One-class Classification in the presence of Point, Collective, and Contextual Anomalies

vorgelegt von Dipl.-Ing. Nico Görnitz geb. in Berlin

von der Fakultät IV—Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
-Dr. rer. nat.-

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender: Prof. Dr. Benjamin Blankertz
Gutachter: Prof. Dr. Klaus-Robert Müller
Gutachter: Prof. Dr. Manfred Opper
Gutachter: Prof. Dr. Marius Micha Kloft

Tag der wissenschaftlichen Aussprache: 9. November 2018

Berlin 2019

To my family.

Acknowledgements

This work was carried out during the years 2010-2017 in the Computational Biology Group at the Friedrich Miescher Laboratory of the Max Planck Society in Tübingen, Germany, the eScience Group of Microsoft Research in Los Angeles, US, and, foremost, the Machine Learning Group at the Berlin Institute of Technology (TU Berlin), Germany.

These past years were a period of personal growth with constant exchange of ideas, the love of learning in an environment encouraging creativity and insight. I am grateful for the experiences to travel and meeting smart and interesting people along this path. I'd like to express my deepest gratitude to my advisors Klaus-Robert Müller and Marius Kloft, who made it possible for me to carry out this work and who contributed to this thesis with their time, ideas and energy.

If it wasn't for Marius Kloft, who attracted me to the world of research with his love for teaching, wit and passion for machine learning research, I may not have pursued this path. Besides him, it was the mentorship of Shinichi Nakajima, Ulf Brefeld, Sören Sonnenburg, Gunnar Rätsch, and Konrad Rieck that helped forming my scientific ideas, which eventually became my Ph.D. thesis. A very special thanks I'd like to address to Klaus-Robert Müller who, in addition, was an avid supporter throughout the whole time.

This work would have not been possible without the fruitful collaboration of my dear colleagues. I took particular pleasure in working with Luiz Alberto Lima, Marina Vidovic, Alexander Bauer, and Seunghak Lee. Furthermore, I would like to thank Jonas Behr, Alexander Binder, Andreas Ziehe, Christian Widmer, Irene Dowding, Regina Bohnert, Georg Zeller, Vipin Sreedharan, Jamal Nasir, Philipp Drewe, Gregoire Montavon, Christoph Lippert, Anne Porbadnigk, Mikio Braun, André Kahles, Géraldine Jean, and Bettina Mieth.

I especially thank my spouse Christina and my son Max who suffered the most from my obsession with unfinished chapters and last-minute changes to the manuscript, for their patience and support. I would like to thank my parents and my siblings, Mandy and Linda, for supporting me in every conceivable way throughout the years.

Finally, I acknowledge financial support of the German Bundesministerium für Bildung und Forschung (BMBF), under the consecutive projects ALICE I and II (01IB10003B and 01IB15001B respectively). I also acknowledge the support by the German Research Foundation (DFG) through the grant DFG MU 987/6-1 and RA 1894/1-1.

Abstract

Anomaly detection has a prominent position in the processing pipeline of any real-world data-driven application. Its central goal is to detect and separate valid data points from malicious—anomalous—ones such that the cleaned data set can be processed further. In many applications, anomalies are even the prime objects of interest and need to be exposed early in order to avoid loss, e.g. in credit card fraud detection.

One-class classification is a machine learning concept that is especially suited for the anomaly detection problem. Intrinsically unsupervised, it aims at providing a concise description of a given data set such that data points generated by a different process can be detected accurately. Prominent machine learning models for one-class classification are one-class support vector machines and the closely related support vector data descriptions.

The contribution of this thesis is the extension of those methods to cope with different scenarios of anomalies:

Point Anomalies Assuming that anomalies are scarce and occur independently of each other, methods for controlling the sparsity of the found solutions in terms of single independent features and groups of features are derived.

Collective Anomalies In this scenario anomalies are assumed to appear as groups of measurements instead of single entries. Techniques from structured output learning are (i) extended to cope with large-scale problems, (ii) employed to derive an unsupervised anomaly detector for groups of measurements that exhibit a latent dependency structure.

Contextual Anomalies Anomalies appear only in specific contexts and data is supposed to carry two signals that contain behavioral and contextual information. Contributions in this scenario consider latent class dependencies and are threefold: (i) the derivation of a method capable of detecting latent class contextual anomalies, (ii) theoretical insight reveal k-means as a special case, and (iii) a method for learning with latent class dependencies when an additional structure is imposed on the latent variables.

The proposed methods are empirically analyzed on a variety of different applications ranging from gene finding to porosity estimation to brain computer interfaces showing promising performance when compared to baseline methods.

Zusammenfassung

Anomalieerkennung nimmt im Verarbeitungsablauf jeder realen Daten-getriebenen Anwendung eine wichtige Stellung ein. Ihre zentrale Aufgabe ist es gültige Daten zu erkennen und von Ungültigen, anomalen Daten, zu trennen sodass der so bereinigte Datensatz weiter verarbeitet werden kann. In vielen Anwendungen sind sogar die Anomalien die interessantesten Objekte und sollten so früh wie möglich erkannt werden um möglichen Verlusten vorzubeugen wie zum Beispiel bei der Prävention von Kreditkartenbetrug.

Einklassen-Klassifikation ist ein Konzept des Maschinellen Lernens, welches besonders geeignet ist um Anomalien zu detektieren. Es handelt sich um intrinsisch unüberwachte Lernverfahren welche darauf abzielen eine genaue Beschreibung eines gegebenen Datensatzes zu liefern, so dass Datenpunkte, die von einem anderen Prozess erzeugt wurden, akkurat erkannt werden können. Die wichtigsten Vertreter dieser Zunft sind die Einklassen-SVM sowie die mir ihr eng verwandte SVDD.

Diese Arbeit leisten einen Beitrag um diese Methoden so zu erweitern, dass sie mit den folgenden, allgemeinen Anomalieszenarien umgehen können:

Punktanomalien Wir nehmen an, dass Anomalien selten sind und unabhängig voneinander auftreten. Wir entwickeln Methoden, welche die Spährlichkeit, die Anzahl der Nullstellen in der Lösung, kontrollieren, basierend dabei auf einzelnen Merkmalen oder Gruppen von Merkmalen.

Kollektivanomalien In diesem Szenario wird angenommen, dass Anomalien in Gruppen von Messungen auftreten anstatt als isolierte Einzelmessung. Wir werden Techniken vom Strukturlernen (i) erweitern, um mit grossen Datenmeengen umgehen zu können, und (ii) anwenden um einen unüberwachten Anomalieerkenner für Gruppen von Messungen zu entwickeln, wenn diese Messungen eine latente Abhängigkeitsstruktur besitzen.

Kontextanomalien Anomalien erscheinen nur in gewissen Kontext und es wird angenommen, dass Daten aus Verhaltens- und Kontextinformationen bestehen. Beiträge in diesen Szenario beschränken sich auf latente Klassenstruktur und sind dreigeteilt: (i) eine Methode zur Erkennung von Anomalien mit latenter Klassenstruktur wird vorgestellt, (ii) theoretische Einsichten, welche zeigen das k-means ein Spezialfall ist, werden vorgestellt, und (iii) eine Methode die mit latenter Klassenstruktur umgehen kann, wenn diese wiederum eine eigene Abhängigkeitsstruktur besitzt, wird entwickelt.

Die vorgestellten Methoden werden empirisch analysiert. Die Anwendungen reichen dabei von Generkennung über Hirn-Computer-Schnittstellen zu Porositätserkennung. Dabei zeigen die vorgestellten Methoden im Vergleich zu Standardmethoden vielversprechende Resultate.

Contents

Chapter 1
Introduction

> 'I provide a service that is unique in this world,' said Dirk. 'The term "holistic" refers to my conviction that what we are concerned with here is the fundamental interconnectedness of all –'
>
> *Douglas Adams (Dirk Gently's Holistic Detective Agency)*

With the abundance of data nowadays, automated tools handling the sheer amount of available and further incoming data are a necessity. Over the past decade, machine learning concepts have become invaluable not only for researchers but also for practitioners in the industry to tackle complex, data-driven problems. Generally phrased, the goal of machine learning is to learn unknown concepts from data given provided label information.

The goal of anomaly detection is, however, to separate valid data points from malicious, anomalous ones. It has a prominent position in the processing pipeline of any real-world data-driven application. Unfortunately, tagging data as anomalous depends very much on the application at hand and can not be generalized easily. Moreover, due to its scarcity and novelty, label information is generally rare, incomplete, or missing altogether. However, finding anomalies is of vital interest in many applications, as they oftentimes translate directly to actionable items, situations that need immediate response such as engine failures. A common approach for detecting the *unlikeliness* is by describing the normal behavior of a system and finding deviations thereof [1–3].

One-class classification [4, 5] is a machine learning concept that is especially suited for anomaly detection. Intrinsically unsupervised, it aims at providing a tight boundary—a concise description—of a given data set such that data points generated by a different process can be detected accurately. Two of the most prominent machine learning models for one-class classification are one-class support vector machines and the closely related support vector data descriptions [6–9]. These methods have been successfully applied to a large number of problems including network intrusion detection [10–12], hyperspectral imagery [13], surface modeling [14], and neurosciences [15, 16].

Despite their success, most machine learning methods treat data points and hence, anomalies, as independent events without taking into account the dependency structure even if it is known. Analyzing data with dependency structure is a challenging effort in machine learning and signal processing that has many important applications; for example, in mobile communication [17], earthquake prediction [18], geosciences data analysis [19, 20], traffic flow modeling [21], and bioinformatics [22].

This thesis, we study the three different classes—from independent events to interconnected entities—of anomalies and develop methods for various settings based on the one-class classification principle:

Point Anomalies Assuming that anomalies are scarce and occur independently of each other, methods for controlling the sparsity of the found solutions in terms of single independent features and groups of features are derived.

Collective Anomalies In this scenario anomalies are assumed to appear as groups of measurements instead of single entries. Techniques from structured output learning are (i) extended to cope with large-scale problems, (ii) employed to derive an unsupervised anomaly detector for groups of measurements that exhibit a latent dependency structure.

Contextual Anomalies Anomalies appear only in specific contexts and data is supposed to carry two signals that contain behavioral and contextual information. Contributions in this scenario consider latent class dependencies and are threefold: (i) a method capable of detecting latent class contextual anomalies, (ii) theoretical insights reveal k-means as special case, and (iii) a method for learning with latent class dependencies when an additional structure is imposed on the latent variables.

This dissertation derives and discusses anomaly detectors based on the one-class classification paradigm for settings involving point, collective, and contextual anomalies.

1.1 A Roadmap through this Thesis

At high level, this thesis is divided into four distinct parts. In the compulsory **Background** part, basic machine learning concepts from kernel machines and optimization as well as an overview of anomaly detection and one-class classification are presented. The following three parts—**Point Outliers**, **Collective Outliers**, and **Contextual Outliers**—contain the original contributions of this thesis.

Chapter 2: Foundations of Machine Learning This chapter reviews fundamental concepts from machine learning that are necessary for the understanding of this thesis. In specific, definitions and theorems for kernel methods and basic (convex) optimization concepts are introduced.

Chapter 3: Anomaly Detection and One-class Classification A comprehensive introduction to anomaly detection and one-class classification is given in this chapter. History and definition, trends and categorization of anomaly detection techniques presented. Evaluation strategies and corresponding measurements are discussed.

Chapter 4: Sparsity-inducing Learning The focus in this chapter is on feature regularization. In specific, we describe techniques for controlling sparsity of singleton features as well as groups of features. The proposed algorithms are applied to applications in BCI-EEG and authorship attribution. No special requirements are imposed on the nature of anomalies.

Chapter 5: Learning with Structured Data In this chapter, anomalies are supposed to appear in groups of measurements rather than single entries. This challenging problem is tackled employing structured output learning concepts. We present a corresponding large-scale optimization scheme for structured output learning and an unsupervised one-class classifier tailored to this scenario. Challenging applications on computational biology problems are presented.

Chapter 6: Learning with Latent Class Dependencies This chapter assumes that anomalies appear only in specific contexts. We develop methods that are capable of contextual anomaly detection when the context comes as latent class dependencies and reveal important special cases. Further, extensions to scenarios when contextual variables exhibit known structural dependencies are proposed.

Chapter 7: Conclusion This chapter concludes this thesis with a brief discussion and outlook.

1.2 Own Contributions and Publications

I had the pleasure to work closely with very skilled scientist that excel in their respective fields. Many times they had a specific problem for an application in mind which enabled me to tailor methods especially to their needs. Generally speaking, idea, derivation, optimization, and implementation of methods are my contributions. This also holds for empirical evaluations on artificially generated data. In all of the presented work, I have been lead author or joint first author (with the exception of Nasir, Görnitz, and Brefeld). In the following, I detail the contributions.

Part II Empirical results on authorship attribution as well as the description thereof, has been done by Jamal Nasir. Experiments on BCI-EEG data including the analysis and result discussion is based on the work of Anne Porbadnigk.

> **Görnitz, N.**, Kloft, M., Rieck, K., Brefeld, U., "Toward Supervised Anomaly Detection", *Journal of Artificial Intelligence Research (JAIR)*, vol. 46, pp. 235–262, 2013
>
> Porbadnigk, A., **Görnitz, N.**, Kloft, M., Müller, K.-R., "Decoding Brain States during Auditory Perception by Supervising Unsupervised Learning.", *Journal of Computing Science and Engineering (JCSE)*, vol. 7, no. 2, pp. 112–121, 2013
>
> Nasir, J. A., **Görnitz, N.**, Brefeld, U., "An Off-the-shelf Approach to Authorship Attribution", in *International Conference on Computational Linguistics (COLING)*, 2014, pp. 895–904

Part III Preparation of the data set and consulting on biological research questions has been done by Georg Zeller. Marius Kloft derived the generalization bounds and the dual formulation presented in Section 5.3. All experiments where carried out by myself.

> **Görnitz, N.**, Braun, M., Kloft, M., "Hidden Markov Anomaly Detection", in *International Conference on Machine Learning (ICML)*, 2015, pp. 1833–1842
>
> **Görnitz, N.**, Widmer, C., Zeller, G., Kahles, A., Sonnenburg, S., Rätsch, G., "Hierarchical Multitask Structured Output Learning for Large-scale Sequence Segmentation", in *Advances in Neural Information Processing Systems (NIPS)*, 2011, pp. 2690–2698
>
> Zeller, G., **Görnitz, N.**, Kahles, A., Behr, J., Mudrakarta, P., Sonnenburg, S., Rätsch, G., "mTim: rapid and accurate transcript reconstruction from RNA-Seq data", *ArXiv*, 2013

Part IV Marius Kloft carried out the generalization bounds presented in Section 6.2. Application to simulated and real porosity prediction has been done by Luis A. Lima. Experiments on BCI data including the analysis and result discussion is based on the work of Anne Porbadnigk. Figures and application of the proposed method to the data have been done by myself.

Görnitz, N., Porbadnigk, A. K., Kloft, M., Binder, A., Sannelli, C., Braun, M., Müller, K.-R., "When brain and behavior disagree: A novel ML approach for handling systematic label noise in EEG data", in *Machine Learning and Interpretation in Neuroimaging Workshop (MLINI)*, 2013

Görnitz, N., Porbadnigk, A. K., Binder, A., Sanelli, C., Braun, M., Müller, K.-R., Kloft, M., "Learning and Evaluation in Presence of Non-i.i.d. Label Noise", in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 33, 2014, pp. 293–302

Porbadnigk, A. K., **Görnitz, N.**, Sannelli, C., Binder, A., Braun, M., Kloft, M., Müller, K.-R., "When Brain and Behavior Disagree: Tackling systematic label noise in EEG data with Machine Learning", in *IEEE International Winter Workshop on Brain-Computer Interface (BCI)*, 2014

Porbadnigk, A. K., **Görnitz, N.**, Sannelli, C., Binder, A., Braun, M., Kloft, M., Müller, K.-R., "Extracting latent brain states — Towards true labels in cognitive neuroscience experiments", *NeuroImage*, vol. 120, pp. 225–253, 2015

Görnitz, N., Lima, L. A., Varella, L. E., Müller, K.-R., Nakajima, S., "Transductive Regression for Data with Latent Dependency Structure", *IEEE Transactions on Neural Networks and Learning (TNNLS)*, 2017

Görnitz, N., Lima, L. A., Müller, K.-R., Kloft, M., Nakajima, S., "Support vector data descriptions and k-means clustering: one class?", *IEEE Transactions on Neural Networks and Learning (TNNLS)*, 2017

Lima, L. A., **Görnitz, N.**, Varella, L. E., Vellasco, M., Müller, K.-R., Nakajima, S., "Porosity Estimation by Semi-supervised Learning with Sparsely Available Labeled Samples", *Computers & Geosciences*, vol. 106, pp. 33–48, 2017

I Background

Chapter 2

Foundations of Machine Learning

This chapter introduces the machine learning concepts needed for understanding the methods developed in this thesis. The overall focus on *classic* concepts and will leave out some of the more prominent techniques nowadays (i.e. deep learning) even though they are likely to play an important role in the near future for one-class classification and anomaly detection. However, these techniques will only be discussed marginally in this thesis. We hereby start with kernel methods and the kernel trick in specific and go on with the basics of (non-)convex optimization theory.

2.1 Kernel Methods

We start the discussion on kernel methods by giving a simple example: consider a linear model $f(\mathbf{x}) = \langle \mathbf{w}, \phi(\mathbf{x}) \rangle$ with a possibly very high dimensional parameter vector $\mathbf{w} \in \mathcal{F}$ and a feature vector $\phi : \mathcal{X} \to \mathcal{F}$ of data point $\mathbf{x} \in \mathcal{X}$ of corresponding dimension. Here, instead of accessing our data points directly, we would like to add a little flexibility by allowing an arbitrary transformation ϕ which maps the data points from the input space $\mathcal{X}$ to some feature space $\mathcal{F}$ (cf. Figure 2.1) which, hopefully, makes the problem more accessible. Further, assume that we are given a sample of size $i = 1, \ldots, n$ of i.i.d. data points $\mathbf{x}_i \in \mathcal{X}$ and corresponding labels $y_i \in \mathbb{R}$ which we will use to fit our parameter vector to produce the least squared error on that given sample,

$$\mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w} \in \mathcal{F}} \sum_{i=1}^{n} \ell(\mathbf{w}, \mathbf{x}_i, y_i) \quad \text{with} \quad \ell(\mathbf{w}, \mathbf{x}, y) := \frac{1}{2}(y - \langle \mathbf{w}, \phi(\mathbf{x}) \rangle)^2 \,.$$

For the sake of simplicity, we employ stochastic gradient descent as an optimization technique. Therefore, at time step t we pick a data point $\mathbf{x}_t$ and the corresponding label y_t from our training sample and update the parameter vector according to the following formula (with $\mathbf{w}^0 = 0$):

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \frac{\partial \ell(\mathbf{w}_t, \mathbf{x}_t, y_t)}{\partial \mathbf{w}} = \mathbf{w}_t + \eta(y_t - \langle \mathbf{w}_t, \phi(\mathbf{x}_t) \rangle)\phi(\mathbf{x}_t) = \mathbf{w}_t + \alpha_t \phi(\mathbf{x}_t) \,.$$

Since our initial objective is convex, gradient descent will (with carefully chosen η) find a local, hence, global minimum. Rather surprisingly, it will attain the optimal value while not leaving the span of the data. Therefore, the optimal parameter vector $\mathbf{w}^*$ can be expressed

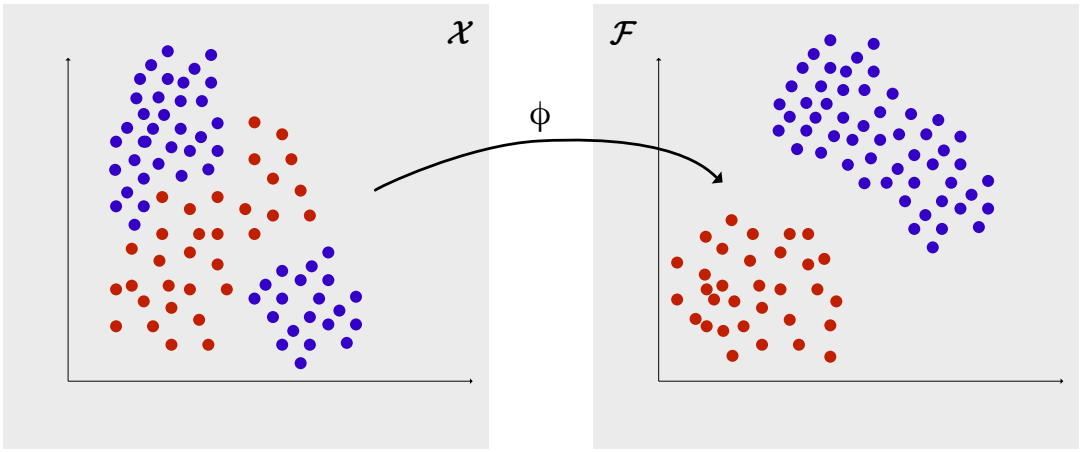

Figure 2.1 – The feature mapping ϕ maps the data points (red and blue dots) from the input space $\mathcal{X}$ (left) to some feature space $\mathcal{F}$ (right). As can be seen, the goal is to simplify the problem for subsequent analysis, e.g. classification.

as a weighted sum of feature vectors,

$$\mathbf{w}^* = \sum_{t=1}^{T} \alpha_t \phi(\mathbf{x}_t) \,. \tag{2.1}$$

Given this expansion, we can also re-write the inner product between parameter vector and feature vector with

$$\langle \mathbf{w}, \phi(\mathbf{x}) \rangle = \sum_t \alpha_t \langle \phi(\mathbf{x}_t), \phi(\mathbf{x}) \rangle = \sum_t \alpha_t k(\mathbf{x}_t, \mathbf{x}) \,,$$

where $k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$ is called a kernel. This gives an alternative view on the above optimization problem, where the key is to find weightings of similarities, as encoded with inner products, between data points. Moreover, we do not need to know the inner workings of the feature map ϕ anymore. To summarize this little example, we can (i) express the optimal solution $\mathbf{w}^*$ as a weighted sum of feature vectors and (ii) access the inner product in terms of similarities between data points. The former property is known as the *representer theorem* and the latter gives rise to the *kernel trick* [33–36]. In fact, these properties form the basis of many successful machine learning models such as support vector machines (SVMs) [37, 38].

Lets now discuss these these findings in more detail. Any discussion about kernel methods and the kernel trick in specific needs to be split in three parts:

i the feature map ϕ which maps a data point from its input space $\mathcal{X}$ into some higher dimensional feature space $\mathcal{F}$;

ii the kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which encodes similarities between feature vectors of corresponding data points $k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$;

iii the reproducing kernel Hilbert space (RKHS) $\mathcal{F}$, a space of functions endowed with a norm.

A very general definition of a kernel is given in Mohri, Rostamizadeh, and Talwalkar.

Definition 1 (Kernels [39]). *A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a kernel over $\mathcal{X}$.*

Albeit this definition allows to encode arbitrary functions, in order to ensure that a decomposition into feature vectors $k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$ exists, k needs to satisfy the following condition:

Theorem 1 (Mercer's condition [39]). *Let $\mathcal{X} \subset \mathbb{R}^N$ be a compact set and let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous and symmetric function. Then, k admits a uniformly convergent expansion of the form*

$$k(\mathbf{x}, \mathbf{x}') = \sum_{i=0}^{\infty} a_i \phi_i(\mathbf{x}) \phi_i(\mathbf{x}') \,,$$

with $a_i > 0$ iff for any square integrable function c ($c \in L_2(\mathcal{X})$), the following condition holds:

$$\int \int_{\mathcal{X} \times \mathcal{X}} c(\mathbf{x}) c(\mathbf{x}') k(\mathbf{x}, \mathbf{x}') d\mathbf{x} d\mathbf{x}' \geq 0 \,.$$

We present another, slightly more general and approachable definition (cf. discussion in [39]) which ensures the existence of the decomposition.

