). Bounds on error expectation for support vector machines. *Neural computation*, 12(9), 2013-2036.

Varma, M., & Ray, D. (2007). Learning the discriminative power-invariance trade-off. In *Computer Vision, 2007 (ICCV 2007), 11th IEEE International Conference on*, 1-8.

Vempala, S. S. (2012). Modeling high-dimensional data: technical perspective. *Communications of the ACM*, 55(2), 112-112.

Wang, Y., Fan, Y., Bhatt, P., & Davatzikos, C. (2010). High-dimensional pattern regression using machine learning: From medical images to continuous clinical variables. *Neuroimage*, 50(4), 1519-1535.

Wellner, J.A., & Zhan, Y. (1996). Bootstrapping Z-estimators. *Technical Report*, 38, University of Washington Department of Statistics.

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences* (Vol. 100). Academic press.

Xu, Z., Jin, R., King, I., & Lyu, M. R. (2009). An extended level method for efficient multiple kernel learning. In *Advances in neural information processing systems (NIPS)*, 21, 1825-1832.

Ylipaavalniemi, J., & Vigário, R. (2008). Analyzing consistency of independent components: An fMRI illustration. *NeuroImage*, 39(1), 169-180.

Ylipaavalniemi, J., & Soppela, J. (2009). Arabica: robust ICA in a pipeline. In *Independent Component Analysis and Signal Separation*, 379-386.

Ylipaavalniemi, J., Savia, E., Malinen, S., Hari, R., Vigário, R., & Kaski, S. (2009). Dependencies between stimuli and spatially independent fMRI sources: Towards brain correlates of natural stimuli. *NeuroImage*, 48(1), 176-185.

Ylipaavalniemi, J., Reyhani, N., & Vigário, R. (2012). Distributional convergence of subspace estimates in FastICA: a bootstrap study. In *Latent Variable Analysis and Signal Separation*, 123-130.

Zhang, T. (2001). Convergence of Large Margin Separable Linear. In *Advances in Neural Information Processing Systems (NIPS)*, 13.

Zien, A., & Ong, C. S. (2007, June). Multiclass multiple kernel learning. In *Proceedings of the 24th international conference on Machine learning (ICML)*, ACM, 1191-1198.

DISSERTATIONS IN INFORMATION AND COMPUTER SCIENCE

Aalto-DD45/2012	Viitaniemi, Ville Visual Category Detection: an Experimental Perspective. 2012.
Aalto-DD51/2012	Hanhjärvi, Sami Multiple Hypothesis Testing in Data Mining. 2012.
Aalto-DD56/2012	Ramkumar, Pavan Advances in Modeling and Characterization of Human Neuromagnetic Oscillations. 2012.
Aalto-DD97/2012	Turunen, Ville T. Morph-Based Speech Retrieval: Indexing Methods and Evaluations of Unsupervised Morphological Analysis. 2012.
Aalto-DD117/2012	Vierinen, Juha On statistical theory of radar measurements. 2012.
Aalto-DD137/2012	Huopaniemi, Ilkka Multivariate Multi-Way Modeling of Multiple High-Dimensional Data Sources. 2012.
Aalto-DD137/2012	Paukkeri, Mari-Sanna Language-and domain independent text mining. 2012.
Aalto-DD133/2012	Ahlroth, Lauri Online Algorithms in Resource Management and Constraint Satisfaction. 2012.
Aalto-DD158/2012	Virpioja, Sami Learning Constructions of Natural Language: Statistical Models and Evaluations. 2012
Aalto-DD20/2013	Pajarinen, Joni Planning under uncertainty for large-scale problems with applications to wireless networking. 2013.

Aalto-DD 47/2013

ISBN 978-952-60-5074-4
ISBN 978-952-60-5075-1 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Information and Computer Science
www.aalto.fi

BUSINESS +
ECONOMY

ART +
DESIGN +
ARCHITECTURE

SCIENCE +
TECHNOLOGY

CROSSOVER

DOCTORAL
DISSERTATIONS