Definition 2 (Positive definite symmetric kernels [39, 40]). *A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be positive definite symmetric (PDS) if for any $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subseteq \mathcal{X}$, the matrix $K = [k(\mathbf{x}_i, \mathbf{x}_j)]_{ij} \in \mathbb{R}^{n \times n}$ is symmetric positive semidefinite (SPSD).*

The matrix K is called the kernel matrix or the Gram matrix associated to K. Hence, we have the following relation between feature maps and kernels: for a specific choice of PDS kernel k, the feature map ϕ is fixed (up to rotation). For a specific choice of feature map ϕ, the corresponding kernel k is fixed. Albeit many possibilities exists for defining an proper kernel k, the following kernels appear frequently:

Linear kernel $k(\mathbf{x}, \mathbf{x}') := \langle \mathbf{x}, \mathbf{x}' \rangle$;

Polynomial kernel $k(\mathbf{x}, \mathbf{x}') := (\langle \mathbf{x}, \mathbf{x}' \rangle + c)^d$ with $c > 0$ and $d \in \mathbb{N}$;

Radial basis function (RBF) kernel $k(\mathbf{x}, \mathbf{x}') := \exp(-\frac{1}{2\sigma^2} \|\mathbf{x} - \mathbf{x}'\|)$.

Now that the relation between feature maps and kernels is clear, we only need a relation between those entities with their respective reproducing kernel Hilbert space (RKHS) which comes in the form of the following theorem as given in Mohri, Rostamizadeh, and Talwalkar:

Theorem 2 (Reproducing kernel Hilbert space (RKHS) [39]). *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PDS kernel. Then, there exists a Hilbert space $\mathcal{F}$ and a mapping ϕ from $\mathcal{X}$ to $\mathcal{F}$ such that:*

$$\forall \, \mathbf{x}, \mathbf{x}' \in \mathcal{X}, \quad k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle.$$

Furthermore, $\mathcal{F}$ has the following property known as the reproducing property:

$$\forall \, f \in \mathcal{F}, \; \forall \mathbf{x} \in \mathcal{X}, \quad f(\mathbf{x}) = \langle f, k(\mathbf{x}, \cdot) \rangle.$$

$\mathcal{F}$ is called a reproducing kernel Hilbert space (RKHS) associated to k.

Finally, we can give a concise description of the existence of the expansion in Eq. (2.1). The original representer theorem was presented by Kimeldorf and Wahba [41] and later refined by, e.g. Schölkopf [42]. In general, the representer theorem states that if a given optimization problem can be rephrased in a specific (very general) form, then the optimal solution of this optimization problem must live in the span of the data.

Theorem 3 (Representer Theorem [39]). *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PDS kernel and $\mathcal{F}$ its corresponding RKHS. Then, for any non-decreasing function $G : \mathbb{R} \to \mathbb{R}$ and any loss function $L : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, the optimization problem*

$$\operatorname*{argmin}_{f \in \mathcal{F}} G(\|f\|_{\mathcal{F}}) + L\left(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)\right)$$

admits a solution of the form $f^* = \sum_{i=1}^n \alpha_i k(\mathbf{x}_i, \cdot) = \sum_{i=1}^n \alpha_i \phi(\mathbf{x}_i)$. *If* G *is further assumed to be increasing, then any solution has this form.*

2.2 Optimization

We give a short introduction to optimization based on the great book of Boyd and Vandenberghe [43]. First, we introduce the necessary concepts while focusing on the general problem. We then turn to the special case of convex optimization and end this section with a discussion on non-convex problems.

At the core of every machine learning method there is an objective that needs to be optimized subject to some constraints. This means that whatever the intent of a method is, it should be expressible as a objective function of some adjustable parameters $\mathbf{x}$:

$$\begin{aligned} \min_{\mathbf{x}} \; & f_0(\mathbf{x}) \\ \text{s.t. } & f_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, m \\ & h_j(\mathbf{x}) = 0, \quad j = 1, \ldots, p\,. \end{aligned} \tag{2.2}$$

We refer to the feasible domain of $\mathbf{x}$ (assumed to be non-empty) as $\mathcal{D}$ and to its optimal value as p^*. The above problem, the primal optimization problem, is a constrained optimization problem which is generally hard to handle. An unconstrained version can be obtained using the notion of Lagrangian functions with penalties on constraint violations.

Definition 3 (Lagrangian)**.** *A function* $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ *of the form*

$$L(\mathbf{x}, \lambda, \nu) = f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i h_i(\mathbf{x})\,,$$

is called the Lagrangian associated with the Problem (2.2)*.*

The variables ν and λ are called the Lagrange multiplier or dual variables.

Definition 4 (Lagrange dual function)**.** *Let* $g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ *be the minimum value of the Lagrangian over* $\mathbf{x}$*. Then for any* $\lambda \in \mathbb{R}^m$ *and* $\nu \in \mathbb{R}^p$

$$g(\lambda, \nu) = \inf_{\mathbf{x} \in \mathcal{D}} L(\mathbf{x}, \lambda, \nu) = \inf_{\mathbf{x} \in \mathcal{D}} \left(f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i h_i(\mathbf{x}) \right)\,.$$

An important property of the above definition is that for any $\lambda \geq 0$ and any ν the Lagrange dual function is a lower bound on the primal optimum p^*, $g(\lambda, \nu) \leq p^*$. The difference between both entities is called the duality gap. If the duality gap is zero, then strong duality holds otherwise we speak of weak duality (which always holds). Maximizing $g(\lambda, \nu)$ wrt. $\lambda \geq 0$ and ν is referred to as the dual problem of Eq. (2.2). There is a strong relation between primal and dual problem that is condensed in the Karush-Kuhn-Tucker (KKT) conditions.

Theorem 4 (KKT conditions)**.** *Let* f_i *and* h_j *for* $i = 0, \ldots, m$ *and* $j = 1, \ldots, p$ *be differentiable. Let further* $\mathbf{x}^*$ *and the pair* (λ^*, ν^*) *be any primal and dual optimal points with zero duality gap. Thus*

$$\nabla f_0(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(\mathbf{x}^*) + \sum_{i=1}^p \nu_i^* \nabla h_i(\mathbf{x}^*) = 0 \quad (stationarity),$$

$$
\begin{aligned}
f_i(\mathbf{x}^*) \leq 0, \quad & i = 1, \ldots, m && (primal\, feasibility),\\
h_i(\mathbf{x}^*) = 0, \quad & i = 1, \ldots, p && (primal\, feasibility),\\
\lambda^* \geq 0, \quad & i = 1, \ldots, m && (dual\, feasibility),\\
\lambda^* f_i(\mathbf{x}^*) = 0, \quad & i = 1, \ldots, m && (complementary\, slackness)
\end{aligned}
$$

are called the Karush-Kuhn-Tucker (KKT) conditions.

We further introduce the concept of the convex conjugate which sometimes helps to generalize certain problems.

Definition 5 (Conjugate function)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$. The function $f^* : \mathbb{R}^n \to \mathbb{R}$, defined as*

$$f^*(\mathbf{y}) = \sup_{\mathbf{x} \in dom f} \left(\langle \mathbf{y}, \mathbf{x} \rangle - f(\mathbf{x}) \right),$$

is called the conjugate of the function f.

It is also known as the convex conjugate or Legendre-Fenchel convex conjugate. If f is differentiable, then f^* is also called the Legendre transform.

The above formulations hold for any optimization problem. However, in the case of convex optimization there are certain properties that are very desirable. Basically, there are two amazing things about convex optimization. First, it is guaranteed to find the *global* optimum in reasonable amount of time. This is what we ultimately care about. The second thing is that strong duality holds. It means that we can check optimality but it allows also optimize the dual problem which might give additional insights into the application at hand or it makes the optimization more efficient. In order to qualify as a convex optimization problem, we need the constraints to fulfill some basic properties (called constraint qualifications). Further, f_i must be convex, h_j must be affine and hence, the feasible set $\mathcal{D}$ must be convex. Hence, we need two more definitions about convex sets and convex functions.

Definition 6 (Convex set)**.** *A set C is said to be convex if the line segment between any two points in C lies in C, i.e. if for any $\mathbf{x}_1, \mathbf{x}_2 \in C$ and any σ with $0 \leq \sigma \leq 1$,*

$$\sigma \mathbf{x}_1 + (1 - \sigma)\mathbf{x}_2 \in C$$

must hold.

Definition 7 (Convex function)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if $dom f$ is a convex set and if for all $\mathbf{x}_1, \mathbf{x}_2 \in dom f$, and σ with $0 \leq \sigma \leq 1$,*

$$f(\sigma \mathbf{x}_1 + (1 - \sigma)\mathbf{x}_2) \leq \sigma f(\mathbf{x}_1) + (1 - \sigma) f(\mathbf{x}_2).$$

However, the question how to find the optimal solution remains. A very simple and general approach is given in Algorithm 1 where we need to (i) find a descent direction (e.g. negative gradient), (ii) find a step length $\theta \geq 0$ (e.g. line search), and (iii) a stopping criterion. Most famous, and probably most widely used, is gradient descent. There are, of course, many

Algorithm 1 General descent method.

while stopping criterion not satisfied **do**
  Determine a descent direction $\mathbf{d}$
  Choose a step length $\theta \geq 0$
  Update $\mathbf{x}^{t+1} = \mathbf{x}^t + \theta \mathbf{d}$
end while

more elaborate optimization algorithms for convex optimization problems [43–48] which might be general purpose or exploit certain special properties (e.g. sparsity). Broadly they can be categorized into first order methods (e.g. gradient descent) and higher order methods (e.g. Newtons method), i.e. where the second derivative is needed. A broad class of algorithms is contained in the proximal methods [49].

Some of those methods might be even applicable to non-convex optimization. However, in this case the best one can get is generally a *local* optimal solution. Certain methods that are based on special problem structure, e.g. assuming that the objective function can be decomposed into a difference of convex functions, showed promising results [50–52]. Despite all the progress made, due to the increased attention to highly non-convex problems imposed by deep neural nets, the working horse of todays optimization algorithms is again... gradient descent.

Chapter 3

Anomaly Detection and One-class Classification

In this chapter, we discuss the fundamentals of anomaly detection and subsequently one-class classification. We start by framing the historical context and state the basic definitions, most fundamental work, the various types of anomalies and learning settings. Further, we attempt to coarsely categorize models and settings before we turn to an in-depth discussion on one-class classification. We proceed by talking about the evaluation and interpretation of anomaly detection and finally, we conclude with an outlook on current and future challenges.

3.1 Introduction

Traditionally, anomaly detection is a very application-driven research subject and methods have been proposed and studied for several decades in statistics, machine learning, data mining, and database systems [2]. Anomaly detection is used nowadays as an umbrella term for wide variety of techniques, settings, and approaches which all share a common goal. This commonality is usually defined over the *unusualness* of observations in a given data set. Hence, the goal is to find, remove, describe or extract (parts of) observations that deviate from rest of the data set significantly. A widely accepted, very general definition of what an anomaly is, was given by Hawkings [53], 1980:

> *An outlier [=anomaly] is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.*

However, there have also been predecessors, e.g. Grubbs, 1969 [54]:

> *An outlying observation, or "outlier" [=anomaly], is one that appears to deviate markedly from other members of the sample in which it occurs.*

These quotations show that the field of anomaly detection is indeed quite old for computer science and statistics. However, importance of anomaly detection sky rocketed only quite recently which was fueled by the advent of the internet, online services, big data and the corresponding economical impact. As of today, practically *all* online services rely heavily on a mix of anomaly detection methods (e.g. fraud detection, intrusion detection, etc. pp.).

Depending on the context and application, the term "anomaly" is often replaced by other substitutions such as outlier, exception, peculiarity, surprise, noise, abnormalities, deviants, discordants. Notably, anomaly and outlier are used interchangeably throughout most of the literature whereas other names might indicate specialized settings (i.e. noise). Generally, three different types of anomalies are considered:

- **point anomalies** are data points that appear isolated from the bulk of the data (cf. Fig. 3.1)
- **contextual anomalies** (sometimes also *conditional* anomalies) are data points whose values itself are only anomalous in a specific contextual relation (cf. Fig. 3.2)
- **collective anomalies** consist of a sequence of data points that only as a group, and not as individual points, can be tagged anomalous (cf. Fig. 3.3)

Of the above, point anomalies have been studied more extensive and many methods readily assume that data points come as independent instances. If, however, data exhibits strong dependency structure, handling data points as independent instance might not suffice anymore. Anomaly detection has tremendous scope for research especially in this area [25, 55–57]. In some situations, depending on the requirements and the analyst's understanding of the problem, it might suffice to phrase collective and contextual anomalies as point anomalies.

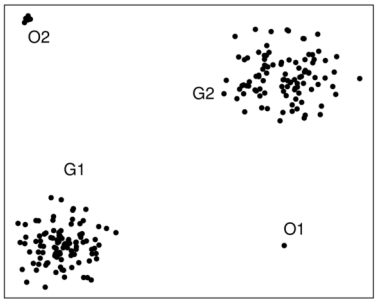

Figure 3.1 – An example of point anomalies. O1 and O2 are outliers which are well isolated from the large data clusters G1 and G2 (distributed under CC BY-SA 4.0 license).

Most anomaly detection methods are loosely based on the Hawkins' outlier and approach the problem by finding strong deviations in data. Few approaches assume that anomalies and/or nominal data can be modeled by the analyst (sometimes referred to as well-defined anomaly distribution, WDAD). These are rare though.

In the end, the output of anomaly detection method should give a clear answer of whether or not a given instance is anomalous. This seems like a binary classification task which, indeed, is handled this way in some cases [58]. However, the task of finding anomalies is complex and in most realistic cases it can not be expected to perfectly separate anomalous data from the nominal data. Furthermore, there are often various degrees of anomalousness within data (e.g. noise is considered a weak anomaly). To cope with more realistic scenarios, most analyst's favor a continuous output such that data points can be ranked from the most nominal to most anomalous data point.

Many methods inherently depend on continuous attributes, they often need further preprocessing to be normalized between a specific range or *whitened* (with features having the same standard deviation). There are, however, other data types which are frequently encountered in data sets. Foremost, binary attributes which can only take on two values $\{0, 1\}$ and (un-)ordered categorical attributes which can take $n \in \mathbb{N}^+$ possible values. If needed, categorical attributes can be converted into a sparse binary vector $\mathbf{x}$ of dimension n with

$\|\mathbf{x}\|_2 = 1$. Binary attributes can further be converted into continuous attributes by, e.g. principle component analysis (PCA).

In search for a Hawkins' outlier, methods need to find deviations from the nominal data while further knowledge about the process of anomalies can not be expected. However, in many practical applications there will be prior knowledge either in terms of expert insights or in existing anomaly samples.

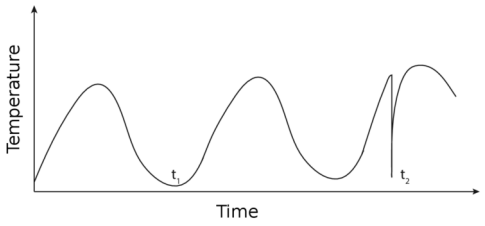

Figure 3.2 – An example of a contextual anomaly. Note that the value of the data instance at t_1 is not anomalous, but in the context of t_2 it is. (distributed under CC BY-SA 4.0 license).

A fully supervised approach assumes that for each data point in the training set a corresponding label is present. Standard binary classifier like support vector machines or logistic regression can be employed and tested accordingly. This approach does work well if *all* anomaly classes are well sampled. Furthermore, usual settings lead to very unbalanced data sets with often $> 99\%$ nominal data. In such cases, binary classifier might return trivial solutions.

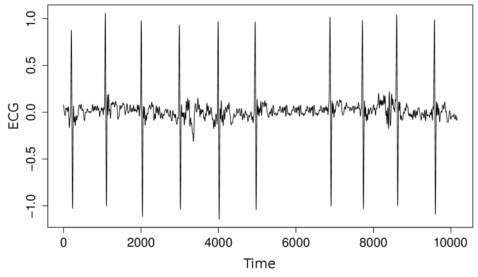

Figure 3.3 – An example of a collective anomaly taken from human electrocardiogram. Note that groups of nominal data points vary in their values significantly, only the absence of a whole group at $t = 6000$ of data points forms an anomaly (distributed under CC BY-SA 4.0 license).

Half-way between supervised and unsupervised learning is semi-supervised learning which can be further split into a bunch of subsettings. In machine learning the general definition for this setting is simply that additional to labeled examples, there are some unlabeled examples. If only inferring the labels for those unlabeled examples is the goal, then this can be tagged a transductive setting. However, in the anomaly detection community, semi-supervised learning often refers to a setting where a data set only containing nominal data is available. Learning with positive and unlabeled examples (LPUE, or PU learning) is the extension when having nominal data as well as a contaminated data set.

Finally, we have a fully unsupervised scenario, where a contaminated data set is available without labels. Unsupervised and semi-supervised learning settings are the most prevalent in the literature.

3.2 Categorization

Although successful anomaly detectors depend on application-specific peculiarities, existing approaches can be roughly categorized depending on the their main idea. In the following, we discuss a list of categories based upon the book of Aggarwal [2]. We would like to point out that such a list is arbitrary and depends very much on the viewpoint and background of the writer. E.g. locality preserving projections (LPP) [59] is a *linear method* that depends

on the *proximity* of data points and tries to find a lower-dimensional subspace of a *high-dimensional* space. Nevertheless, it is a convenient way of coarsely separating methods to understand their main approach on outlier detection.

Extreme Value Theory Extreme value theory (EVT) is concerned with the limit behavior (as the number of samples goes to infinity) of sample extremes, i.e. data points that have very high or very low values. This is much like the central limit theorem (CLT), which models the limit behavior of sample sums. Indeed, both theories have been developed roughly at the same time. Extreme value theory was pioneered by Leonard Tippett in the first half of the last century. Working with cotton, he noticed that the weakest fibres controls the strength of a thread. Together with R. A. Fisher, Tippet obtained three asymptotic limits describing the distributions of extremes which was later put down in a book by Emil Julius Gumbel. Originally developed as an univariate theory, EVT was later extended to multivariate settings. However, it is developed to find outliers at the borders of the data which might not be helpful as a direct method but as a post-processing step for anomaly scores. Samples for learning the appropriate model (generalized extreme value distribution or the generalized Pareto distribution) are either selected based on the block-maxima (BM) approach or the more recently proposed peaks over threshold (POT).

Probabilistic and Statistical Models Statistical models for anomaly detection comprises tail inequalities and tail confidence tests. For Bayesian probabilistic models, parameters (and hyper-parameter) are modeled by probability distributions and the result itself is, again, a probability distribution called the posterior distribution. This is (often) in contrast to frequentist approaches, where we are usually interested in point estimates through, e.g. maximum likelihood (ML) or maximum a posteriori (MAP) hence, the best possible model. There are two key challenges to overcome: (a) modeling the dependencies between random variables and choosing the right probability distributions for the various parameters, and (b) inference, i.e. deriving the posterior distribution given observations. On the positive side, modeling data dependencies is a natural thing and since the result is a probability distribution, it can be neatly used to separate high-density regions from potential outlier regions. The downside, however, is that probability distributions must be chosen carefully to fully reflect the reality. As simple as it sounds, this is often not the case. Due to its complexity for deriving the posterior function, simpler or matching (i.e. conjugate) distributions are often used.

Linear Models Modeling the nominal class as a linear model and finding strong deviations from this model is an appealing approach since many prominent and powerful methods used in machine learning are based upon linear models. This list includes regression methods such as (ordinary/regularized) least squares regression, least absolute shrinkage and selection operator (LASSO) [60], and Gaussian processes as well as binary classifier such as logistic regression and support vector machines. The list continuous with, e.g. principle component analysis (PCA), independent component analysis (ICA), and, what will be very important to this thesis, one-class support vector machines (OC-SVMs) [7, 61]. Interestingly though, support vector data description (SVDD), which models a hypersphere around the nominal class, would technically not belong to this category as it is based on a quadratic model.

Proximity-based Models Proximity-based approaches split into three groups: (a) distance-based methods, (b) density-based methods, and (c) cluster-based methods. Distance-based approach will measure similarity based upon the k-nearest neighbor distances. The rational behind is that anomalies are data points with much larger distances to its nearest neighbors

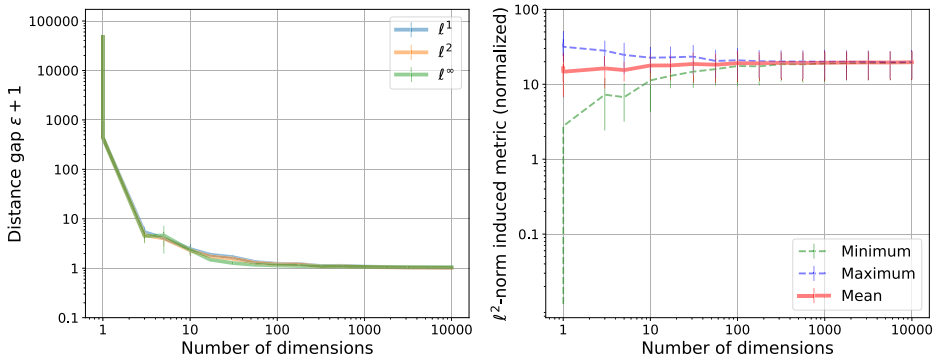

Figure 3.4 – Data was realized from isotropic Gaussian distributions (various cluster with random means and variance) with increasing dimensionality and uniformly distributed query points. The distance gap $\epsilon+1$ is reported on three distinct ℓ^p-norm induced metrics ($p \in \{1, 2, \infty\}$, left figure) as well as the minimum, maximum, and mean euclidean distance for increasing number of dimensions (right figure). Both figures show that with increasing number of dimensions minimum and maximum distances concentrate quickly.

than nominal data. Computationally, calculating k-nearest neighbors requires the computation of a pairwise distance matrix. This can be prohibitive for a large number of data points. Luckily though, new techniques such as locality sensitive hashing reduce the effort significantly. Approaches in (b) will utilize a number of data instances within the proximity of a data point to estimate its local density [3]. As usual, low density instances will be reported as more anomalous as high density points. One of the most prominent representative of this group is local outlier factor (LOF) [62] which was heavily used as a vehicle for other methods, e.g. local outlier probabilities (LoOP) [63]. The goal of methods in (c) is to partition the given data into subsets, where instances within a subset are *similar*. Here, approaches can return hard assignments such the famous k-means algorithm as introduced by MacQueen in 1967 [64] or hierarchical clustering approaches such as single-linkage clustering, or, soft assignments, e.g. Gaussian mixture models. Anomalies can then be identified as data points with large deviation from the nominal data within clusters.

High-dimensional Models High-dimensional models must be seen in historical context. In the past, researchers have noticed that *traditional* models for anomaly detection, namely nearest-neighbor-based models, degrade in performance as the number of dimensions or features increases. This is blamed on the infamous *curse of dimensionality*. In 1999, Beyer, Goldstein, Ramakrishnan, and Shaft condensed the reason into a theorem stating that if certain conditions are met, then, ultimately, the minimum distance and the maximum distance within a data set will concentrate (cf. Figure 3.4).

In detail, Theorem 1 in [65] states that as the dimensionality of data sets increases, the distance between some query point and its nearest neighbor D_{min} and its furthest neighbor D_{max} will converge to the same value,

$$\lim_{dims \to \infty} P\left(\frac{D_{max}}{D_{min}} - 1 \leq \epsilon\right) = 1$$

for any $\epsilon > 0$.

Regularization-based methods circumvent those problems by effectively restricting the model class. However, it is still an area of active research [66, 67] today and there is still a need for models that reduce the dimensionality of the presented data set for, e.g. visualization and summarization. Reducing the dimensionality can also increase the detection performance,

if misleading information, i.e. noise, is significantly reduced. A large class of methods in this category will extract a subspace from the original space. One of the most prominent approaches is kernel principal component analysis (kernel PCA) [33, 35].

Information Theoretic Models Information theoretic models are based on the minimum description length principle (MDL) which was developed by Rissanen, 1978 [68]. A recent introduction can be found in the book of Peter Grünwald [69] which is based on a previous tutorial [70]. The fundamental idea behind is, to see data compression as a way of learning functions. Moreover, in the context of this work, as a way to identify outliers. Given a set of hypotheses $H \in \mathcal{H}$ and a data set $X \in \mathcal{X}$, use the hypotheses that compressed X the most. Finding outliers then translates to finding patterns or data points that can not be compressed.

3.3 One-class Classification

The term *one-class classification* first appeared in the works published by Moya and Hush [1]. The following quote is taken from their original research paper [5] and describes the essence of an one-class classifier:

> *We call a classifier that can recognize new examples of target patterns and distinguish those from non-target patterns a one-class classifier.*

Note that this definition is quite distinct from the binary (or multi-class) classification setting, where the classifier is only required to distinguish between the target class and one specific form of non-target class (the opposing class) and not all possible non-target classes.

Translated to the anomaly detection setting, an one-class classifier distinguishes between the nominal data (=target pattern) and any sort of anomaly (=non-target pattern). It is this behavior that makes an one-class classifier especially suited for anomaly detection. One-class classifier are usually trained either on nominal data only (semi-supervised) or on an contaminated set (unsupervised) consisting of nominal data and some anomalies.

Interestingly, long before the work of Moya and Hush, T.C. Minter (1975) published his work on *single-class classification*. In his work [73],

> *(...) a Bayes classifier will be presented which classifies samples into the "class of interest" or the "other" classes but requires only labeled training samples for the "class of interest" to design the classifier. Thus, this classifier minimizes the need for ground truth. For these reasons, the classifier will be referred to as a single-class classifier.*

He, therefore, presented a first version of a semi-supervised one-class classifier [2].

Given the definition of the Hawkins' outlier, anomalies will be quite distinct from the nominal class. Hence, we presume that anomalies will occur in the tails of the nominal class probability density, i.e. in the low probability regions. Successful training of an one-class classifier, therefore, comprises of learning a tight description around the high probability regions of the presented data set. Hence, we would like to learn the density level set containing *most* of the data instances. To do so, there are two possible options. First, one can estimate the distribution and then cut-off at the desired density level set. Second, one attempts to estimate a binary classifier that can tag whether or not a given data instance belong to the desired density level set or not. Approaches building on the first principle are called *plug-in* estimators while approaches of the second principle are called *direct* estimators. One-class classifiers as discussed here, will be based on the second principle.

[1] *before* 1996 as claimed by Wikipedia, c.f. [4, 71, 72]

[2] The "other" classes mustn't contain the "class of interest".

Lets start with some theory behind. In 1992, Einmahl and Mason [74] presented a generalization of the quantile function based on the estimation of minimum volume sets (MVS) that has the following form:

$$U(\alpha) = \inf\{\lambda(C) : P(C) \geq \alpha,\ C \in \mathcal{C}\}$$

with P being the distribution, $\mathcal{C}$ measurable subsets of the input space, λ a measure (real-valued function going from $\mathcal{C}$ to $\mathbb{R}$). Parameterized by $0 < \alpha < 1$, MVS are the smallest volumes containing a probability mass of α, i.e. if $\alpha = 1$ then the corresponding MVS contains the support of the density (all non-zero probability elements). In the empirical case ($P(C) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_C(\mathbf{x}_i)$ and Lebesque measure), α controls the fraction of data points lying outside of the MVS. So, we get the smallest volume (a *tight description*) of the data with a $1 - \alpha$ fraction of data points that will not be included. Moreover, Wolfgang Polonik showed that under some assumptions, minimum volume sets are indeed density level sets [75, 76].

More formally, we are given a set of input instances $x_1, \ldots, x_n \in \mathcal{X}$, which are commonly assumed to be realized from independent and identically distributed (i.i.d.) random variables $X_1, \ldots, X_n \sim P$, where P is a potentially unknown measure of probability. The aim is to find a set containing the most typical instances under the measure P, and instances lying outside of the set are declared as anomalies. The task of anomaly detection can be formally phrased within the framework of density level set estimation [75, 77, 78] as follows. Denoting by X another i.i.d. copy according to P, the theoretically optimal nominal set is $L_\nu := \{x \in \mathcal{X} : p(x) \geq b_\nu\}$ for $\nu \in]0, 1[$ and b_ν such that $P(X \notin L_\nu) = \nu$, which is called the ν *density level set* and can be interpreted as follows: L_ν contains the most likely inputs under the density p, while rare or untypical data ("anomalies") are modeled to lie outside of L_ν. The parameter ν indicates the fraction of outliers in the model.

The aim is to compute, based on the data $x_1, \ldots, x_n \in \mathcal{X}$, a good approximation of L_ν, that is, to determine a function $f : \mathcal{X} \to \mathbb{R}$ giving rise to an estimated density level set $\hat{L}_\nu := \{x \in \mathcal{X} : f(x) \geq 0\}$. It is desirable that $\hat{L}_\nu$ closely approximates the true density level set L_ν, i.e., $\hat{L}_\nu$ converges to L_ν in probability, that is,

$$P(\hat{L}_\nu \backslash L_\nu \cup L_\nu \backslash \hat{L}_\nu) \to 0 \text{ for } n \to \infty.$$

This implies that $\hat{L}_\nu$ has asymptotically probability mass ν, that is, $P(X \notin \hat{L}_\nu) \to \nu$ for $n \to \infty$.

In the following, we focus on the two most prominent kernel-based [33] one-class classifiers. Other approaches include, e.g. Bayesian data description [79], Gaussian processes [80], neural networks [5], and random forest [81]. Further, many existing approach can be, with little changes, used as one-class classifier, i.e. kernel density estimation, Gaussian mixture models, k-means.

3.3.1 One-class Support Vector Machine

A, if not *the*, classic approach to kernel-based one-class classification is the one-class support vector machine (OC-SVM) [7, 8, 82]. Even today, the OC-SVM is among the most prominent and successful anomaly detectors. The OC-SVM is based on linear models $f_{\mathbf{w},\rho}(\mathbf{x}) = \langle \mathbf{w}, \phi(\mathbf{x}) \rangle - \rho$, where the data is mapped into a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ via a feature map $\phi : \mathcal{X} \to \mathcal{H}$. It subsequently separates a fraction of $1 - \nu$ many inputs

from the origin with maximum margin:

$$\begin{aligned} \max_{\mathbf{w},\rho,\xi\geq 0} \quad & \frac{1}{2}\|\mathbf{w}\|^2 - \rho + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & \xi_i \geq -f_{\mathbf{w},\rho}(\mathbf{x}_i) \quad \forall\, i = 1,\ldots,n. \end{aligned} \tag{Primal OC-SVM}$$

The corresponding dual OP has the following form:

$$\begin{aligned} \max_{0\leq\alpha\leq\frac{1}{n\nu}} \quad & -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j \mathsf{k}(\mathbf{x}_i,\mathbf{x}_j) \\ \text{subject to} \quad & \sum_{i=1}^{n}\alpha_i = 1 \end{aligned} \tag{Dual OC-SVM}$$

with expansions $\mathbf{w} = \sum_{i=1}^{n}\alpha_i\phi(\mathbf{x}_i)$. Further properties will be discussed in this thesis where appropriate.

Due to its success, there is a vast literature building atop of OC-SVMs. Rätsch et al. [83] showed that boosting-like algorithm can be constructed solving a ℓ_1-norm regularized one-class SVM. Lee and Scott [84] gave an answer for the problem of calculating the one-class SVM solution path when ν varies between 0 and 1. The relation to density estimation was shown in [85]. Estimating minimum volume sets has been investigated by [86]. Vert and Vert [87] actually showed that the one-class SVM with RBF kernel is a *consistent* density level set estimator. The construction of hierarchical level sets has been investigated by [88, 89]. Recent works comprises extensions towards group anomaly detection [90] and, of course, deep learning [91].

3.3.2 Support Vector Data Description

The goal is to find a model $f\colon \mathcal{X} \to \mathbb{R}$ and a density level set $L := \{\mathbf{x}\colon f(\mathbf{x}) \leq 0\}$ containing most of the regular data points, while for anomalies and outliers $\mathbf{x} \notin L$ holds. In case of the support vector data description (SVDD) method, $f_{\mathbf{c},R}(\mathbf{x}) = \|\mathbf{c} - \phi(\mathbf{x})\|^2 - R^2$ and parameter estimation corresponds to solving a quadratically constrained quadratic program (QCQP) as originally proposed by Tax [92]:

$$\begin{aligned} \min_{R,\mathbf{c},\xi\geq 0} \quad & R^2 + C\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & \xi_i \geq f_{\mathbf{c},R}(\mathbf{x}_i) \quad \forall\, i = 1,\ldots,n \end{aligned} \tag{Primal SVDD}$$
$$\tag{3.1}$$

That allows for the following simple geometric interpretation: a ball with minimum radius R is computed that comprises most the regular data points, while all points lying outside of the normality radius are declared being anomalous. This is a more direct link to the minimum volume estimation discussion above. Here, the sets $C \in \mathcal{C}$ are hyper-spheres in the reproducing kernel Hilbert space (RKHS). The corresponding dual problem of the above OP is:

$$\begin{aligned} \max_{0\leq\alpha\leq C} \quad & \sum_{i=1}^{n}\alpha_i\mathsf{k}(\mathbf{x}_i,\mathbf{x}_i) - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\mathsf{k}(\mathbf{x}_i,\mathbf{x}_j) \\ \text{subject to} \quad & \sum_{i=1}^{n}\alpha_i = 1 \end{aligned} \tag{Dual SVDD}$$

with expansions $\mathbf{c} = \sum_{i=1}^{n} \alpha_i \phi(\mathbf{x}_i)$.

As well as with the OC-SVM, more details and properties will be discussed in this thesis if necessary, i.e. Chapter 6.3 presents a in-depth discussion of the properties of the SVDD and some of the OC-SVM. It is noteworthy to mention that a strong connection between both methods exists which is the reason why in the literature the term OC-SVM is sometimes used for the SVDD.

Due to its simple interpretation, SVDDs became very popular in the literature and lots of application but also theoretical contributions have been made. The solution path of the SVDD was analyzed by [93]. A Bayesian approach to data description was proposed by [79].

3.4 Model Selection and Evaluation

One of the most crucial parts, not only for anomaly detection but basically for any learning machine in any setting, is, the evaluation and model selection part. Selecting the best hyper-parameter and measuring how good a trained model generalizes is the essence before putting it in production. Unfortunately, due to the nature of the anomalies, only few, if any, will be known. Hence, a serious problem in anomaly detection in general is the absence of sufficient information about anomalous classes.

In unsupervised anomaly detection, models are often combined instead of *selected* (model averaging vs. model selection). Even though that still leaves the problem of estimating the generalization error untouched. However, only in rare cases will be absolutely no information available about existing anomalies and evaluation can be attempted using the few available labeled examples.

Even if enough labeled samples are available, the class balances will be extremely skewed and some measures, e.g. classification error, will not be suited to reflect the state of generalization error appropriately. To circumvent those problems, error measures that are transient to class imbalances are used. Namely, the area under the ROC curve (AUC or AUROC) and area under the precision recall curve (AUPR) [94].

Few works consider estimation of performance measures with missing information. In case of PU learning, Hajizadeh, Li, Dollevoet, and Tax [95] introduced a measure, PULP, that works without explicitly given negative labels. PULP is based on the calculation of the probability of true positives for some random positive predictions. Another very interesting approach is given in [96], where the author shows that based on mass volume curves and excess mass curves, evaluation of unsupervised anomaly detectors can be done without the help of test samples.

Tax and Müller [97] tackle the problem of model selection of one-class classifier with only considering the nominal test samples. Thomas, Clémençon, Feuillard, and Gramfort [98] generalize this solution for model selection and base their decision upon estimation of mass volume curves.

3.5 Summary and Discussion

The problem of anomaly detection arises in many application scenarios and is often a vital part of the data processing pipeline. Moreover, anomalies itself pose a valuable information source as they often translate to actionable items. Techniques for anomaly detection have been investigated for more than half a century now with new challenges arising today, i.e. due to the shear amount of data (Big Data) or new types of complex data (Social Networks). Especially when data has dependency structure, it often poses a complex task as anomalies might not occur as isolated points (*point* anomalies) instead they only appear in groups of data points (*collective* anomalies) or within some context (*contextual* anomalies).

A specific approach to anomaly detection is one-class classification where a classifier is trained on a single class (the nominal data) alone. A promising family of one-class classifiers is the one-class support vector machine (OC-SVM) and the support vector data description (SVDD).

In this thesis, we propose extensions upon the framework of OC-SVM and SVDD to tackle specific problems with with a special focus on leveraging dependency structure within features and samples.

We are hereby relying on specialized formulations and slight deviations from the standard techniques. In order to avoid notational clutter, we chose to introduce the type of classifier and the corresponding formulation (e.g. constraint vs. unconstraint, primal vs. dual, generalized loss functions and/or regularizer) used within each part separately.

II Point Anomalies

Chapter 4

Sparsity-inducing Regularization

One of the most famous and oldest disputes in statistics and machine learning is about the origins of the ordinary least squares method. While Adrien Marie Legendre published his work in 1805 [99] and Carl Friedrich Gauss 4 years later in 1809 [100], the latter claimed that he developed his method in 1795, almost 10 years before Legendre published his work. Recent work [101] suggests that Gauss was indeed right, albeit, a definitive answer to this century old question will probably never given. However, the method and its countless variants still remain the working horse in the data science community.

One particular successful extension, ridge regression [102], added a squared regularizer to the main objective. The results of this simple extension was a more stable optimization as well as more accurate predictions. However, these types of regularizations also result in a dense representation, meaning that each and every variable is used for prediction even if the correlation to the effect is negligible.

Albeit dense representations generally have slightly higher prediction performance on real datasets, a parsimonious representation would have several advantages. Besides space and computational advantages due to the smaller number of variables involved in the process, those representations might unravel the driving forces, the main causes behind the examined problem and would make the model and the data much more accessible to interpretation. The underlying assumption of *sparse models* is that target values can be accurately described using only a small subset of input variables. Hence, the model is supposed to learn a mapping with zero influence weights for variables that do not improve prediction performance. Robert Tibsherani [60] formed ridge regression into a *sparse model* by simply replacing the squared regularizer with an ℓ_1 regularizer. The resulting model, the lasso, hence became very successful with applications in, e.g. computation biology [103].

Structured sparse models [48, 104] are a natural extension of the *independent* regularization of single variables by considering *dependencies* among those. The kind of structure is hereby usually defined *a priori*. A common setting includes group sparsity where, instead of single variables, disjunct groups of variables are weighted against each other. Of course,

there are numerous extensions towards overlapping groups, hierarchical or graph dependencies among groups of features.

However, group sparsity has a nice interpretation in terms of multiple kernel learning [105–108]. As the name suggests, the goal is to learn a convex combination of multiple kernels to either select the most promising group of features (sparse setting) or to leverage all combined information of heterogeneous data representations (non-sparse setting). In this chapter, we aim at deriving a variant of the one-class SVM [82] that can smoothly transition into a sparse model using ℓ_p-norm regularization as well as a ℓ_p-norm regularized multiple kernel learning variant of the semi-supervised one-class SVM as given in [12] that also can be applied to sparse as well as non-sparse MKL settings. Hereby, the ℓ_p-norm ensures that both, high prediction accuracy and interpretability can be achieved.

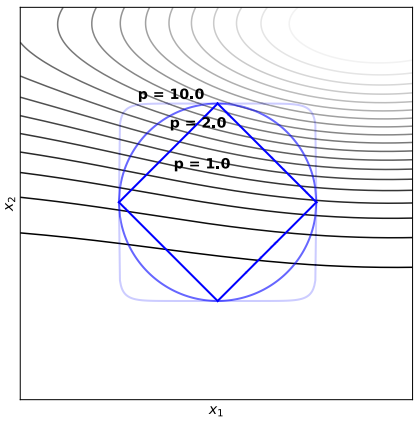

Figure 4.1 – Grey lines represent density level sets of the objective function of interest. Blue lines indicate the ℓ_p-norms for $p = \{1, 2, 10\}$. Maximum values are achieved in the corners for $p = 1$ (sparse solution with $x_1 = 0$ and $x_2 = 1$), at about 80° for $p = 2$ (non-sparse but $x_1 << x_2$), and at almost exactly 45° for $p = 10$ ($x_1 \approx x_2$).

In the following sections, after discussing the problem setting in more detail (cf. 4.1), we introduce a variant of the one-class SVM with ℓ_p-norm regularization that also admits sparse reconstruction (cf. Section 4.2) and a particular variant that admits sparse reconstruction of (non-overlapping) groups of features (Section 4.3). Before we conclude the chapter in Section 4.5, we will employ the developed methods to applications in Section 4.4.

4.1 Preliminaries

We build our approaches on the paradigms of support vector learning [33, 37] and one-class classification; that is, we are given n data points $\mathbf{x}_1, \ldots, \mathbf{x}_n$, where $\mathbf{x}_i$ lies in some input space $\mathbb{R}^d$, and the goal is to find a model $f : \mathbb{R}^d \to \mathbb{R}$ and a density level-set $D_\rho = \{\mathbf{x} : f(\mathbf{x}) \geq \rho\}$ encompassing the normal data, i.e., $\mathbf{x} \in D_\rho$, while for outliers $\mathbf{x}' \notin D_\rho$ holds. In this chapter, we consider linear models of the form

$$f(\mathbf{x}) = \langle \mathbf{w}, \phi(\mathbf{x}) \rangle \tag{4.1}$$

for some feature function $\phi : \mathbb{R}^d \to \mathcal{F}$ mapping the data into some feature space.

Our aim is a parsimonious representation of the model for single features (cf. Section 4.2) as well as disjunct groups of features (cf. Section 4.3 without compromising on accuracy. That is, we introduce a tunable parameter that smoothly controls the sparseness of the solution. We consider the Minkowski ℓ_p-norm, where $\|\mathbf{w}\|_p = \left(\sum_{i=1}^{d} |w_i|^p \right)^{1/p}$ with $p \geq 1$. An illustration of the impact of various norms on the optimization outcome is given in Figure 4.1.

4.2 Sparsity-inducing One-class SVM

In this section, we will introduce an ℓ_p-norm regularization for the (primal) one-class SVM as given in Equation Primal OC-SVM. This will allow us to smoothly transition from the dense solutions given by the ℓ_2-norm regularized one-class SVM towards sparse solutions with ℓ_1 regularizer. A variant of the ℓ_1-norm regularized one-class SVM was derived by Rätsch, Mika,

Schölkopf, and Müller [109, 110] in the context of boosting relying on Ivanov regularization and barrier methods for optimization.

Another way of smooth transitions between sparse and dense solutions is the elastic net approach that has been introduced by [111]. Here, two regularizer, one ℓ_1 and the other ℓ_2, are weighted against each other. This solution is slightly more complex and does need one more hyper-parameter to adjust. Therefore, we resort to the ℓ_p-norm solution. In detail, our contributions are the following:

- we propose a variant of the primal one-class SVM with ℓ_p-norm regularization;
- we derive a corresponding unconstrained optimization problem and solver,
- for the special case of $p = 1$ are able to derived a re-formulation that allows very efficient optimization;
- a semi-supervised learning extension for $p = 1$ that leverages negative labels is proposed.

In section, we never leave the primal space for optimization and therefore, we assume that transformations have been applied to the data so we can set $\phi \mapsto id_{\mathbb{R}^d}$ which will allow us to discard the function and avoid notational clutter.

Lets begin with a slightly generalized version of the primal one-class SVM problem in Equation (Primal OC-SVM):

$$\begin{aligned} \min_{\mathbf{w},\rho,\xi} \quad & \Omega(\mathbf{w}) + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \rho \\ \text{s.t.} \quad & \langle \mathbf{w}, \mathbf{x}_i \rangle \geq \rho - \xi_i, \quad \xi_i \geq 0, \quad \forall i \in \{1, \ldots, n\} \end{aligned} \tag{4.2}$$

where $\Omega(\mathbf{w})$ is a smooth regularizer and $\nu \in]0, 1]$ a hyper-parameter controlling the 'size' of the level set (the lower ν the larger the level set). Once the optimal parameters $\mathbf{w}^*$ and ρ^* are found, these are plugged into (4.1), and new instances $\mathbf{x}$ are classified according to $\text{sign}(f(\mathbf{x}) - \rho^*)$.

The learning machine (4.2) has been intensively studied for the choice of the regularizer $\Omega(\mathbf{w}) := \frac{1}{2}\|\mathbf{w}\|_2^2$, which leads to *dense* optimal weight vectors $\mathbf{w}^*$, i.e., the entries of $\mathbf{w}^*$ are strictly different from zero (except in pathological cases) and thus hinder feature selection and interpretability. In contrast, we build the methodology used in this section on more general regularizers of the form

$$\Omega(\mathbf{w}) := \|\mathbf{w}\|_p \ ,$$

where $\|\mathbf{w}\|_p = \left(\sum_{i=1}^{d} |w_i|^p\right)^{1/p}$ denotes the Minkowski ℓ_p-norm. Solving optimization problem (4.2) can be tedious due to the various constraints and non-smooth terms. However, we can easily re-write the above optimization problem by substituting ξ_i. Note that $\xi_i \geq \rho - \langle \mathbf{w}, \mathbf{x}_i \rangle$ and $\xi_i \geq 0$ which leads to $\xi_i \geq \max(0, \rho - \langle \mathbf{w}, \mathbf{x}_i \rangle)$. Minimization ensures equality and hence, we arrive at

$$\min_{\mathbf{w},\rho} \quad f_p(\mathbf{w}, \rho) = \|\mathbf{w}\|_p - \rho + \frac{1}{\nu n}\sum_{i=1}^{n}\max(0, \rho - \langle \mathbf{w}, \mathbf{x}_i \rangle) \ . \tag{4.3}$$

Algorithm 2 Subgradient descent solver for ℓ_p-norm one-class SVM (Eq. (4.2))

Require: $\{\mathbf{x}\}_{i=1}^n$, ν, p, and step size/rule α
  Initialize $\mathbf{w}^0$ and ρ^0
  Set $f_{best} = +\inf$, $\mathbf{w}_{best} = \mathbf{0}$, and $k = 0$
  while no convergence and $k \leq$ max_iter **do**
    $f^k = f_p(\mathbf{w}^k, \rho^k)$ (cf. Eq.4.3)
    if $f^k < f_{best}$ **then**
      $f_{best} = f^k$
      $\mathbf{w}_{best} = \mathbf{w}^k$
    end if
    $\mathbf{w}^{k+1} = \mathbf{w}^k - \alpha \frac{\partial f_p(\mathbf{w}^k, \rho^k)}{\partial \mathbf{w}^k}$ (cf. Eq.4.4)
    $\rho^{k+1} = \rho^k - \alpha \frac{\partial f_p(\mathbf{w}^k, \rho^k)}{\partial \rho^k}$ (cf. Eq.4.4)
    $k = k + 1$
  end while
  return $\mathbf{w}_{best}$

The above optimization problem can be readily solved using standard techniques such as sub-gradient descent where only sub-gradients w.r.t. $\mathbf{w}$ and ρ need to be assessed by

$$\frac{\partial f_p(\mathbf{w}, \rho)}{\partial \mathbf{w}} = \frac{\mathbf{w} \odot |\mathbf{w}|^{p-2}}{\|\mathbf{w}\|_p^{p-1}} + \frac{1}{\nu n} \sum_{i=1}^{n} \begin{cases} -\mathbf{x}_i & \text{for } 0 < \rho - \langle \mathbf{w}, \mathbf{x}_i \rangle \\ 0 & \text{else} \end{cases}, \tag{4.4}$$

$$\frac{\partial f_p(\mathbf{w}, \rho)}{\partial \rho} = -1 + \frac{1}{\nu n} \sum_{i=1}^{n} \begin{cases} 1 & \text{for } 0 < \rho - \langle \mathbf{w}, \mathbf{x}_i \rangle \\ 0 & \text{else} \end{cases}. \tag{4.5}$$

Here, $\odot$ denotes the Hadamard product. The resulting sub-gradient descent solver is given in Alg. 2. We give an example of the impact of $1 \leq p \leq 4$ for a very simple setup consisting of 20-dimensional correlated Gaussian variables in Figure 4.2. To circumvent numerical issues, we report the sum of absolute values normalized by its respective maximum component $\sum_i |\mathbf{w}_i| / \max_j |\mathbf{w}_j|$. As we expected, parsimony in the solution vector correlates with p.

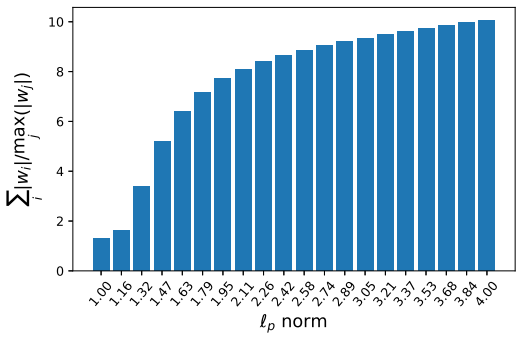

Figure 4.2 – Impact of p on the sparseness of the solution. To avoid numerical issues, we report the sum of absolute values normalized by the maximum component $\sum_i |\mathbf{w}_i| / \max_j |\mathbf{w}_j|$.

Very sparse solutions Now, we focus more on the limiting case $p = 1$, which is likely to lead to very sparse solutions: suppose we minimize an objective function $g(\mathbf{w})$ subject to $\|\mathbf{w}\|_1 \leq 1$; then, the optimal solution is attained when the level sets of the objective function 'hit' the norm constraint. If the objective function is convex, the point of intersection is usually at one of the corners of the constraint and thus has sparse coordinates (cf. Figure 4.1). In linear methods, each dimension in the solution often corresponds to a measurable cause. The benefit of increasing the sparseness in the solution vector lies in the fact that the solution now becomes interpretable. I.e. when no ground truth is available, interpretability is mandatory. An elegant way to solve (4.2) for $\Omega(\mathbf{w}) = \|\mathbf{w}\|_1$ is to set $\mathbf{w} = \mathbf{w}^+ - \mathbf{w}^-$, substituting $\|\mathbf{w}\|_1 = \sum_d \mathbf{w}_d^+ + \mathbf{w}_d^-$, and to optimize over $\mathbf{w}^+, \mathbf{w}^- \geq \mathbf{0}$ instead of $\mathbf{w}$.

To enhance numerical stability of sparse one-class learning, we propose to consider the following sparsity-inducing one-class learning formulation:

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & \|\mathbf{w}\|_1 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & \langle \mathbf{w}, \mathbf{x}_i\rangle \geq 1-\xi_i, \quad \xi_i \geq 0, \quad \forall i \in \{1,\ldots,n\} \end{aligned} \tag{4.6}$$

(which is reminiscent of the very well known 2-class C-SVM given by Cortes and Vapnik [37] or sparse Fisher by Mika et al [112]). The following theorem shows that (4.2) is an exact re-formulation of (4.6).

Theorem 5. *Let* $\Omega(\mathbf{w}) = \|\mathbf{w}\|_1$ *and denote the optimal solution of* (4.2) *and* (4.6) *by* $(\mathbf{w}^*_\nu, \rho^*_\nu, \xi^*_\nu)$ *with* $\rho^*_\nu > 0$ *and* $(\widetilde{\mathbf{w}}^*_C, \widetilde{\xi}^*_C)$*, respectively. Then, for any* $\nu \in]0,1]$*, setting* $C := \frac{1}{\nu n}$*, it holds*

$$\mathbf{w}^*_\nu = \rho^*_\nu \widetilde{\mathbf{w}}_C\,,$$

i.e., the weight vectors output by (4.2) *and* (4.6) *are, besides a scaling factor, equivalent.*

Proof. Let $(\mathbf{w}^*, \rho^*, \xi^*)$ be optimal in (4.2). It follows that $(\mathbf{w}^*, \rho^*)$ is optimal in the corresponding unconstrained formulation:

$$(\mathbf{w}^*, \rho^*) \;=\; \operatorname*{argmin}_{\mathbf{w},\rho} \;\; \|\mathbf{w}\|_1 \;+\; \frac{1}{\nu n}\sum_{i=1}^{n}\max(0, \rho - \langle \mathbf{w}, \mathbf{x}_i\rangle) \;-\; \rho\,.$$

Note that thus $\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w}} \|\mathbf{w}\|_1 + \frac{1}{\nu n}\sum_{i=1}^{n}\max(0, \rho^* - \mathbf{w}^\top \mathbf{x}_i)$. Now denote $\widetilde{\mathbf{w}}^* := \operatorname{argmin}_{\widetilde{\mathbf{w}}} \; \rho^* \|\widetilde{\mathbf{w}}\|_1 + \frac{1}{\nu n}\sum_{i=1}^{n}\max(0, \rho^* - \langle \rho^*\widetilde{\mathbf{w}}, \mathbf{x}_i\rangle)$. By a variable substitution $\mathbf{w} = \rho^*\widetilde{\mathbf{w}}$, we observe that $\mathbf{w}^* = \rho^*\widetilde{\mathbf{w}}^*$ and hence $\mathbf{w}^*/\rho^*$ is optimal in $\min_{\widetilde{\mathbf{w}}} \; \|\widetilde{\mathbf{w}}\|_1 + \frac{1}{\nu n}\sum_{i=1}^{n}\max(0, 1 - \langle \widetilde{\mathbf{w}}, \mathbf{x}_i\rangle)$ (because ρ^* is positive), which, setting $C := \frac{1}{\nu n}$ is the unconstrained version of (4.6) (and thus equivalent). Thus $\mathbf{w}^*/\rho^*$ is optimal in (4.6), which was to show. □

Semi-supervised Learning In exploratory data analysis with large amounts of data, unsupervised methods are often applied in an iterative manner to reveal properties of interest. To incorporate accumulated knowledge, we would like to include labeled information into the optimization problem. For dense models, this has been done in Görnitz, Kloft, Rieck, and Brefeld [12]. However, we need to retain the sparseness of the solution and don't want to increase the number of hyper-parameters that need to be adjusted on-the-fly (of which there are four in [12]). We can achieve this by only adding negative labels hence, having unlabeled and negatively labeled data available. Moreover, when compared to [12], there will be no margin as well as uniform influence for each data point.

We wish to include negatively labeled instances $\mathbf{x}_{n+1}, \ldots, \mathbf{x}_m$ (i.e., instances of which we already know that they are outliers) into the learning machine (4.6). A simple and effective way to do so is to constrain the negatively labeled instances to lie outside of the density level set: $\langle \mathbf{w}, \mathbf{x}_i\rangle \leq 1 + \xi_i, \quad \xi_i \geq 0, \quad \forall i \in \{n+1, \ldots, m\}$. The resulting linear program

$$\begin{aligned} \min_{\mathbf{w}^+,\mathbf{w}^-,\xi} \quad & \sum_{j=1}^{d}(\mathbf{w}^+_j + \mathbf{w}^-_j) + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & \langle \mathbf{w}^+ - \mathbf{w}^-, \mathbf{x}_i\rangle \geq 1-\xi_i, \quad \xi_i \geq 0, \quad \forall i \in \{1,\ldots,n\} \\ & \langle \mathbf{w}^+ - \mathbf{w}^-, \mathbf{x}_i\rangle \leq 1+\xi_i, \quad \xi_i \geq 0, \quad \forall i \in \{n+1,\ldots,m\} \\ & \mathbf{w}^+ \geq 0, \quad \mathbf{w}^- \geq 0 \end{aligned} \tag{4.7}$$

can be efficiently solved using off-the-shelf solver such as MOSEK[1].

A note on $p < 1$ Very sparse solution can be obtained by optimizing over the ℓ_0-norm which gives the number of non-zero elements and other $p < 1$-norms. However, these are no proper norms anymore and the resulting optimization algorithm would be non-convex or even a combinatorial problem (for ℓ_0). Hence, applying Algorithm 2 does not guarantee convergence to a meaningful optimum anymore. Using ℓ_1 instead can be viewed as a convex surrogate for the actual sparse solution.

4.3 Inducing Group-sparsity

Group sparsity, which was studied in the context of lasso first [113], is the problem of identifying and selecting groups of features instead of single features. Unlike standard lasso, features within groups are regularized using the ℓ_2-norm while on group level a ℓ_1 regularizer is used. The result leads to a parsimonious selection of groups with dense features rather than a parsimonious representation of single features. A comprehensive survey on optimizing structured sparsity models using penalties can be found in [48].

Multiple kernel learning (MKL) is especially useful for heterogeneous data where a single similarity measure might not suffice to efficiently capture signals from data. Combining various kernels and weighting them accordingly to achieve improved performance is the goal of MKL. Hereby, each kernel represents a group of features and when optimized to yield sparse weightings [107, 114, 115], can be viewed as a group sparsity optimization problem. In fact, the relation between MKL and group sparsity is well-known [48, 116].

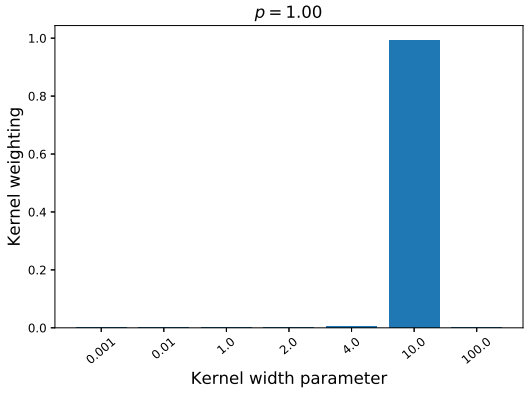

Figure 4.3 – Impact of p on the sparseness of the solution. Here, the weighting of seven RBF kernels with various widths have been learned setting $p = 1$.

Recently, non-sparse multiple kernel learning using ℓ_p-norm regularization has been shown to outperform sparse counterparts by some margin [105, 117, 118]. The additional hyper-parameter p can be adjusted to trade-off accuracy and sparsity of the solution. In the limiting case of $p = 1$, equivalence to the group sparsity problem is ensured.

There are various works that extend to multiple kernel learning idea to one-class classifiers. Among the most prominent are Kloft, Brefeld, Düssel, Gehl, and Laskov [119] who extended the support vector data description (SVDD) to regularize groups of features with an ℓ_1 loss in the context of network intrusion detection. The same extension had been applied to one-class SVMs in [107]. Our contribution, on the other hand, provides a trade-off parameter p that controls the sparseness of the solution and, additionally, is able to handle labeled examples.

Our contributions in this section are:

- we extend the (convex) semi-supervised anomaly detector (SSAD) [12] to handle multiple kernels and to automatically adjusting their weighting using ℓ_p-norm regularization;

[1] http://www.mosek.com/

- we propose a corresponding block coordinate descent solver in the spirit of [105] that alternates between solving the SSAD problem using the most recent kernel mixture and analytically updating the mixture coefficients.

We show exemplary the impact of the trade-off parameter p on the sparseness of the found solution in Figure 4.3 and Figure 4.4. Here, seven RBF kernels with various widths (x-axis) are generated from Gaussian data points. Figure 4.3 shows that setting $p = 1$ leads to sparse solutions and only a single kernel receives most of the weight. In Figure 4.4 however, the non-sparse solution using $p = 2$ still picks the very same kernel to receive most of the weight while distributing large portions of the total weight to other kernels as well.

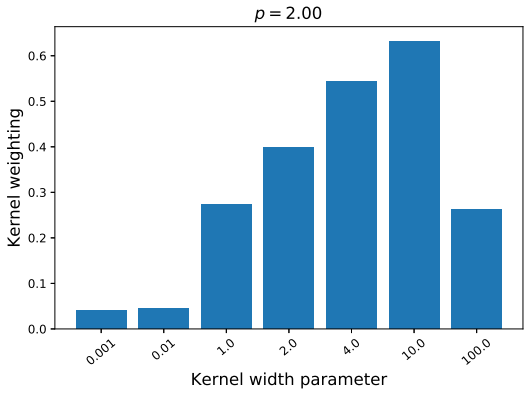

Figure 4.4 – Impact of p on the sparseness of the solution. Here, the weighting of seven RBF kernels with various widths have been learned setting $p = 2$.

In cases were few labeled examples are available, using only unlabeled data for the estimation of the parameter of the one-class SVM is usually leading to less accurate models as when fully exploiting labeled information [12]. Therefore in addition to n unlabeled examples $\mathbf{x}_1, \ldots, \mathbf{x}_n$, we include m labeled examples $(\mathbf{x}_{n+1}, y_{n+1}), \ldots, (\mathbf{x}_{n+m}, y_{n+m})$. Labels $y_i \in \{+1, -1\}$ are considered binary, that is in case $y_i = +1$, the entry $\mathbf{x}_i$ belongs to the nominal class. To combine sums and hence, improve readability, we introduce labels $y_i = +1 \; \forall \, i = 1, \ldots, n$ for all unlabeled examples and an indicator function $\mathbf{1}_c \equiv [c > n]$ to mask labeled examples; the function $\mathbf{1}_c$ simply returns 1 if $c > n$ and 0 otherwise.

A semi-supervised generalization of the one-class SVM model is the convex version of the semi-supervised anomaly detection framework (SSAD) [12] which we will use with an L_2-regularizer together with the hinge-loss. Let γ be the margin for the labeled examples and κ, η_u, and η_l trade-off parameters. For avoiding notational clutter, we introduce the example-wise regularization hyper-parameters

$$\eta_i = \begin{cases} \eta_u & \text{for } i = 1, \ldots, n \\ \eta_l & \text{else} \end{cases}$$

which allows us to shorten the optimization problem to

$$\begin{aligned} \min_{\mathbf{w}, \rho, \gamma \geq 0, \xi \geq 0} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 - \rho - \kappa\gamma + \sum_{i=1}^{n+m} \eta_i \xi_i \\ \text{s.t.} \quad & y_i \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle \geq y_i \rho + \mathbf{1}_i \gamma - \xi_i \quad \forall \, i = 1, \ldots, n+m \; . \end{aligned} \tag{4.8}$$

The solution $\mathbf{w}$ admits a dual representation and can be written as

$$\mathbf{w} = \sum_{i=1}^{n+m} \alpha_i y_i \phi(\mathbf{x}_i)$$

and hence, the decision function depends only on inner products of the input examples which paves the way for kernel functions $k_\phi(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$ (see [33] for an introduction to

kernels). It holds

$$f(\mathbf{x}) = \sum_{i=1}^{n+m} \alpha_i y_i \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle - \rho = \sum_{i=1}^{n+m} \alpha_i y_i k_\phi(\mathbf{x}_i, \mathbf{x}) - \rho.$$

We omit the subscript ϕ in the remainder to not clutter notation unnecessarily.

A kernel represents a similarity measure for single features or groups of features. If inputs are of very distinct nature, e.g. continuous and discrete values, a single similarity measure might not be sufficient. In such cases, we would like to incorporate multiple feature descriptions into the learning problem. Those would be represented by their respective kernel. To fully exploit the provided set of kernels, we aim to learn a weighted combination of T kernels with mixing coefficients $\mathbf{d} = (\beta_1, \ldots, \beta_T)$:

$$k_{\text{MKL}}(\mathbf{x}, \mathbf{x}') := \sum_{t=1}^{T} \beta_t k_t(\mathbf{x}, \mathbf{x}') = \sum_{t=1}^{T} \beta_t \langle \phi_t(\mathbf{x}), \phi_t(\mathbf{x}') \rangle = \sum_{t=1}^{T} \langle \sqrt{\beta_t}\phi_t(\mathbf{x}), \sqrt{\beta_t}\phi_t(\mathbf{x}') \rangle.$$

In general, properties of the mixing coefficients include (i) non-negativity, hence $\beta_t \geq 0$ and (ii) normalization $\|\mathbf{d}\|_p = 1$. Recent work [105] suggests to use the more general p-norm instead of a common 1-norm [107, 120, 121]. The latter usually leads to sparse mixing coefficients whereas p-norm with $1 \leq p \leq +\infty$ admits sparsity adjustments for the problem at hand and thus adds flexibility. Incorporating multiple feature representations in our model (4.8) leads to

$$f_{\text{MKL}}(\mathbf{x}) = \sum_{t=1}^{T} \langle \hat{\mathbf{w}}_t, \sqrt{\beta_t}\phi_t(\mathbf{x}) \rangle - \rho = \sum_{t=1}^{T} \sqrt{\beta_t} \langle \hat{\mathbf{w}}_t, \phi_t(\mathbf{x}) \rangle - \rho.$$

Due to technical reasons, i.e. to preserve convexity, we substitute the model parameters $\mathbf{w}_t = \sqrt{\beta_t}\hat{\mathbf{w}}_t$ and arrive at the revised primal MKL-SSAD optimisation problem:

$$\begin{aligned} \min_{\{\mathbf{w}_t\}_{t=1}^T, \rho, \gamma \geq 0, \xi \geq 0} \quad & \frac{1}{2} \sum_{t=1}^{T} \frac{1}{\beta_t} \|\mathbf{w}_t\|_2^2 - \rho - \kappa\gamma + \sum_{i=1}^{n+m} \eta_i \xi_i \\ \text{s.t.} \quad & y_i \sum_{t=1}^{T} \langle \mathbf{w}_t, \phi_t(\mathbf{x}_i) \rangle \geq y_i \rho + \mathbf{1}_i \gamma - \xi_i \quad i = 1, \ldots, n+m \\ & \|\mathbf{d}\|_p^2 \leq 1\,, \quad \mathbf{d} \geq 0. \end{aligned} \tag{4.9}$$

[105] prove the equivalence of Tikhonov and Ivanov regularisation which allows to move the regulariser on the mixing coefficients in the objective function. We will exploit this relation on various occasions in this section. Deriving the Lagrange dual problem, we arrive at the intermediate saddle point problem

$$\begin{aligned} \max_{\alpha} \min_{\{\mathbf{w}_t\}_{t=1}^T, \mathbf{d} \geq 0} \quad & \frac{\lambda}{2} \|\mathbf{d}\|_p^2 + \frac{1}{2} \sum_{t=1}^{T} \frac{1}{\beta_t} \|\mathbf{w}_t\|_2^2 - \sum_{i=1}^{n+m} \alpha_i y_i \sum_{t=1}^{T} \langle \mathbf{w}_t, \phi_t(\mathbf{x}_i) \rangle \\ \text{s.t.} \quad & \kappa \leq \sum_{i=1}^{n+m} \mathbf{1}_i \alpha_i\,, \quad 1 = \sum_{i=1}^{n+m} y_i \alpha_i\,, \quad 0 \leq \alpha_i \leq \eta_i\,, \quad i = 1, \ldots, n+m. \end{aligned} \tag{4.10}$$

We are solving the optimisation problem in a block-coordinate descent fashion by alternating between $\mathbf{w}$ and $\mathbf{d}$. This enables us to compute the latter analytically assuming fixed variables

Algorithm 3 Proposed optimization algorithm for MKL-SSAD (4.9)

Require: $\mathbf{x}, \mathbf{y}, \eta_u, \eta_l, \kappa$ & $p-$ norm
  Initialize kernel mixture coefficients such that $\|\mathbf{d}^{z=0}\|_p = 1$
  while Until Convergence **do**
    Step 1: solve the convex SSAD problem as stated in Eqn. (4.12)
    $\alpha^{z+1} = \text{argmax}_{\alpha: 0 \leq \alpha_i \leq \eta_i} J(\alpha, \mathbf{d}^z)$ s.t. $\kappa \leq \sum_{i=1}^{n+m} \mathbf{1}_i \alpha_i$ and $1 = \sum_{i=1}^{n+m} y_i \alpha_i$
    Step 2: optimize the weights according to Eqn. (4.11)
    $\mathbf{d}^{z+1} = \text{argmin}_{\mathbf{d} \geq 0} J(\alpha^{z+1}, \mathbf{d})$ s.t. $\|\mathbf{d}\|_p^2 \leq 1$
    z=z+1
  end while
  return Trained parameter vector $\alpha^\star$, weights $\mathbf{d}^*$

$\mathbf{w}$ and setting the partial derivative to zero:

$$\lambda \beta_t^{p-1} \|\mathbf{d}\|_p^{2-p} - \frac{\|\mathbf{w}_t\|_2^2}{\beta_t^2} = 0.$$

Therefore, given $\Upsilon \geq 0$ we get

$$\beta_t = \Upsilon \|\mathbf{w}_t\|_2^{\frac{2}{p+1}}.$$

Furthermore, it holds that at any optimal point $\|\mathbf{d}\|_p = 1$ and solving for Υ gives $\Upsilon = 1/(\sum_{t=1}^T \|\mathbf{w}_t\|_2^{\frac{2p}{p+1}})^{\frac{1}{p}}$. Putting things together, gives the analytical update rule

$$\beta_t = \frac{\|\mathbf{w}_t\|_2^{\frac{2}{p+1}}}{(\sum_{t=1}^T \|\mathbf{w}_t\|_2^{\frac{2p}{p+1}})^{\frac{1}{p}}} \tag{4.11}$$

which, since only norms are involved, ensures non-negativity for the mixing coefficients. Substituting $\mathbf{w}_t$ using the representer theorem $\mathbf{w}_t = \beta_t \sum_{i=1}^{n+m} \alpha_i \mathbf{y}_i \phi_t(\mathbf{x}_i)$ yields the final optimisation problem for MKL-SSAD:

$$\begin{aligned} \max_{\alpha} \min_{\mathbf{d}} & \overbrace{-\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \sum_{t=1}^T \beta_t k_t(\mathbf{x}_i, \mathbf{x}_j)}^{=: J(\alpha, \mathbf{d})} \\ \text{s.t.} \quad & \kappa \leq \sum_{i=1}^{n+m} \mathbf{1}_i \alpha_i \,, \quad 1 = \sum_{i=1}^{n+m} y_i \alpha_i \,, \quad 0 \leq \alpha_i \leq \eta_i \,, \quad i = 1, \ldots, n+m \\ & \|\mathbf{d}\|_p^2 \leq 1 \,. \end{aligned} \tag{4.12}$$

As a block-coordinate descent method, we can iteratively alternate between the two optimization blocks and every limit point of Algorithm 3 is a globally optimal point (cf. [105]). Algorithm 3 summarizes the proposed optimization procedure. To be comparable, kernels need to be centered and normalized.

4.4 Applications

We present two applications that show the benefits of our introduced methods. First, we will employ the techniques from Section 4.2 is exploratory data analysis in a EEG-BCI setting where no ground truth is available. However, the results given by applying our methodology

was assessed and analyzed by a domain expert. In the second application we will attempt to find authorships of disputed documents. To achieve state-of-the-art performance, which will be tested on a related dataset were labels are known, various feature representations of text need to be mixed to achieve highest possible accuracy.

4.4.1 Analysis of Brain States

The last years have seen a rise in interest in using Electroencephalography-based Brain Computer Interfacing (EEG-BCI) methodology for investigating non-medical questions beyond the purpose of communication and control. One of these novel applications is to examine how signal quality is being processed neurally, which is of particular interest for industry, besides providing neuroscientific insights. As for most behavioural experiments in the neurosciences, the assessment of a given stimulus by a subject is required. Based on an EEG study on speech quality of phonemes, we will first discuss the information contained in the neural correlate of this judgement. Typically, this is done by analyzing the data along behavioral responses/labels. However, participants in such complex experiments often guess at the threshold of perception. This leads to labels that are only partly correct and oftentimes random, which is a problematic scenario for using supervised learning. Therefore, we propose a novel supervised-unsupervised learning scheme based on techniques from Section 4.2, that aims to differentiate true labels from random ones in a data-driven way. We show that this approach provides a more crisp view of the brain states that experimenters are looking for, besides discovering additional brain states to which the classical analysis is blind.

EEG Experiment and Classical Analysis Understanding which levels of quality loss are still perceived by users is a crucial question for any provider of signal quality. Conventionally, behavioral tests are used for this purpose, asking participants directly for their rating. Recent work has proposed to complement this approach by also recording a user's neural response to a stimulus, as the neural response may differ from the behavioral response [122–124].

Eleven participants (mean age 25) took part in this study, for whom both behavioral and neural response was recorded using 64-channel EEG. Participants performed an auditory discrimination task in which they had to press a button whenever they detected an auditory stimulus of degraded quality (target). Stimuli were presented in an oddball paradigm, using the undisturbed phoneme /a/ as non-target (NT, 70% of stimuli). Among these stimuli of high quality, the participant had to find instances when the phoneme was superimposed with signal-correlated noise. Participants were instructed to indicate by button press, if they noticed a deviation in the stimulus. Four noisy target stimuli were used, T1-T4, consisting of the phoneme /a/ superimposed with decreasing levels of signal-correlated noise (targets; 6% per class). In an additional 6% of trials, the phoneme /i/ was presented as control stimulus (C; target). The noise levels of the target stimuli (T1-T4) were chosen separately for each participant, in order to account for individual differences in sensitivity to noise, aiming at perception rates of 100%, 75%, 25% and 0%, respectively. For this purpose, a pre-test was run; the resulting signal-to-noise ratios (SNR) for the deviant stimuli were set to 5, 21, 24 and 28 dB on average (mean perception rate in the experiment: 99%, 46%, 22% and 7%). The disturbed auditory stimuli were created using a Modulated Noise Reference Unit (MNRU). Target stimuli that were detected by the participant are referred to as 'hits' (true positives) and the others as 'misses' (false positives).
Each stimulus had a duration of 160 ms with 1000 ms stimulus onset asynchrony. Per participant, 8 to 12 blocks were recorded with 300 stimuli each. The button presses of the participants were recorded using a parallel port computer keyboard. For stimulus presentation, in-ear headphones by Sennheiser were used. EEG was recorded using a Brain Products (Munich, Germany) EEG system with 64 electrodes (AF3-4, 7-8; FAF1-2; Fz, 3-10; Fp1-2; FFC1-2,

5-8; FT7-10; FCz, 1-6; CFC5-8; Cz, 3-6; CCP7-8; CP1-2, 5-6; T7-8; TP7-10; P3-4, Pz, 7-8; POz; O1-2 and the right mastoid) and a BrainAmp EEG amplifier. Electrodes were placed according to the international 10-10 system. The tip of the nose was chosen as a reference site and a forehead ground electrode. EEG data were sampled at a rate of 100 Hz. In the following, we investigate event-related potentials (ERPs), i.e. the differential signal between the voltage at a given electrode position and the reference electrode.

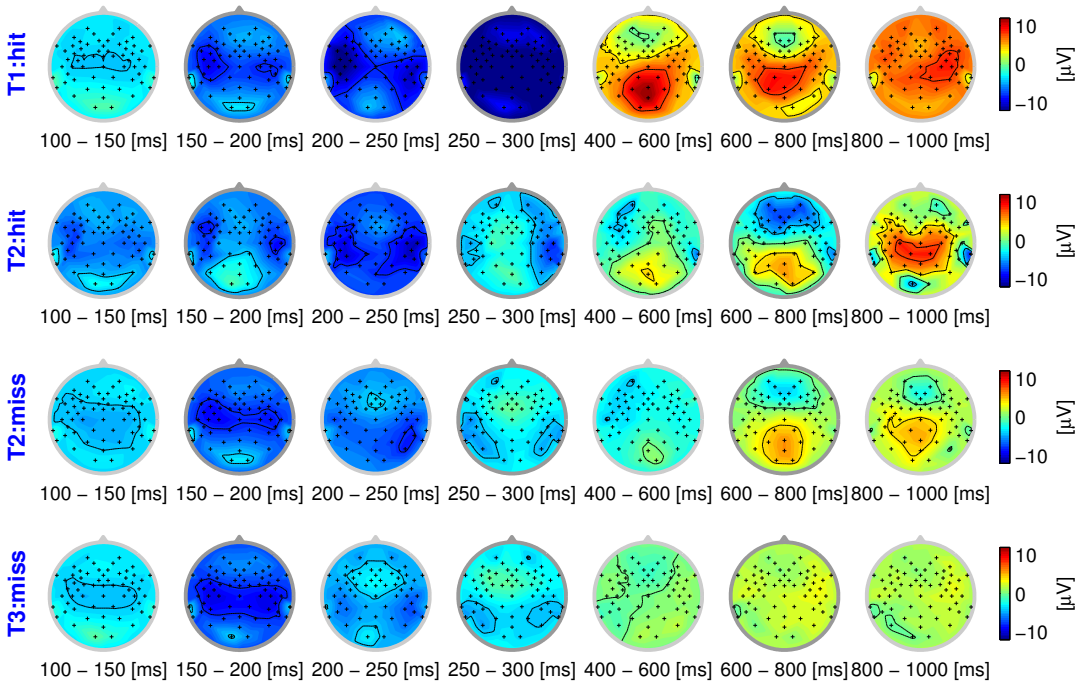

Figure 4.5 – Scalp distribution of ERPs for different stimuli in seven time intervals, grouped by their behavioral label [hit/miss] (participant vp=1). The maps represent a top view on the head with nose pointing upwards.

The behavioral responses of the participants provide labels for each trial, seemingly indicating whether the stimulus was perceived as disturbed or not. However, these labels can be assumed to be confounded with label noise to a large degree, in particular at the threshold of perception (stimulus T2). As a first step, we take these spurious labels as ground truth and analyze the event-related potentials in these groups. If the behavioral response indicates that the quality degradation is processed (hits), the resulting ERP activation pattern can be characterized by two components: early sensory and late cognitive processing stages. Figure 4.5 shows the spatial distribution of the ERPs as scalp distributions (head seen from above, nose pointing upwards), averaged over seven time intervals. The figure shows data exemplarily for one participant (vp 1). The top row shows the averaged neural response to a strong degradation that was noticed behaviorally (T1 hit). The four early intervals represent sensory processing of the stimulus (100–300ms post stimulus), which is reflected in a temporal negativity above the auditory cortices. In contrast, the last three intervals can be assumed to reflect cognitive processing (400–1000ms post stimulus). This elicits an occipital positivity, commonly referred to as P3 component. This component is elicited as a neural reaction to deviating stimuli in an oddball paradigm [125]. In our study, a P3 can be expected to occur when a participant notices that the quality of a stimulus is degraded. Generally speaking, the stronger the degradation, the higher the amplitude of the EEG signal, in particular that of the P3 component. This effect becomes obvious when comparing the first two rows of the figure,

with a much weaker activation during late intervals for stimulus T2 (weak degradation) compared to T1 (strong degradation). In contrast, the last row shows the neural processing of a stimulus with a subtle degradation that is not noticed on a behavioral level (T3 miss). While sensory processing still causes activity in the early intervals, there is no notable cognitive component.

Exploratory Data-driven Analysis While the topography of the averaged ERPs seemed to show a consistent picture so far, the presence of label noise becomes very obvious for the stimulus at the threshold of perception (T2). As the ratings of participants become unreliable to the point of guessing, grouping according to behavioral labels becomes conspicuously confounded, as can be seen in the second and third row of Figure 4.5. Even though the participant gave different ratings in these cases, the neural activation is strikingly similar. While the presence of label noise is obvious for this stimulus, the labels of the other classes can be expected to be confounded as well, just to a lesser degree. In the following, we will attempt to infer the correct labels in a data-driven way by using a novel learning methodology, in order to obtain an unbiased view of the EEG data.

We now turn to the details of the proposed supervised-unsupervised processing pipeline. The motivation behind this approach: even though it may seem that other methods (e.g. kernelized methods) could be more suitable for this problem, EEG data is well separable by linear classification (for a comparison of linear vs. nonlinear methods, cf. [126]). As discussed previously, the missing ground truth compels us to rely solely on interpretability of the results which can be achieved easily by applying *linear* and *sparse* methods. The inspection of the results of applying step 1 shows that there is a high chance of finding trials confounded by measurement noise (faulty electrodes) characterized by high amplitudes and/or drifts which we denote as artifacts. Therefore, we deliberately force the method to exclude such examples and search for other features by including the highest-ranked data points as outliers in a semi-supervised manner. Typically we chose 5 examples of each end of the spectrum to explicitly retain outlier labels (cf. Algorithm 4). We divide into three classes: core, plateau and outlier class. These classes occur naturally when applying the sparse one-class methods described in the previous section. Examples belonging to the plateau class are orthogonal to the core and outlier class and lie on the decision boundary. Hence, for division simple thresholding is sufficient.

Algorithm 4 Processing Pipeline

1: Given $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} = \mathcal{X}$ solve Eq. (4.7) preserving $\mathbf{w}_1^*$
2: Calculate the anomaly score $f_1(\mathbf{x}_i) = \langle \mathbf{w}_1^*, \mathbf{x}_i \rangle - 1$ for $i = \{1, \ldots, n\}$
3: Select subset $L_k \subseteq \mathcal{X}$ with $|L_k| = k$ and $L_k = \{\mathbf{x} \in \mathcal{X} \mid |f_1(\mathbf{x}_i)| \geq |f_1(\mathbf{x}_j)| \, \forall i \neq j\}$
4: Semi-supervised learning with elements from L_k tagged as outliers resulting in $\mathbf{w}_2^*$
5: Again, calculate the anomaly score $f_2(\mathbf{x}_i) = \langle \mathbf{w}_2^*, \mathbf{x}_i \rangle - 1$ for $i = \{1, \ldots, n\}$
6: Selecting the most confident examples $S = \{\mathbf{x} \in \mathcal{X} \mid f_2(\mathbf{x}_i) \geq 0, \, i = 1, \ldots, n\}$
7: Applying Eqn. (4.7) on S again, returns the final solution $f_3(\mathbf{x}) = \langle \mathbf{w}_3^*, \mathbf{x} \rangle$
8: Now, the sets $P_{\text{outlier}} = \{\mathbf{x} \in \mathcal{X} \mid f_3(\mathbf{x}_i) < 0, \, i = 1, \ldots, n\}$, $P_{\text{plateau}} = \{\mathbf{x} \in \mathcal{X} \mid f_3(\mathbf{x}_i) = 0, \, i = 1, \ldots, n\}$ and $P_{\text{core}} = \{\mathbf{x} \in \mathcal{X} \mid f_3(\mathbf{x}_i) > 0, \, i = 1, \ldots, n\}$ can be analyzed

The feature inputs are based on the time series of the ERPs. We first reduced the dimensionality of the data (cf [127]). Hence, we calculated the mean of the ERP signal within the seven neurophysiologically plausible intervals shown in Figure 4.5 (for each electrode and trial). For this, the EEG signal from 61 recorded electrodes was used (omitting the Fp and EO electrodes). Thus, the dimensionality of the data was reduced from 6400 (100 data points x

61 electrodes) to 427 features (7 data points x 61 electrodes). These features were then used as input for the processing pipeline.

The supervised-unsupervised learning approach groups the trials into three classes: a core class, an outlier class and a plateau class. These three classes can be seen exemplarily in Figure 4.6 for one participant (vp=1) and the stimulus at the threshold of perception (T2). Again, the scalp distribution of ERPs are shown in the seven intervals that were also used as input features. Remarkably, the core class (row 2) finds a very typical representation of hits with distinct auditory processing (first intervals) and a strong P3 component (last two intervals), suggesting that the degradation was consciously processed. This pattern is subdued in the plateau class (row 3), where the auditory cortices still show a strong activation, but only a very subtle P3 is visible, indicating that the degradation was processed on a sensory level, but not noticed by the participant. Finally, there is virtually no activation in early or late components for the outlier class (row 4), suggesting at most subliminal processing of the stimulus. This distinction is by far more cogent than that based on behavioral labels, where two classes were assumed (hit/miss) that were obviously confounded (middle rows of Figure 4.5). Not only does the algorithm find plausible classes, it also does so on the basis of neurophysiologically plausible features: As can be seen in the top row of the figure, the active features reflect the bi-temporal neural activity in early processing stages (auditory) and the occipital activity in late processing stages (cognitive).

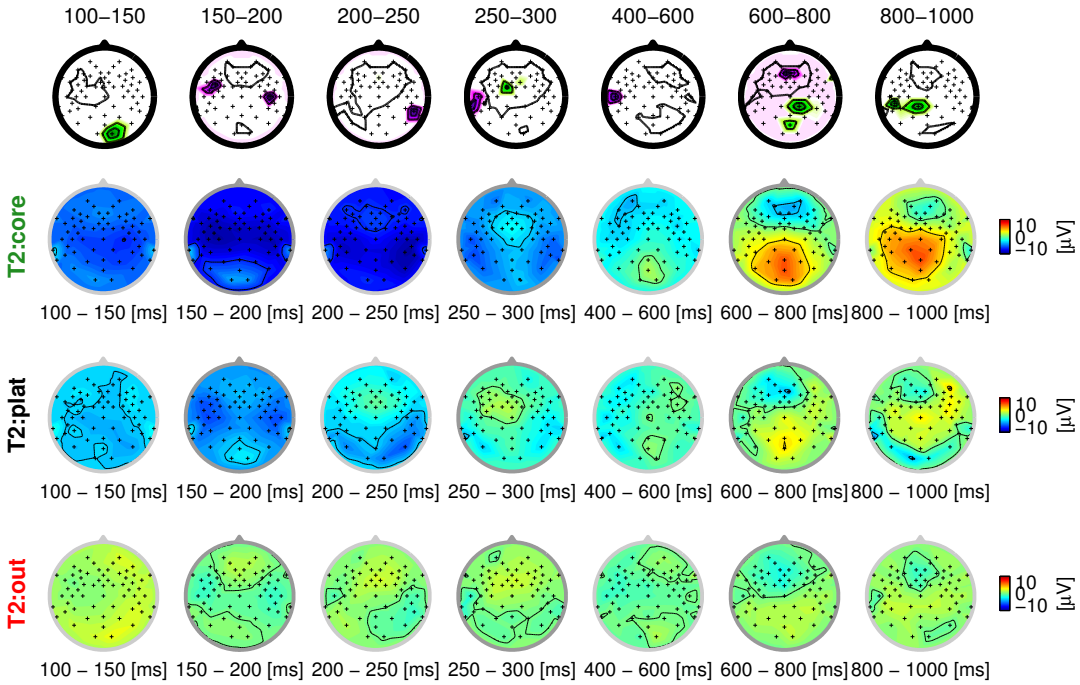

Figure 4.6 – Top row: weights of features (filter) assigned in the last step of the Algorithm (4). Bottom rows: scalp plots of the trials that are grouped into the core, plateau and outlier class (vp=1, T2).

Across all participants and stimuli, the trials grouped into the core class show a distinct representation of how the stimulus is processed, including both sensory and cognitive components ('neural hit') or only sensory processing ('neural miss'). For obvious degradations (C, T1), it is always the 'neural hit' that is found, while the algorithm rather assigns 'neural misses' to this class for subtle degradations. This is reasonable, as neural misses can be assumed to be predominant in those classes (the same is true for hits). In almost all cases (participants/stimuli), the outlier class represents trials that reflect a mental state other than these clear hit/miss patterns. Mostly, these are trials with very subdued activation (60% of

trials show an amplitude lower than +/-5μV on average), which indicates that the stimulus was processed at most at a subliminal level. Finally, the plateau class, where the EEG signal is orthogonal to the features chosen by the algorithm, contains a cluster of trials that differ most widely among participants. These either reflect measurement noise or eye artifacts (40%), a subdued pattern of neural hits/misses (30%) or a mental state other than that (20%). Figure 4.7 summarizes these results based on visual inspection.

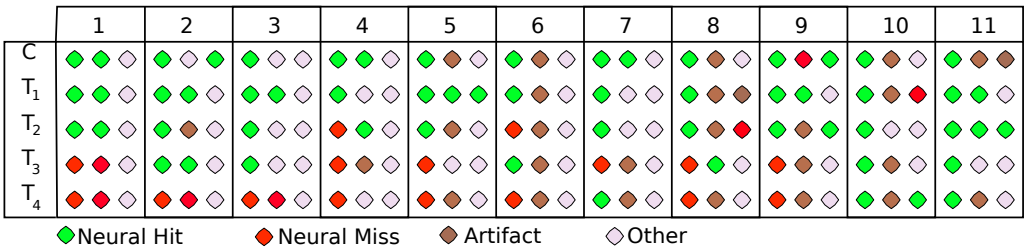

Figure 4.7 – Overview over all participants (x-axis) and stimuli (y-axis): neural pattern of core, plateau and outlier classes (column 1-3), based on visual inspection.

The motivation behind our approach is to find a coherent way to handle dependent label noise that is composed of a mixture of random labels and accurate ones. Figure 4.8 provides an insight into these ratios, as far as our approach can reveal them. First, it shows that the behavioral perception rate in black, i.e. the percentage of trials that were labeled as hits by the participants. As can be seen, the perception rate is high (almost 100%) for stimuli C/T1 and then drops markedly for stimuli T2-4 (left to right). Underneath these values, the figure shows which percentage of these behaviorial hits are assigned to the core, plateau or outlier class (ratios shown in gray, orange and white). This could be interpreted as the quantitative mixture of random labels and accurate ones.

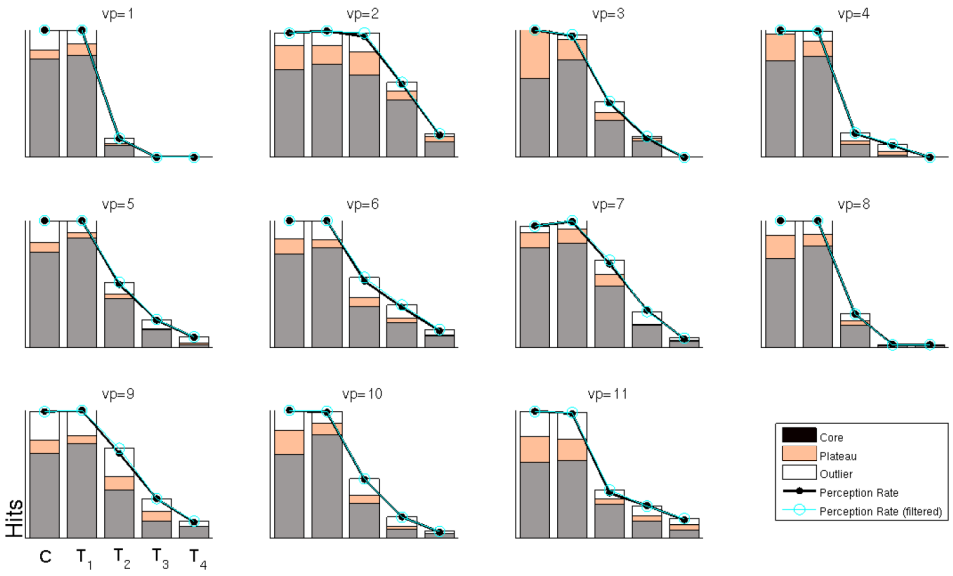

Figure 4.8 – Behavioral perception rate for all participants and target stimuli (C,T1-T4 from left to right), with the ratio of how many of these trials are grouped into core, plateau and outlier class (gray, orange and white box).

Application Outcome Analyzing EEG signals robustly, despite their high non-stationarity (cf. [128–131]), their multimodal nature, and the obviously noisy signal characteristics [11], is a major challenge that necessitates machine learning. However, in particular in complex cognitive tasks, the behavioral ratings given by participants are often unreliable, thus introducing label noise. Although in practice, independent label noise can be handled by most vanilla supervised learning algorithms, they can fail miserably in case of *dependent* label noise. This set-up is rather common in behavioural experiments where a subject is required to assess a given stimulus; in this work we have analyzed data from speech signal quality judgements. Near perception threshold, the behavioral responses of subjects provide labels that are noisy through a subjective assessment of the auditory signal. There are two reasons for this: (a) the subjects guess, i.e. the labels are random, (b) a very weakly correlated perception of a change in audio signal quality is reported that gives rise to a faint structure in the noisy labels. Computing the neural correlates of behaviour requires labels that reflect the task as clean as possible. To achieve this, we proposed a novel supervised-unsupervised learning procedure, that first removes artifactual trials from the experiment and then infers which of the remaining labels are reliable and which are random. Once these more reliable labels are in place, a better and more meaningful experimental evaluation of the neural correlates in our audio signal quality application can be performed. Moreover, our approach allows for defining groupings of trials that reflect more fine-grained cognitive states. The interesting point to note furthermore is that in this manner a neural correlate may occasionally be even more sensitive than the conscious behavioural one.

4.4.2 Authorship Attribution

Automatically attributing a piece of text to its author is one of the oldest problems studied in linguistics [132]. Despite being an old problem, authorship attribution is still highly topical and todays applications range from plagiarism detection [133], identifying the origin of anonymous harassments in emails, blogs, and chat rooms [134] to copyright and estate issues as well as resolving historical questions of disputed authorship [135, 136].

Intrinsically, the goal of authorship detection is to identify the characteristic traits of an author. The idea is that, these traits distinguish 'her' from other authors in terms of writing style, use of words, etc. Thus, prior work often focuses on designing and extracting features from text to capture these traits. There is a great deal of features proposed for authorship detection, including word or character n-grams [137, 138], part-of-speech [139], probabilistic context-free grammars [140], or linguistic features [141]. However, indicative features for one author do not necessarily help to characterise another. A major problem in authorship detection is therefore to find the right set of features for a given task at hand [142].

Algorithmically, a variety of different models have been studied in the context of authorship detection, ranging from probabilistic approaches [143] and similarity-based methods [144] to vector space models [136, 145]. The approaches either treat documents as independent (instance-based) or concatenate documents by the same author (profile-based). Intuitively, the latter is helpful if an author has a concise way of expressing herself so that the concatenated document allows to extract a statistic that is sufficient for capturing her style. On the other hand, instance-based approaches are better suited for expressive authors and have advantages in sparse data scenarios.

Another aspect in authorship attribution is the application scenario of the final model. In transductive (in-sample) settings, the unlabeled documents of interest are already included in the training process and the model does not necessarily perform well on new and unseen texts. By contrast, inductive (out-of-sample) scenarios generally allow to learn models that can be applied to any future text but require larger training samples to achieve accurate performances.

Here, we employ the techniques developed in Section 4.3. This remedies the above mentioned problems by fusing existing techniques: (i) We cast authorship attribution as an anomaly detection problem where one model is learned for every author. The idea is to identify a concise region in feature space that contains (most of) the documents of the author of interest while other documents are considered outliers. Thus, the model can be viewed as a profile-based approach in feature space while the data is treated on an instance-based level. (ii) We remedy the in-sample / out-of-sample problem by using a semi-supervised extension of the commonly unsupervised outlier detection framework. By doing so, we may include authorship labels for the already known documents and leave the disputed ones unlabeled. (iii) Finally, as the proposed approach is a member of the multiple kernel learning family this automatically includes a mathematically well founded feature selection framework that renders the method generally applicable. The optimal solution is given by a (possibly sparse) linear mixture of kernel functions.

Empirically, we observe that our approach consistently outperforms baseline competitors or confirms common knowledge with respect to the authorship of disputed articles. The main advantage of the method however lies in its simplicity. Practitioners do not need to take critical design choices in terms of which features to use and which not. By contrast, all features (kernels) can be used in the optimization and the method itself finds the optimal combination for the problem at hand.

Related Work Authorship attribution using linguistic and stylistic features has a long tradition and can be dated back to the nineteenth century. As a first attempt, Mendenhall [132] uses features based on word lengths to characterize the plays of Shakespeare. Later in the first half of the 20th century, different textual statistics, such as Zipf's distribution [146] and Yule's k-statistic [147] have been proposed to quantify textual style. Study by Mosteller et al. is one of the most influential modern work in authorship attribution[135]. They use a Bayesian approach to analyze frequencies of a small set of function words. Until the late 1990s, research in *stylometry* has been dominated by feature engineering to quantify writing style [148] and about 1,000 different measures have been proposed [149].

Document representation is essential for author attribution tasks. Features aim to capture characteristic traits of authors that persist across topics. Traditional stylometric features include function and high-frequency words, hapax legomena, Yules k-statistic, syllable distributions, sentence length, word length and word frequencies, vocabulary richness functions, syntactic and analysis. Many studies combine features of different types using multivariate analyses. Some researchers use punctuation symbols while others experiment with n-grams [150]. Grammatical style markers with natural language processing techniques are also used to extract features from the documents.

Also in terms of technical approaches, authorship attribution has been studied with a wide range of different approaches. The deployed techniques can be broadly divided into three categories: machine learning [150], multivariate/cluster analysis [151], and natural language processing [152]. Principal components analysis (PCA) is one of the widely used techniques for authorship studies, for instance, [153] apply PCA to identify the authorship of unknown articles that have been attributed to Stephen Crane. In addition, machine learning-based approaches, including support vector machines [150], are frequently used to discriminate documents by different authors. An excellent survey on the diversity of approaches for authorship detection is provided by [154].

Results on the Reuters 50-50 Data set We use a subset of the Reuters 50-50 data set[2] to evaluate the performance of the aforementioned approaches. The reduced data contains 1000

[2] `https://archive.ics.uci.edu/ml/datasets/Reuter_50_50`

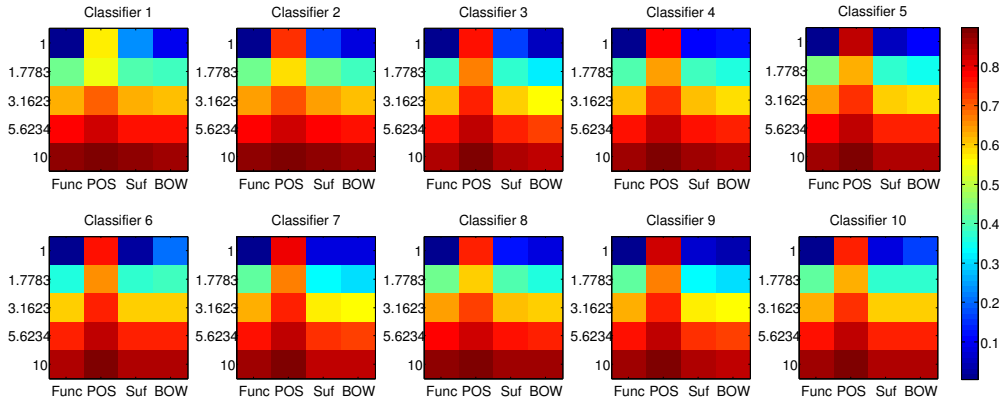

Figure 4.9 – Kernel mixture coefficients for the 10 classes

articles written by 10 authors, Aaron Pressman, Alan Crosby, Alexander Smith, Benjamin Kang Lim, Bernard Hickey, Brad Dorfman, Darren Schuettler, David Lawder, Edna Fernandes, and Eric Auchard.

We split the data into training (90%) and test (10%) sets and conduct a 10-fold cross-validation on the training set for model selection. The best performing models are then evaluated on the test set. We compare the performance of our approach with different p-norms to the SSAD which uses one kernel at a time. We use $p \in \{1, 1.7783, 3.1623, 5.6234, 10\}$. Note that $p = 10$ approximates the sum-kernel that would result by simply adding-up the four kernels.

The results in terms of averaged micro- and macro-F_1 measures are shown in Table 4.1. MKL consistently outperforms the single-kernel baseline for all p-norms. That is, instead of extensively experimenting with SSAD and different kernel functions and parameter selections, a single run with our MKL already leads to better performances in both metrics. Figure 4.9 visualizes the resulting mixing coefficients for the 10 authors/classifiers. While the models are very similar at first sight, small deviations indicate differences in the style of the authors.

Table 4.1 – F-scores for the subset of Reuters 50-50

	p-norm MKL					SSAD			
	1	1.7783	3.1623	5.6234	10	func-word	POS	Suffix-3	BOW
F_{micro}	73.46	73.08	73.84	73.89	74.23	63.08	54.62	70.01	72.85
F_{macro}	79.23	78.86	79.63	79.76	80.07	68.66	58.03	74.01	78.09

Revisiting the Federalist Papers The Federalist Papers are a series of 85 articles and essays written during 1787–1788. They were published anonymously to persuade the citizens of the State of New York to ratify the Constitution. Later, these papers were credited to Alexander Hamilton, John Jay, and James Madison; 73 of the documents are uniquely associated with one of the three authors while the remaining 12, also known as the disputed papers, have been claimed by both, Hamilton and Madison. Three of the 73 articles are considered joint work by Hamilton and Madison. Previous studies often assign all 12 disputed papers to Madison which we assume as ground-truth in the remainder [135, 136].

To confirm or refuse these previous findings, we conduct an experiment using the following four kernels as document representation: the first kernel is made of 484 function words taken from [155], the second contains part-of-speech (POS) tags, the third is assembled by 3-letter suffixes, the last one simply a bag-of-words (BOW) kernel. We compare the performance of our approach (MKL) with semi-supervised anomaly detection (SSAD) [12]. As before, the baseline cannot use all kernels at a time and are evaluated on every kernel separately. For simplicity, we show only the MKL results for parameter $p = 2$ as all other p-norms that we tried out lead to the same result.

We randomly divide the undisputed papers into training (80%) and holdout (20%) and use the 12 disputed papers for testing. We make sure that training sets contain at least three examples of every author and two articles written jointly by Hamilton and Madison. Otherwise we discard and draw again. We repeat experiments five times with randomly drawn training and holdout sets and report on averaged accuracies for the disputed test set.

The results are shown in Table 4.2. The one-class SVM and SVDD constantly credit the 12 disputed articles as joint work by Hamilton and Madison. The outcome of SSAD highly depends on the kernel function; while the part-of-speech kernel distributes the papers on Jay (3) and Hamilton and Madison (9), respectively, the bag-of-words kernel assigns all documents to Hamilton. By contrast, SVDD with function word and suffix-3 kernels attribute the articles to Madison. The same outcome is observed for our novel MKL that also credits the 12 papers to Madison. Thus, MKL and SSAD with function words and BoW kernel confirm todays assumption that all 12 papers have been written by Madison. However, chosing SSAD as the base classifier in the absence of prior knowledge leaves much room for interpretations and the user in the need of deciding between three solutions, depending on which kernel she prefers. By using our MKL, selecting features and or kernel functions is no longer necessary as the learning algorithm itself picks the right combination of kernels for the problem at hand. Thus, the more kernels are being used, the richer the decision space for the MKL.

Table 4.2 – Results for the disputed articles of the Federalist papers.

	kernel	H&M	M	J	H
MKL	(all)	0	**12**	0	0
SSAD	484fw	0	**12**	0	0
	POS	9	0	3	0
	Suffix3	0	**12**	0	0
	BoW	0	0	0	12

Application Outcome Our empirical results show the robustness of our approach as it consistently outperforms baseline competitors on a subset of Reuters 50-50 or confirms common knowledge wrt the authorship of disputed articles of the Federalist Papers. The main advantage of the method however lies in its simplicity. Practitioners do not need to take critical design choices in terms of which features to use and which not. By contrast, all features (kernels) can be used in the optimization and the method itself finds the optimal combination for the problem at hand.

4.5 Summary and Discussion

In this chapter, we introduced two techniques that allow to smoothly transition from dense solutions towards sparse solutions for unsupervised and semi-supervised one-class support vector machines based on ℓ_p-norm regularization. While in Section 4.2 we focused on controlling individual features, in Section 4.3 we turned our attention to groups of features. Further, we proposed corresponding optimization schemes and, in case of individual features and $p = 1$, a highly optimized linear program formulation. We applied the proposed

techniques on applications in EEG-BCI as well as authorship attribution were, in both cases, domain experts applied and analyzed the proposed methods.

While the results indicate that both methods work reasonable in their respective application, none of them is perfect and there is a number of drawbacks that we discuss here.

Limits of ℓ_p-norm regularized OC-SVM In Section 4.2, the proposed solver based on the sub-gradient descent algorithm does need a number of parameters including the choice of the step lengths. For diminishing and constant step length the algorithm has been shown to converge. In practice, however, this solver doesn't work very reliable and needs to be adjusted manually in order to converge in reasonable time or with reasonable error. More advanced methods based on proximal algorithms [49] are likely to be much more faster and less manual. Furthermore, the choice of p, besides $p = 1$ and $p = 2$, has no intrinsic rationale and can only be tested on a hold-out data set. Finally, when a parsimonious solution is the goal, we actually would like to solve the respective problem using ℓ_0-(pseudo)norm regularization. However, below $p < 1$ the optimization problem becomes non-convex and is not guaranteed to give meaningful results anymore. Finally, one of the most important properties of the one-class SVM is the ability to employ kernels. This flexibility is unfortunately lost in our approach.

Limits of ℓ_p-norm MKL SSAD The proposed approach iterates between finding the optimal weighting between kernels and solving the SSAD optimization problem which is computationally demanding especially when a large number of kernels is employed. However, finding the optimal weighting for multiple kernels can also be attempted in a heuristic fashion or using cross-validation techniques (in case of a smaller number of kernels). Moreover, it seems that a uniform combination of kernels is a reasonable choice as indicated in Table 4.1. However, the proposed MKL approach is a very systematic approach to incorporate, e.g. continuous and discrete features into a single feature description.

Source code and resources for the proposed methods are available on github [a]. Parts of this chapter are based on:

> **Görnitz, N.**, Kloft, M., Rieck, K., Brefeld, U., "Toward Supervised Anomaly Detection", *Journal of Artificial Intelligence Research (JAIR)*, vol. 46, pp. 235–262, 2013
>
> Porbadnigk, A., **Görnitz, N.**, Kloft, M., Müller, K.-R., "Decoding Brain States during Auditory Perception by Supervising Unsupervised Learning.", *Journal of Computing Science and Engineering (JCSE)*, vol. 7, no. 2, pp. 112–121, 2013
>
> Nasir, J. A., **Görnitz, N.**, Brefeld, U., "An Off-the-shelf Approach to Authorship Attribution", in *International Conference on Computational Linguistics (COLING)*, 2014, pp. 895–904

[a] https://github.com/nicococo/tilitools

III Collective Anomalies

Chapter 5

Learning with Structured Data

Structured data is ubiquitous: time series, graphs such as social networks, or trees for dependency parsing. Learning with that kind of data is a challenge that every machine learner faces on a daily basis. However, there are is a massive discrepancy in predicting structured outputs vs. using structured input.

Learning with structured input is a common setting and generally boils down to selecting an appropriate feature embedding for the bespoken structures. E.g. sequence data might be embedded using n-gram techniques or a sliding window approach [156]. Once features are fixed, standard techniques, such as binary support vector machines, can be applied without any changes.

Structured output methods, on the other hand, make complex predictions on groups or collections of data points while exploiting structural information of the input data. Complex prediction might include sequence, tree, or graph-like outputs hence, multiple interdependent variables (cf. Figure 5.1). This is in gross contrast to standard prediction methods that output single labels, e.g. classification or regression methods. Most prominent approaches include structure output support vector machines (SSVMs) [157], conditional random fields (CRFs) [158], and structured perceptrons [159] and have been applied to applications ranging from computational biology [27, 160] to speech recognition [161]. A comprehensive overview on the prediction of structured data is given in Bakir, Hofmann, Schölkopf, Smola, Taskar, and S.V.N. Vishwanathan [162].

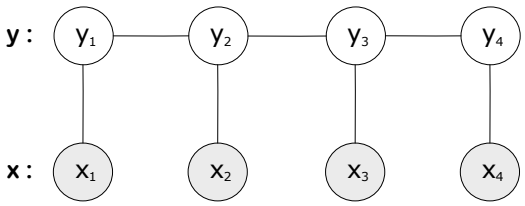

Figure 5.1 – A hidden Markov model as undirected graphical model (Markov random field, MRF). A sequence of observations $\mathbf{x}$ (grey circles) is coupled with corresponding latent labels $\mathbf{y}$.

Inferring interdependent labels ties a group of input data points together and provides a rich information source that would be well suited for collective (group) anomaly detection problems. However, such methods are generally supervised and require extensive labeling

on a very fine-grained level which is prohibitive in anomaly detection settings. Furthermore, they are notoriously hard to solve and only applicable for smaller and medium sized problems.

In this chapter, we derive an unsupervised anomaly detection method based on the structured output learning paradigm that is able to reliable spot anomalous groups of data points when those exhibit discrete label dependency structure (Section 5.3). Further, we introduce the supervised structured output learning paradigm and derive an efficient optimization method that enables large-scale learning (Section 5.2). Empirical evaluations on challenging computational biology tasks are presented in Section 5.4. To further differentiate both scenarios, we will use $\mathbf{z} \in \mathcal{Z}$ instead of $\mathbf{y} \in \mathcal{Y}$ in the unsupervised setting.

5.1 Preliminaries

In the supervised learning case, we assume that ground truth label structures and corresponding observations $\{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\} \in \mathcal{X} \times \mathcal{Y}$ are given for training purposes and hence, to adjust the parameter vector $\mathbf{w} \in \mathcal{H}$. Furthermore, we are interested in finding a predictor that is able to output the corresponding label structure $\hat{\mathbf{y}} \in \mathcal{Y}$ given the observations $\mathbf{x} \in \mathcal{X}$ and a joint feature map $\Psi : \mathcal{X} \times \mathcal{Y} \to \mathcal{H}$ that captures the dependencies between input and output (e.g., [157, 159]):

$$\hat{\mathbf{y}}(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}} \langle \mathbf{w}, \Psi(\mathbf{x}, \mathbf{y}) \rangle. \tag{5.1}$$

While this approach can be applied to very general structures, e.g. sequences, trees, and graphs, in this chapter, a special focus is set on large-scale sequence learning. That is, we assume that each data entry $\mathbf{x} \in \mathcal{X}$ contains a large number of samples that form a hidden Markov sequence. An example of such a model is given in Figure 5.1. Note that we are given a set of such observations $\mathbf{x} \in \mathcal{X}$ of variable length.

In the unsupervised anomaly detection setting, we will employ the same sort of technique but without given labels for training. That is, we are given only the observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \mathcal{X}$ and the desired output is, as standard in anomaly detection, an anomaly score for observation $\mathbf{x}$ (each consisting of a group of data points). However, the proposed method, *latent structure anomaly detector*, is build atop of the structure output principle (cf. Eq. (5.1)).

5.2 Large-scale Structured Output Learning

In contrast to binary classification, elements from the output space $\mathcal{Y}$ (e.g., sequences, trees, or graphs) of structured output problems have an inherent structure which makes more sophisticated, problem-specific loss functions desirable. The loss between the true label $\mathbf{y} \in \mathcal{Y}$ and the predicted label $\hat{\mathbf{y}} \in \mathcal{Y}$ is measured by a loss function $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$. A widely used approach to predict $\hat{\mathbf{y}} \in \mathcal{Y}$ is the use of a linearly parametrized model as given in Eq. (5.1). Generally, we are interested in finding the predictor with minimum risk given the data distribution P,

$$\mathcal{R}(\hat{\mathbf{y}}) = \int_{\mathcal{X} \times \mathcal{Y}} \Delta(\mathbf{y}, \hat{\mathbf{y}}(\mathbf{x})) dP(\mathbf{x}, \mathbf{y}).$$

The most common approaches to estimate the model parameters $\mathbf{w}$ are based on structured output SVMs (e.g., [157, 163]) and conditional random fields (e.g., [158]; see also [164]). Here we follow the approach taken in [157, 160], where estimating the parameter vector $\mathbf{w}$

using the margin rescaling variant amounts to solving the following optimization problem

$$\begin{aligned} \min_{\mathbf{w}\in\mathcal{H},\xi\geq 0} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & \langle \mathbf{w}, \Psi(\mathbf{x}_i,\mathbf{y}_i) - \Psi(\mathbf{x}_i,\bar{\mathbf{y}})\rangle \geq \Delta(\mathbf{y}_i,\bar{\mathbf{y}}) - \xi_i \quad \forall\, i,\, \bar{\mathbf{y}} \in \mathcal{Y}. \end{aligned} \tag{5.2}$$

Instead of taking all possible configurations $\bar{\mathbf{y}} \in \mathcal{Y}$ into account (which would be computational infeasible even for very small problems), a standard method for optimization iteratively computes the maximum violator $\max_{\bar{\mathbf{y}}\in\mathcal{Y}} \Delta(\mathbf{y}_i,\bar{\mathbf{y}}) - \langle \mathbf{w}, \Psi(\mathbf{x}_i,\mathbf{y}_i)\rangle$ based on the intermediate solution for $\mathbf{w} \in \mathcal{H}$ and then solves the resulting optimization problem. This way, a new constraint is generated per iteration for each example until convergence (cf. Algorithm 5). These kinds of optimization algorithms is referred to as column generation. For our purposes, we re-formulate the above problem into an unconstrained optimization problem and replace the loss function as well as the regularizer by some place-holders:

$$\min_{\mathbf{w}\in\mathcal{H}} \quad \left\{ R(\mathbf{w}) + C\sum_{i=1}^{n} \ell(\max_{\bar{\mathbf{y}}\in\mathcal{Y}} \langle \mathbf{w}, \Psi(\mathbf{x}_i,\bar{\mathbf{y}})\rangle + \Delta(\mathbf{y}_i,\bar{\mathbf{y}}) - \langle \mathbf{w}, \Psi(\mathbf{x}_i,\mathbf{y}_i)\rangle) \right\}, \tag{5.3}$$

where ℓ is the loss function and $R(\mathbf{x})$ the regularizer. For $\ell(a) = \max(0,a)$ and $R(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$ we obtain the structured output support vector machine with margin rescaling and hinge-loss as shown in Eq. (5.2).

In [26] the authors propose a hierarchical multitask structured output method that showed promising results in computational biology tasks. It turns out that we can now combine the structured output formulation with hierarchical multitask learning in a straight-forward way. We extend the regularizer $R(\mathbf{w})$ in (5.3) with a γ-parametrized convex combination of a multitask regularizer $\frac{1}{2}\,||\mathbf{w} - \mathbf{w}_p||_2^2$ with the original term. When $R(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$ and omitting constant terms, we arrive at $R_{p,\gamma}(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2 - \gamma\langle \mathbf{w}, \mathbf{w}_p\rangle$. Thus we can apply the mentioned hierarchical multitask learning approach and solve for every node the following optimization problem:

$$\min_{\mathbf{w}\in\mathcal{H}} \quad \left\{ R_{p,\gamma}(\mathbf{w}) + C\sum_{i=1}^{n} \ell(\max_{\bar{\mathbf{y}}\in\mathcal{Y}} \langle \mathbf{w}, \Psi(\mathbf{x}_i,\bar{\mathbf{y}})\rangle + \Delta(\mathbf{y}_i,\bar{\mathbf{y}}) - \langle \mathbf{w}, \Psi(\mathbf{x}_i,\mathbf{y}_i)\rangle) \right\} \tag{5.4}$$

A major difficulty remains: solving the resulting optimization problems which now can become considerably larger than for the single-task case.

Column generation techniques often converge slowly. Moreover, the size of the restricted optimization problems grows steadily and solving them becomes more expensive in each iteration. Simple gradient descent or second order methods can not be directly applied as alternatives, because (5.4) is continuous but non-smooth. Our approach is instead based on bundle methods for regularized risk minimization as proposed in [165, 166] and [167]. In case of SVMs, this further relates to the OCAS method introduced in [168]. In order to achieve fast convergence, we use a variant of these methods adapted to structured output learning that is suitable for hierarchical multitask learning.

We consider the objective function $J(\mathbf{w}) = R_{p,\gamma}(\mathbf{w}) + L(\mathbf{w})$, where

$$L(\mathbf{w}) := C\sum_{i=1}^{n} \ell(\max_{\bar{\mathbf{y}}\in\mathcal{Y}} \{\langle \mathbf{w}, \Psi(\mathbf{x}_i,\bar{\mathbf{y}})\rangle + \Delta(\mathbf{y}_i,\bar{\mathbf{y}})\} - \langle \mathbf{w}, \Psi(\mathbf{x}_i,\mathbf{y}_i)\rangle).$$

Many machine learning methods adjust their parameters by minimizing directly the empirical loss. On the contrary for structure output support vector machines, evaluating the empirical loss $L(\mathbf{w})$ implies the invocation of the embedded argmax-function. We argue that it

Algorithm 5 Column Generation Method for Structured Output Learning

$\mathbf{w}^{(1)} = \mathbf{w}_p$
$k = 1$ and $\Gamma_i = \emptyset \quad \forall i$
repeat
  for $i = 1, .., n$ **do**
    $\mathbf{y}^* = \text{argmax}_{\mathbf{y} \in \mathcal{Y}} \{\langle \mathbf{w}^{(k)}, \Psi(\mathbf{x}_i, \mathbf{y}) \rangle + \Delta(\mathbf{y}_i, \mathbf{y})\}$
    if $\langle \mathbf{w}^{(k)}, \Psi(\mathbf{x}_i, \mathbf{y}^*) \rangle + \Delta(\mathbf{y}_i, \mathbf{y}^*) > \max_{(\Psi, \Delta) \in \Gamma_i} \{\langle \mathbf{w}^{(k)}, \Psi \rangle + \Delta\}$ **then**
      $\Gamma_i = \Gamma_i \cup (\Psi(\mathbf{x}_i, \mathbf{y}^*), \Delta(\mathbf{y}_i, \mathbf{y}^*))$
    end if
  end for
  $\mathbf{w}^{(k)} = \underset{\mathbf{w} \in \mathcal{H}}{\text{argmin}} \left\{ R_{p,\gamma}(\mathbf{w}) + C \sum_{i=1}^{n} \ell \left(\max_{(\Psi, \Delta) \in \Gamma_i} \{\langle \mathbf{w}, \Psi \rangle + \Delta\} - \langle \mathbf{w}, \Psi(\mathbf{x}_i, y_i) \rangle \right) \right\}$
  $k = k + 1$
until no changes in (Ψ_i, Δ_i) $\forall i$

is the most expensive step in general and therefore effective methods as e.g. line search are practically prohibitive. Instead, we propose to optimize an estimate of the empirical loss $\hat{L}(\mathbf{w})$, which can be computed efficiently. We define the estimated empirical loss $\hat{L}(\mathbf{w})$ as

$$\hat{L}(\mathbf{w}) := C \sum_{i=1}^{N} \ell \left(\max_{(\Psi, \Delta) \in \Gamma_i} \{\langle \mathbf{w}, \Psi \rangle + \Delta\} - \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}_i) \rangle \right).$$

Accordingly, we define the estimated objective function as $\hat{J}(\mathbf{w}) = R_{p,\gamma}(\mathbf{w}) + \hat{L}(\mathbf{w})$. It is easy to verify that $J(\mathbf{w}) \geq \hat{J}(\mathbf{w})$. Γ_i is a set of pairs $(\Psi(\mathbf{x}_i, \mathbf{y}), \Delta(\mathbf{y}_i, \mathbf{y}))$ defined by a suitably chosen, growing subset of $\mathcal{Y}$, such that $\hat{L}(\mathbf{w}) \to L(\mathbf{w})$ (cf. Algorithm 6).

In general, bundle methods are extensions of cutting plane methods that employ a proximal operator [49] to stabilize the solution of the approximated function. In the framework of regularized risk minimization, a natural input to the proximal operator is given by the regularizer. We apply this approach to the objective $\hat{J}(\mathbf{w})$ and solve

$$\min_{\mathbf{w} \in \mathcal{H}} R_{p,\gamma}(\mathbf{w}) + \max_{i \in I} \{\langle \mathbf{a}_i, \mathbf{w} \rangle + b_i\}, \quad (5.5)$$

where the set of cutting planes $\mathbf{a}_i$, b_i lower bound $\hat{L}$. As proposed in [166, 167], we use a set I of limited size. Moreover, we calculate an aggregation cutting plane $\bar{\boldsymbol{a}}$, $\bar{b}$ that lower bounds the estimated empirical loss $\hat{L}$. To be able to solve the primal optimization problem in (5.5) in the dual space as proposed by [166, 167], we adopt an elegant strategy described in [167] to obtain the aggregated cutting plane $(\bar{\boldsymbol{a}}', \bar{b}')$ using the dual solution α of (5.5):

$$\bar{\boldsymbol{a}}' = \sum_{i \in I} \alpha_j \mathbf{a}_i \qquad \text{and} \qquad \bar{b}' = \sum_{i \in I} \alpha_i b_i. \quad (5.6)$$

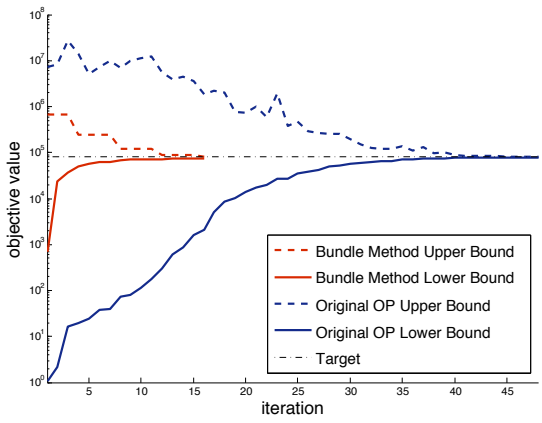

Figure 5.2 – Convergence behavior of the proposed method (red, cf. Alg. (6)) against the standard column generation approach (blue, cf. Alg. (5)).

The following two formulations reach the same minimum when optimized with respect to $\mathbf{w}$:

$$\min_{\mathbf{w}\in\mathcal{H}} \left\{ R_p(\mathbf{w}) + \max_{i\in I} \langle \mathbf{a}_i, \mathbf{w}\rangle + b_i \right\} = \min_{\mathbf{w}\in\mathcal{H}} \{R_p(\mathbf{w}) + \langle \bar{\boldsymbol{a}}', \mathbf{w}\rangle + \bar{b}'\}.$$

This new aggregated plane can be used as an additional cutting plane in the next iteration step. We therefore have a monotonically increasing lower bound on the estimated empirical loss and can remove previously generated cutting planes without compromising convergence (see [167] for details).

The algorithm is able to handle any (non-)smooth convex loss function ℓ, since only the subgradient needs to be computed. This can be done efficiently for the hinge-loss, squared hinge-loss, Huber-loss, and logistic-loss.

The resulting optimization algorithm is outlined in Algorithm 6. There are several improvements possible: For instance, one can bypass updating the empirical risk estimates in line 6, when $L(\mathbf{w}^{(k)}) - \hat{L}(\mathbf{w}^{(k)}) \leq \epsilon$. Finally, while Algorithm 6 was formulated in primal space, it is easy to reformulate in dual variables making it independent of the dimensionality of $\mathbf{w} \in \mathcal{H}$.

Algorithm 6 Bundle Methods for Structured Output Algorithm

$S \geq 1$: maximal size of the bundle set
$\theta > 0$: line-search trade-off (cf. [168] for details)
$\mathbf{w}^{(1)} = \mathbf{w}_p$
$k = 1$ and $\bar{\boldsymbol{a}} = \mathbf{0}, \bar{b} = 0, \Gamma_i = \emptyset \quad \forall i$
repeat
  for $i = 1, .., n$ **do**
    $\mathbf{y}^* = \operatorname{argmax}_{\mathbf{y}\in\mathcal{Y}} \{\langle \mathbf{w}^{(k)}, \Psi(\mathbf{x}_i, \mathbf{y})\rangle + \Delta(\mathbf{y}_i, \mathbf{y})\}$
    if $\ell\left(\max_{\mathbf{y}\in\mathcal{Y}} \{\langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y})\rangle + \Delta(\mathbf{y}_i, \mathbf{y})\}\right) > \ell\left(\max_{(\Psi,\Delta)\in\Gamma_i} \langle \mathbf{w}, \Psi\rangle + \Delta\right)$ **then**
      $\Gamma_i = \Gamma_i \cup (\Psi(\mathbf{x}_i, \mathbf{y}^*), \Delta(\mathbf{y}_i, \mathbf{y}^*))$
    end if
    Compute $\mathbf{a}_k \in \partial_{\mathbf{w}} \hat{L}(\mathbf{w}^{(k)})$
    Compute $b_k = \hat{L}(\mathbf{w}^{(k)}) - \langle \mathbf{w}^{(k)}, \mathbf{a}_k\rangle$
    $\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w}\in\mathcal{H}} \left\{ R_{p,\gamma}(\mathbf{w}) + \max\left(\max_{(k-S)_+<i\leq k} \{\langle \mathbf{a}_i, \mathbf{w}\rangle + b_i\}, \langle \bar{\boldsymbol{a}}, \mathbf{w}\rangle + \bar{b}\right)\right\}$
    Update $\bar{\boldsymbol{a}}, \bar{b}$ according to (5.6)
    $\eta^* = \operatorname{argmin}_{\eta\in\mathbb{R}} \hat{J}(\mathbf{w}^* + \eta(\mathbf{w}^* - \mathbf{w}^{(k)}))$
    $\mathbf{w}^{(k+1)} = (1-\theta)\mathbf{w}^* + \theta\eta^*(\mathbf{w}^* - \mathbf{w}^{(k)})$
    $k = k + 1$
  end for
until $L(\mathbf{w}^{(k)}) - \hat{L}(\mathbf{w}^{(k)}) \leq \epsilon$ and $J(\mathbf{w}^{(k)}) - J_k(\mathbf{w}^{(k)}) \leq \epsilon$

An illustration of the convergence behavior of the proposed bundle method is shown in Fig. 5.2. Here, it can be clearly seen that column generation needs many more iterations and hence, much more time in order to converge.

5.3 Latent Structure Anomaly Detection

Building upon the previously discussed structured output methods, we will now exploit these techniques for unsupervised collective anomaly detection. As always, we start with the standard formulation of the one-class SVM,

$$\begin{aligned} \min_{\mathbf{w},\rho,\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \rho \\ \text{s.t.} \quad & \langle \mathbf{w}, \mathbf{x}_i \rangle \geq \rho - \xi_i, \quad \xi_i \geq 0, \quad \forall i \in \{1, \ldots, n\}, \end{aligned} \tag{5.7}$$

and take a closer look at the constraints. Here, the value of the linear function is forced to stay above threshold ρ for the bulk of the data. If we abstract from the linear function and allow arbitrary scoring functions $s_{\mathbf{w}} : \mathcal{X} \to \mathbb{R}$ the constraints will change to $s_{\mathbf{w}}(\mathbf{x}_i) \geq \rho - \xi_i$. Assuming an adequate (parameterized) model of the probability of a given sample $\mathbf{x}$, $s_{\mathbf{w}}(\mathbf{x}) = \log p_{\mathbf{w}}(\mathbf{x})$ is available. As a result, the one-class SVM picks high-scoring, high-probability samples as nominal and low-scoring, low-probability samples as anomalies which is exactly as intended. Extending the idea to structures $\mathbf{z} \in \mathcal{Z}$ gives

$$p_{\mathbf{w}}(\mathbf{x}, \mathbf{z}) = p_{\mathbf{w}}(\mathbf{x}|\mathbf{z})p(\mathbf{z}) = \frac{1}{Z(z)} \exp(\langle \mathbf{w}, \Psi(\mathbf{x}, \mathbf{z}) \rangle) \cdot \eta \exp(\delta(\mathbf{z})), \tag{5.8}$$

where we assume log-linear models for the conditional probability $p(\mathbf{x}|\mathbf{z})$ and some Gaussian prior over pre-defined penalties $\delta : \mathcal{Z} \to \mathbb{R}^-$ with corresponding normalizations η. Of course, structure are supposed to be unknown and a standard way of achieving this, is to marginalize over all possible configurations of $\mathbf{z} \in \mathcal{Z}$. However, we argue that *maximum a posteriori* will be beneficial wrt. computational efforts. Hence, taking the log and discarding unnecessary terms, we end up with the final scoring function

$$s_{\mathbf{w}}(\mathbf{x}) = \max_{\mathbf{z} \in \mathcal{Z}} \langle \mathbf{w}, \Psi(\mathbf{x}, \mathbf{z}) \rangle + \delta(\mathbf{z}) \tag{5.9}$$

In the problem setting of *latent structure anomaly detection*, we extend the expressiveness of the one-class SVM as given in Eq. 5.7 by considering models of the form $f_{\mathbf{w},\rho}(x) = \max_{z \in \mathcal{Z}} \langle \mathbf{w}, \Psi(x, z) \rangle + \delta(z) - \rho$, where $\Psi : \mathcal{X} \times \mathcal{Z} \to \mathcal{H}$ is a joint feature map into a reproducing kernel Hilbert space $\mathcal{H}$ that corresponds to a kernel function $k : (\mathcal{X} \times \mathcal{Z}) \times (\mathcal{X} \times \mathcal{Z}) \to \mathbb{R}$. This is a principled way of approaching the encoding problem for arbitrary dependencies between x and z as it is common in the structured output literature [157]. Albeit, it has been already used to encode hidden Markov and hidden semi-Markov models [26, 160], it is not restricted to those and has been applied to Markov random fields [169], weighted context-free grammars and taxonomies [157]. Here, the maximization step for the latent variable z acts as a frequentist's equivalent to marginalization in basic probability theory [169].

Employing the above notation, we phrase the primal optimization problem of latent anomaly detection as follows:

Problem 6 (Primal latent anomaly detection optimization problem)**.** *Given a monotonically non-decreasing loss function $l : \mathbb{R} \to \mathbb{R}$, minimize, with respect to $\mathbf{w} \in \mathcal{H}$ and $\rho \in \mathbb{R}$,*

$$\frac{1}{2}\|\mathbf{w}\|^2 - \rho + \frac{1}{\nu n}\sum_{i=1}^{n} l\Big(\rho - \max_{z \in \mathcal{Z}} \big(\langle \mathbf{w}, \Psi(x_i, z) \rangle + \delta(z)\big)\Big). \tag{P}$$

The interpretation of the above formulation is as follows. The loss function could be, e.g., $l(t) = \max(0, t)$, in which case the above detection method extends the one-class support vector machine [7] to the latent domain (this is extensively discussed in the upcoming Section 5.3). Variants of this detection method can be obtained from the above general formulation by employing different loss functions, e.g., of logistic or exponential type ($l(t) = \log(1 + \exp(t))$ and $l(t) = \exp(t)$, respectively). It is important to note that, when contrasted to the classical kernel-based hypothesis model $f_{\mathbf{w},\rho}(\phi(x)) = \langle \mathbf{w}, \phi(x) \rangle -$

ρ, the above detection method employs a latent hypothesis model of the form $f_{\mathbf{w},\rho}(x) = \max_{z\in\mathcal{Z}}\langle \mathbf{w}, \Psi(x,z)\rangle + \delta(z) - \rho$, which allows for additional flexibility.

To obtain a dual representation of the Problem 6, we start by equivalently re-writing (P) as

$$\begin{aligned}\min_{\mathbf{w}\in\mathcal{H},\rho\in\mathbb{R},\xi\in\mathbb{R}^n}\ & \frac{1}{2}\left\|\mathbf{w}\right\|^2 - \rho + \frac{1}{\nu n}\sum_{i=1}^{n} l(\xi_i)\\ \text{s.t.}\ \ & \xi_i \geq \rho - \max_{z\in\mathcal{Z}}\big(\,\langle \mathbf{w}, \Psi(x_i,z)\rangle + \delta(z)\big),\ \forall i\end{aligned}$$

Denote, for all $\alpha \in \mathbb{R}^n$ with $\alpha \geq 0$,[1] the Lagrangian by

$$\mathcal{L}(\mathbf{w},\rho,\xi,\alpha) := \frac{1}{2}\left\|\mathbf{w}\right\|^2 - \rho + \frac{1}{\nu n}\sum_{i=1}^{n} l(\xi_i) + \sum_{i=1}^{n}\alpha_i\Big(\rho - \xi_i - \max_{z\in\mathcal{Z}}\big(\,\langle \mathbf{w}, \Psi(x_i,z)\rangle + \delta(z)\big)\Big).$$

By *weak* duality (e.g., [43], Chapter 5),

$$\begin{aligned}\text{Eq. (P)} \ \geq\ \max_{\alpha:\alpha\geq 0}\ \min_{\mathbf{w}\in\mathcal{H},\rho\in\mathbb{R},\xi\in\mathbb{R}^n}\ \mathcal{L}(\mathbf{w},\rho,\xi,\alpha) \ =\ \max_{\alpha:\alpha\geq 0}\ \Bigg(& -\frac{1}{\nu n}\sum_{i=1}^{n}\max_{\xi_i\in\mathbb{R}}\Big(\alpha_i\nu n\xi_i - l(\xi_i)\Big)\\ & + \min_{\rho\in\mathbb{R}}\rho\Big(-1+\sum_{i=1}^{n}\alpha_i\Big) - \underbrace{\max_{\substack{\mathbf{w}\in\mathcal{H},z_i\in\mathcal{Z}\\ i=1,\ldots,n}}\sum_{i=1}^{n}\alpha_i\big(\,\langle \mathbf{w}, \Psi(x_i,z_i)\rangle + \delta(z_i)\big) - \frac{1}{2}\left\|\mathbf{w}\right\|^2}_{(*)}\ \Bigg)\end{aligned}$$

Let $\mathbf{w}^\alpha$ and $(z_i^\alpha)_{i=1,\ldots,n}$ be the maximizing arguments in $(*)$. Thus $\max_{z_i\in\mathcal{Z}}\langle \mathbf{w}^\alpha, \Psi(x_i,z_i)\rangle + \delta(z_i) = \langle \mathbf{w}^\alpha, \Psi(x_i,z_i^\alpha)\rangle + \delta(z_i^\alpha)$, and $\max_{z_i\in\mathcal{Z}}\langle \mathbf{w}, \Psi(x_i,z_i)\rangle + \delta(z_i) \geq \langle \mathbf{w}, \Psi(x_i,z_i^\alpha)\rangle + \delta(z_i^\alpha)$ for all $\mathbf{w}\in\mathcal{H}$ and $i = 1,\ldots,n$. Hence, for all $\alpha\in\mathbb{R}^n_+$,

$$(*) \ =\ \max_{\mathbf{w}\in\mathcal{H}}\sum_{i=1}^{n}\alpha_i\big(\,\langle \mathbf{w}, \Psi(x_i,z_i^\alpha)\rangle + \delta(z_i^\alpha)\big) - \frac{1}{2}\left\|\mathbf{w}\right\|^2,$$

from which it follows $\mathbf{w}^\alpha = \sum_{i=1}^{n}\alpha_i\Psi(x_i,z_i^\alpha)$, and thus

$$(*) = \max_{\substack{z_i\in\mathcal{Z}\\ i=1,\ldots,n}}\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j k\Big((x_i,z_i),(x_j,z_j)\Big) + \sum_{i=1}^{n}\alpha_i\delta(z_i).$$

Hence,

$$\begin{aligned}&\max_{\alpha:\alpha\geq 0}\ \min_{\mathbf{w}\in\mathcal{H},\rho\in\mathbb{R},\xi\in\mathbb{R}^n}\ \mathcal{L}(\mathbf{w},\rho,\xi,\alpha) \ =\ \max_{\alpha:\alpha\geq 0}\ \Bigg(-\frac{1}{\nu n}\sum_{i=1}^{n}\max_{\xi_i\in\mathbb{R}}\Big(\alpha_i\nu n\xi_i - l(\xi_i)\Big)\\ &+ \min_{\rho\in\mathbb{R}}\rho\Big(-1+\sum_{i=1}^{n}\alpha_i\Big) - \max_{\substack{z_i\in\mathcal{Z}\\ i=1,\ldots,n}}\Bigg(\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j k\Big((x_i,z_i),(x_j,z_j)\Big) + \sum_{i=1}^{n}\alpha_i\delta(z_i)\Bigg)\Bigg)\\ &\overset{(\dagger)}{=} \max_{\alpha:\alpha\geq 0,\sum_{i=1}^{n}\alpha_i=1}\ \Bigg(-\frac{1}{\nu n}\sum_{i=1}^{n} l^*(\alpha_i\nu n)\\ &\qquad\qquad - \max_{\substack{z_i\in\mathcal{Z}\\ i=1,\ldots,n}}\Bigg(\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j k\Big((x_i,z_i),(x_j,z_j)\Big) + \sum_{i=1}^{n}\alpha_i\delta(z_i)\Bigg)\Bigg)\end{aligned}$$

[1] For vectors $x\in\mathbb{R}^n$, we denote by $\mathbf{x}\geq 0$ as the component-wise inequalities $x_i\geq 0$, $i = 1,\ldots,n$.

where for (†) we employ the notion of the Fenchel-Legendre convex conjugate function $f^*(\mathbf{a}) := \sup_{\mathbf{b}} \langle \mathbf{a}, \mathbf{b} \rangle - f(\mathbf{b})$ [170] and exploit that the function $\mathbf{w} \mapsto \frac{1}{2} \|\cdot\|^2$ is self-conjugated; as well as we observe that $\min_{\rho \in \mathbb{R}} \rho\Big(-1 + \sum_{i=1}^n \alpha_i\Big) = 0$ if $\sum_{i=1}^n \alpha_i = 1$ and $-\infty$ else-wise, which enforces the constraint $\sum_{i=1}^n \alpha_i = 1$ when maximizing with respect to α. Thus we obtain the following dual optimization problem of (P).

Problem 7 (Dual latent anomaly detection optimization problem). *Given a monotonically non-decreasing loss function $l : \mathbb{R} \to \mathbb{R}$, and denoting by the $l^* : \mathbb{R} \to \mathbb{R}$ the dual loss function, maximize, with respect to $\alpha \in \mathbb{R}^n$ and subject to $\alpha \geq 0$ and $\sum_{i=1}^n \alpha_i = 1$,*

$$- \min_{\substack{z_i \in \mathcal{Z} \\ i=1,\ldots,n}} \left(\frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j k\big((x_i, z_i), (x_j, z_j)\big) + \sum_{i=1}^n \alpha_i \delta(z_i) \right) - \frac{1}{\nu n} \sum_{i=1}^n l^*(\alpha_i \nu n) \,. \tag{D}$$

The minimization over $z \in \mathcal{Z}$ can be expanded into slack variables, so the above dual becomes a quadratically constrained program (QCQP) with $n \cdot |\mathcal{Z}|$ many quadratic constraints.

By the above dualization the prediction function can be written as

$$f(x) = \max_{z \in \mathcal{Z}} \left(\sum_{i=1}^n \alpha_i k\big((x_i, z_i^\alpha), (x, z)\big) + \delta(z) \right) - \rho.$$

where ρ can be calibrated by line search such that exactly a fraction of $1 - \nu$ training points satisfy $f(x_i) \geq 0$. The corresponding estimated density-level set is given by $\hat{L}_\nu := \{x \in \mathcal{X} : f(x) \geq 0\}$.

For the theoretical analysis, we consider a slight variation of latent anomaly detection,

$$\begin{aligned} \min_{\mathbf{w} \in \mathcal{H}} \quad & \frac{1}{n} \sum_{i=1}^n l\Big(1 - \max_{z \in \mathcal{Z}} \big(\langle \mathbf{w}, \Psi(x_i, z) \rangle + \delta(z) \big)\Big) \\ \text{s.t.} \quad & \|\mathbf{w}\| \leq C \,. \end{aligned} \tag{5.10}$$

For the important choice of $l(t) = \max(0, t)$ studied in Section 5.3, the above reformulation is equivalent to the original problem (P), in the sense for any choice of ν in (P), there exists a choice of $C > 0$ in (5.10) such that both problems have the same solution in the variable $\mathbf{w}$. This is shown in Supplementary Material A.1.

To analyze (5.10) theoretically, note that (5.10) corresponds to performing empirical risk minimization (ERM), $\hat{f} := \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n l(f(x_i))$ over the class $\mathcal{F} := \{f_{\mathbf{w}} = \big(x \mapsto 1 - \max_{z \in \mathcal{Z}} (\langle \mathbf{w}, \Psi(x, z) \rangle + \delta(z))\big) : \|\mathbf{w}\| \leq C\}$. In the following theorem, we show that the solution of (5.10) has asymptotically the same loss as the theoretically optimal quantity $f^* := \operatorname{argmin}_{f \in \mathcal{F}} E\, l(f(X))$.

Theorem 8 (Latent anomaly detection generalization bound). *The following generalization bound holds for the latent anomaly detection method* (A.1). *Let $l : \mathbb{R} \to \mathbb{R}$ be a non-negative and L-Lipschitz continuous loss function. Denote $A := \max_{z \in \mathcal{Z}} |\delta(z)|$ and $B := \max_{x \in \mathcal{X}, z \in \mathcal{Z}} \|\Psi(x, z)\|$. With probability at least $1 - \epsilon$ over the draw of the sample, the generalization error is bounded as:*

$$E\, l(\hat{f}) - E\, l(f^*) \;\leq\; 8L \frac{1 + A + BC\,|\mathcal{Z}|}{\sqrt{n}} \;+\; L(1 + A + BC) \sqrt{\frac{2 \log(2/\epsilon)}{n}} \,.$$

Proof. The full proof is shown in supplemental material A.1. □

While the present analysis considers a worst-case bound that is independent of the structure of the latent space $\mathcal{Z}$, it would be interesting to analyze the bound also for special choices of the joint feature map and discrete loss functions. Such an analysis was presented

in McAllester and Keshet [171], who showed asymptotic consistency of the update direction of a perception-like structured prediction algorithm.

Note that the requirements on the loss function are, in particular, fulfilled by the loss $l(t) = \max(0, t)$, which is employed both by the one-class SVM and by the proposed hidden Markov anomaly detector that is introduced in Section 5.3 below. Indeed in that case, l is non-negative and Lipschitz continuous with constant $L = 1$.

Hidden Markov Anomaly Detection In this section, we derive the proposed *hidden Markov anomaly detection* (HMAD) methodology that is capable of dealing with sequence data that exhibits latent state structure. We therefore need to settle for an appropriate loss function l and a joint feature map $\Psi(x, z)$.

Setting $l(t) := \max(0, t)$, we can derive a latent version of the one-class support vector machine (OC-SVM) [7]. Contrary to [172], structures need not to be known. We derive the latent version of the OC-SVM as follows.

Problem 9 (Primal latent OC-SVM optimization problem)**.** *Given the monotonically non-decreasing hinge loss function $l : \mathbb{R} \to \mathbb{R}$, $l(t) = \max(0, t)$, minimize, with respect to $\mathbf{w} \in \mathcal{H}$ and $\rho \in \mathbb{R}$,*

$$\frac{1}{2}\|\mathbf{w}\|^2 - \rho + \frac{1}{\nu n}\sum_{i=1}^{n} \max\Big(0, \rho - \max_{z \in \mathcal{Z}}\big(\langle \mathbf{w}, \Psi(x_i, z)\rangle + \delta(z)\big)\Big)\,. \tag{P′}$$

It is easy to check that the dual loss of $l(t) = \max(0, t)$ is the function $l^*(t) = 0$ if $0 \leq t \leq 1$ and ∞ else, and thus the corresponding dual optimization problem is as follows.

Problem 10 (Dual latent one-class SVM optimization problem)**.** *Given the monotonically non-decreasing hinge loss function $l : \mathbb{R} \to \mathbb{R}$, $l(t) = \max(0, t)$, and denoting by $l^* : \mathbb{R} \to \mathbb{R}$ the dual hinge loss function, maximize, with respect to $\alpha \in \mathbb{R}^n$ and subject to $0 \leq \alpha \leq \frac{1}{\nu n}$ and $\sum_{i=1}^n \alpha_i = 1$,*

$$-\min_{\substack{z_i \in \mathcal{Z} \\ i = 1, \ldots, n}} \left(\frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j k\big((x_i, z_i^\alpha), (x_j, z_j^\alpha)\big) - \sum_{i=1}^{n} \alpha_i \delta(z_i^\alpha)\right) \tag{D′}$$

In hidden Markov anomaly detection, we are interested in inferring the hidden state sequence $z = (z^1, \ldots, z^T) \in \mathcal{Z}$, with single entries $z^t \in \mathcal{Y}$, associated with an observed feature sequence $x = (x^1, \ldots, x^T)$, i.e., each element of the sequence is a feature vector $x^t = (x_l^t)_{l=1,\ldots,d} \in \mathbb{R}^d$. Hidden Markov models have been introduced as a certain class of probability density functions P with chain-like factorization [161] and parameters $\mathbf{w}$:

$$P(x, z|\mathbf{w}) = \pi(x^1, z^1|\mathbf{w}) \prod_{t=2}^{T} \big(P(z^t|z^{t-1}, \mathbf{w}) P(x^t|z^t, \mathbf{w})\big)\,. \tag{5.11}$$

Based on the corresponding log-probability and conditioned on the inputs, $\log P(z|x) = \log \pi(z^1, x^1|\mathbf{w}) + \sum_{t=2}^{T} \log P(z^t|z^{t-1}, \mathbf{w}) + \log P(z^t|x^t, \mathbf{w})$, we introduce the matching scoring function $G\colon \mathcal{X} \times \mathcal{Z} \times \mathcal{H} \to \mathbb{R}$ that decomposes into $G^{\text{trans}}\colon \mathcal{Y} \times \mathcal{Y} \times \mathcal{H} \to \mathbb{R}$ and $G^{\text{em}}\colon \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \to \mathbb{R}$:

$$\log P(z|x) = G(x, z, \mathbf{w}) = \sum_{t=2}^{T} G^{\text{trans}}(z^t, z^{t-1}, \mathbf{w}) + \sum_{t=1}^{T} G^{\text{em}}(x^t, z^t, \mathbf{w}), \tag{5.12}$$

such that $G(x, z, \mathbf{w}) \propto \langle \mathbf{w}, \Psi(x, z)\rangle$. This motivates defining a joint feature map as follows:

Definition 8 (Hidden Markov joint feature map)**.** *Given a feature map $\phi : \mathcal{X} \to \mathcal{F}$, define the* Hidden Markov joint feature map $\Psi : \mathcal{X} \times \mathcal{Z} \to \mathcal{H}$ *as*

$$\Psi(x, z) = \begin{pmatrix} (\sum_{t=2}^{T} \mathbf{1}[z^t = i \wedge z^{t-1} = j])_{i,j \in \mathcal{Y}}, \\ (\sum_{t=1}^{T} \mathbf{1}[z^t = i]\, \phi(x^t))_{i \in \mathcal{Y}} \end{pmatrix}.$$

To better understand the above feature map, observe that the weight vector $\mathbf{w}$, which is $\mathbf{w} = (\mathbf{w}^{\text{em}}, \mathbf{w}^{\text{trans}})$, decomposes into a transition vector $\mathbf{w}^{\text{trans}} = (\mathbf{w}^{\text{trans}}_{i,j})_{i,j \in \mathcal{Y}}$ and an emission vector $\mathbf{w}^{\text{em}} = (\mathbf{w}^{\text{em}}_i)_{i \in \mathcal{Y}}$, so the linear model becomes

$$\langle \mathbf{w}, \Psi(x, z) \rangle = \sum_{t=2}^{T} \sum_{i,j \in \mathcal{Y}} \mathbf{1}[z^t = i \ \wedge \ z^{t-1} = j]\, \mathbf{w}^{\text{trans}}_{i,j} + \sum_{t=1}^{T} \sum_{i \in \mathcal{Y}} \mathbf{1}[z^t = i] \langle \mathbf{w}^{\text{em}}_i, \phi(x^t) \rangle \ ,$$

which is reminiscent of the log probability associated with HMMs and given by (5.12).

Definition 9 (Hidden Markov anomaly detection (HMAD))**.** Hidden Markov anomaly detection (HMAD) *is defined as the latent OC-SVM (Problem 9 and 10) together with the hidden Markov joint feature map (Definition 8).*

Note that thus, because of the specific form of the joint feature map occuring in HMAD, the problem of maximizing over the latent variables in Eqn. (P′) can be solved by finding the most probable state sequence of the corresponding hidden Markov model, which can be efficiently computed using, e.g., Viterbi's algorithm [161].

Similar to its non-structured counterpart, the structured one-class SVM enjoys interesting properties, as we show below. Recall that for an input x and prediction function f the following cases can occur:

1. $f(x) > 0$ (then x is strictly inside the density level set)
2. $f(x) = 0$ (then x is right at the boundary of the set)
3. $f(x) < 0$ (then x is outside of the density level set, i.e., x is an outlier)

The following theorem shows that the parameter ν controls the number of outliers.

Theorem 11. *The following statements hold for the structured one-class SVM and the induced decision function f:*

(a) The fraction of outliers (inputs x_i with $f(x_i) < 0$) is upper bounded by ν.

(b) The fraction of inputs lying strictly inside the density level set (inputs x_i with $f(x_i) > 0$) is upper bounded by $1 - \nu$.

The theorem is proven in Appendix A.2 and shows that the quantity ν can be interpreted as the fraction of outliers predicted by the learning algorithm. In particular this shows, together with theoretical analysis, that for well behaved problems (where there is no probability mass exactly on the decision region and where the true decision boundary is contained in the hypothesis set, e.g., via the use of universal kernels [173]), the estimated density level set $\hat{L}_\nu$ asymptotically equals the truly underlying density level set L_ν: $P(\hat{L}_\nu \backslash L_\nu \cup L_\nu \backslash \hat{L}_\nu) \to 0$ for $n \to \infty$.

Optimization Algorithm A first difficulty occurring when trying to solve the optimization problem (P′) consists in the function $g : (\mathbf{w}, \rho) \mapsto \rho - \max_{z\in\mathcal{Z}} \big(\langle \mathbf{w}, \Psi(x_i, z)\rangle + \delta(z)\big)$, which is concave and thus renders the optimization problem non-convex. However, note that any concave function $h : \mathbb{R} \to \mathbb{R}$ can be decomposed into convex and concave parts, $\max(0, h(x)) = \max(0, -h(x)) + h(x)$. Hence, putting $g(\mathbf{w}, \rho) = \rho - \max_{z\in\mathcal{Z}} \big(\langle \mathbf{w}, \Psi(x_i, z)\rangle + \delta(z)\big)$, we can write Eq. (P′) $= \frac{1}{2}\|\mathbf{w}\|^2 - \rho + \frac{1}{\nu n}\sum_{i=1}^{n} \Big(\max\big(0, -g(\mathbf{w}, \rho)\big) + g(\mathbf{w}, \rho)\Big)$. The above decomposition consists of a convex term followed by a concave term, which admits the optimization framework of DC programming (difference of convex functions) [46]. Although the function $-g$ is not differentiable, it admits, at any point $(\mathbf{w}_0, \rho_0) \in \mathcal{H} \times \mathbb{R}$, a subdifferential

$$\begin{aligned}\partial_{(\mathbf{w}_0,\rho_0)} g(\mathbf{w}_0, \rho_0) := \{\mathbf{v} \in \mathcal{H} \times \mathbb{R} : g(\mathbf{w}, \rho) - g(\mathbf{w}_0, \rho_0) \\ \geq \langle \mathbf{v}, (\mathbf{w}, \rho) - (\mathbf{w}_0, \rho_0)\rangle , \forall (\mathbf{w}, \rho) \in \mathcal{H} \times \mathbb{R}\} .\end{aligned}$$

One can verify—using the sub-differentiability of the maximum operator—that, for any $z \in \mathcal{Z}$, the point $(\Psi(x_i, z), -1)$ is contained in the subdifferential $\partial_{(\mathbf{w}_0,\rho_0)} g(\mathbf{w}_0, \rho_0)$. Thus, we can linearly approximate, for any $z \in \mathcal{Z}$, via $g(x) \approx \langle \mathbf{w}, \Psi(x_i, z)\rangle + \delta(z) - \rho$. In the optimization algorithm we will thus construct a sequence of variables $(\mathbf{w}^t, \rho^t, z^t)$, $t = 1, 2, 3, \ldots$, where we use this approximation with z chosen as $z^t = \operatorname{argmax}_{z\in\mathcal{Z}} \big\langle \mathbf{w}^{t-1}, \Psi(x_i, z)\big\rangle + \delta(z)$, where $\mathbf{w}^{t-1}$ is conveniently computed by solving a regular one-class SVM problem. The resulting optimization algorithm is described in Algorithm 7.

Algorithm 7 Hidden Markov Anomaly Detection

input data $x_1, \ldots, x_n$
put $t = 0$ and initialize $\mathbf{w}^t$ (e.g., randomly)
repeat
  t:=t+1
  for $i = 1, \ldots, n$ **do**
    $z_i^t := \operatorname{argmax}_{z\in\mathcal{Z}} \langle \mathbf{w}^{t-1}, \Psi(\mathbf{x}_i, z)\rangle + \delta(z)$
    (i.e. use Viterbi algorithm)
  end for
  let $(\mathbf{w}^t, \rho^t)$ be the optimal arguments when solving one-class SVM with $\phi(\mathbf{x}_i) := \Psi(\mathbf{x}_i, z_i^t)$
until $\forall\, i = 1, \ldots, n : \; z_i^t = z_i^{t-1}$
Return optimal model parameters $\mathbf{w} := \mathbf{w}^t$, $\rho = \rho^t$, and $z_i := z_i^t \quad \forall\, i = 1, \ldots, N$

Despite the non-convex nature of the optimization problem, we found in our experiments that the algorithm tends to converge often faster than the standard column-generation approach of the supervised structured SVM [157], since no storage of constraints is necessary, which in turn leads to constant time and space complexity for each iteration of Algorithm 7.

5.4 Evaluation and Applications

In the following section, we apply the derived methods to applications from computational biology. First, supervised large-scale structured output methods are applied to the transcript identification problem using RNA-seq data on real eucaryotic model organism and compared against state-of-the-art techniques. Unsupervised collective anomaly detection is then evaluated on controlled data and on procarytic gene detection.

5.4.1 Transcript Identification for Eucaryotic Organisms

High-throughput sequencing technology applied to cellular mRNA (RNA-Seq) has revolutionized transcriptome studies [174–176] (among many others). In contrast to microarray platforms, which it has replaced in many applications, RNA-Seq can not only be used to accurately quantify known transcripts, but also to reveal the precise structure of transcripts at single-nucleotide resolution. RNA-Seq based transcript reconstruction has therefore become a valuable tool for the completion of genome annotations [177] (for instance) and further enabled subsequent analyses of differentially expressed genes [178], transcript isoforms [179, 180] and exons [181], all of which generally rely on correctly inferred transcript inventories. De novo transcript reconstruction is thus a pivotal step in the analysis of RNA-Seq data.

There are two conceptually different strategies to approach this problem: one can either assemble transcripts directly from RNA-Seq reads using methodology that originated from genome assembly approaches [182, 183]. Alternatively, the problem can be decomposed into two steps: RNA-Seq reads are first aligned to the genome of origin followed by the actual transcript reconstruction on the basis of these alignments. While the first, assembly-based strategy does not require a high-quality genome sequence and is thus applicable to non-model organisms, it is arguably addressing a more difficult problem than the latter, mapping-based approach. Consequently, transcripts, in particular ones with low expression, may be more accurately reconstructed by methods implementing the mapping-based approach [184, 185] (see also [182, 183] for a comparison). The performance of mapping-based methods however strongly depends on the quality of the RNA-Seq read alignments. Considerable attention has therefore been payed to solve the problem of correctly aligning RNA fragments across splice junctions [186, 187].

Following the mapping-based paradigm, we developed a novel machine learning-based method, which we call mTim: **m**argin-based **t**ranscript **i**nference **m**ethod. In contrast to algorithmic transcript assembly [184, 185], we formalize the problem as a supervised label sequence learning task and apply state-of-the-art techniques, namely Hidden Markov support vector machines (HM-SVMs) [157]. This way of approaching the problem is similar to recently developed gene finders [188], and mTim is indeed a hybrid method that can utilize both, RNA-Seq read alignments and characteristic features of the genome sequence, e.g. around splice sites [189]. However, mTim's emphasis is on inference from aligned RNA-Seq reads, and its model is only augmented by a few genic sequence motif sensors [188], which can moreover be disabled. We thus make weak assumptions, if any, about the inferred transcripts: importantly, we do not model protein-coding sequences (CDS) and are thus able to predict noncoding transcripts as well as coding ones with similar expression.

The task of reconstructing the exon-intron structure of expressed genes can be converted into a label sequence learning problem, where we attempt to label each nucleotide in the genome as either intergenic, exonic or intronic. Our prior knowledge about what constitutes a valid gene structure is incorporated into a state model to restrict the space of possible labelings to valid ones.

Starting from a naive state model that would consist of a single state for each of the atomic labels, exonic, intronic, and intergenic, we extended it as follows (see Figure 5.3): first, we devised a strand-specific model. Second, we created expression-dependent submodels. This allows us to maintain several parameter sets, each of which is optimized for transcripts with a certain read support. Due to non-uniform read coverage along transcripts, transitions between expression levels proved useful in practice. Finally, the simple model was extended by states that mark segment boundaries (e.g. when transitioning from exon to intron), as this facilitates boundary recognition from features such as spliced reads (Fig. 5.3).

Feature Derivation The inference of transcript structures is based on sequences of observations or features derived from RNA-Seq read alignments and predicted splice sites. Specifically, we derive the following position-wise features from RNA-Seq alignments:

- number of reads aligned at the given position, indicating an exon.
- a gradient of the read coverage; high absolute values correspond to sharp in- or decreases in coverage typical of the start and end of exonic regions, respectively
- number of reads that are spliced over the given position (strand-specific), thus indicating an intronic position.
- number of spliced reads supporting a donor splice site at the given position (strand-specific).
- number of spliced reads supporting an acceptor splice site at the given position (strand-specific).
- number of paired-read alignments for which the insert spanned the given position (only used if read pair information is available, strand-specific), an indicator of transcript connectivity.

Additionally, we derive features from the genome sequence around a given position such as strand-specific donor and acceptor splice site prediction.

As a ground truth for guiding the supervised training process, annotated gene models with a portion of the surrounding intergenic region are excised and converted into label sequences by assigning one of the above atomic labels to each nucleotide (see color coding in Figure 5.3). In the presence of alternative transcripts, this labeling was based on a single isoform (the one that was best supported by RNA-Seq reads), and additionally a mask of alternative transcript regions was generated to avoid that learning the correct alternatives is penalized during training.

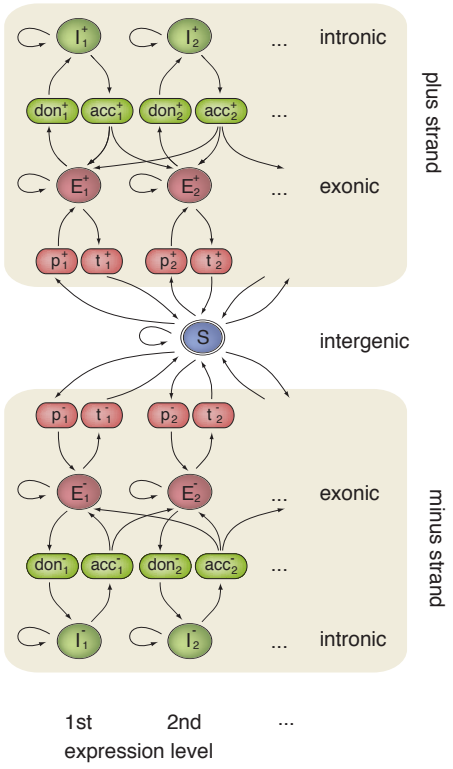

Figure 5.3 – State model used by mTim. The first and last nucleotide of introns and transcripts were modeled with particular care: The former are associated with splice site signals at exon-intron junctions (states denoted `acc` and `don`), whereas the latter correspond to transcript start and end (denoted `p` and `t`, respectively). The model is strand-specific and consists of expression-specific sub-models.

Data Preparation For the following computational experiments we used RNA-Seq data from well-studied model organism for which high-quality annotations exist, because these can not only be used for training, but also to assess the accuracy of the inferred transcripts.

We aligned RNA-Seq reads to the genome using the splice-aware alignment tool PalMapper [187]

Primary RNA-Seq alignments were filtered with the goal to reduce the number of alignment errors. To this end, we used a small subset of annotated introns to define an optimal choice of parameters for filtering criteria such as maximal number of edit operations

(mismatches, insertions, and deletions), minimal length of the shortest aligned segment in a spliced alignment, and the minimal number of alignments supporting an intron. The chosen filter settings maximize the F-Score (harmonic mean of precision and recall) between the annotation set and the introns contained in the filtered alignments.

Donor and acceptor splice sites were predicted from the genome sequence following a published protocol [189]. In summary, this method cuts out genomic sequences around all potential splice donor and acceptor site (exhibiting the two-nucleotide consensus sequence) and applies SVM classifiers with string kernels to recognize annotated splice sites. Trained classifiers are subsequently used to generate whole-genome predictions which were subsequently transformed into probabilistic confidence values [189].

From the RNA-seq read alignments we then generated the above-listed coverage and splice-site features and derived a label sequence from the corresponding gene annotations (see above for details).

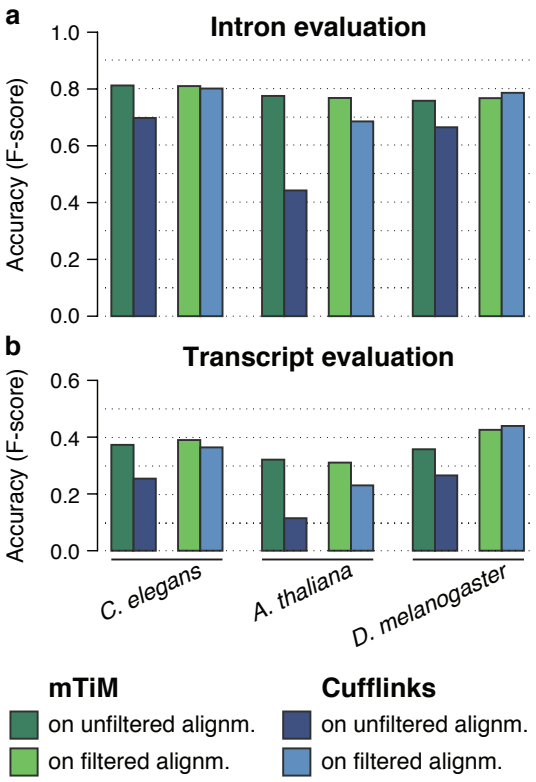

Figure 5.4 – Comparison of mTim and Cufflinks. **(a)** Assessment of the total number of introns whose boundaries were correctly predicted at single-nucleotide precision. **(b)** Evaluation of the number of gene loci for which at least one transcript isoform was predicted correctly (all introns correct).

To be able to assess the impact of alignment quality on subsequent transcript inference, we used unfiltered alignments in a first set of experiments and subsequently repeated these using filtered RNA-seq alignments as input to assess the improvement of transcript inference with improved alignment quality.

To generate transcript models from these read alignments, the mTim pipeline proceeds through the following steps:

1. Definition of genome chunks; importantly, chunks are defined based on read coverage only without using any annotation information.

2. Partitioning genome chunks into subsets for cross validation.

3. Training on chunks from the training set using known (annotated) gene models as ground truth.

4. Application of the trained mTim models to predict transcript structures on test chunks.

Using cross-validation, we obtain unbiased estimates of mTim's transcript reconstruction accuracy for data it had not seen during training.

To compare mTim's prediction to the state of the art in alignment-guided transcript inference, we also applied Cufflinks with default parameter settings to the same unfiltered and filtered RNA-seq alignment data.

Results and Discussion To evaluate its performance, we applied mTim to RNA-Seq data from model species. We chose three organisms, *Chaenorhabditis elegans* (nematode worm), *Arabidopsis thaliana* (thale cress) and *Drosophila melanogaster* (fruit fly), whose genomes

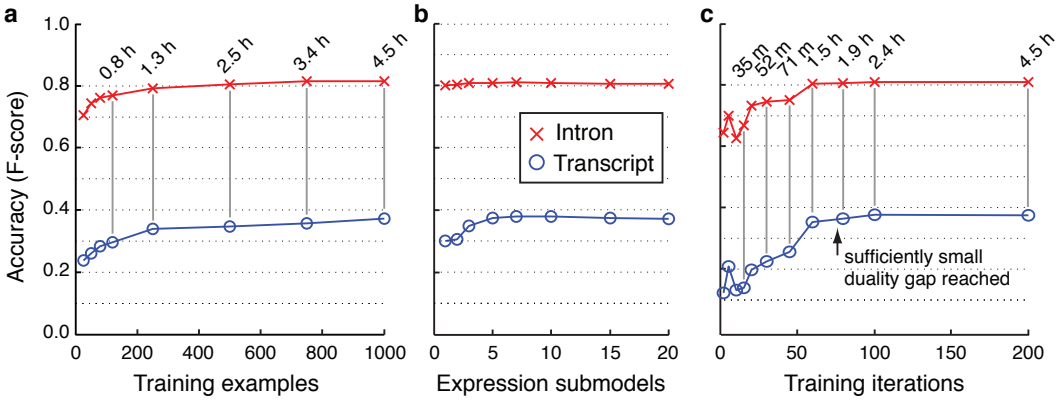

Figure 5.5 – Opimizing mTim's performance. **(a)** The HM-SVM learning algorithm utilized training data efficiently and accuracy quickly reached a plateau. **(b)** Expression-specific submodels (see Fig. 5.3 and Methods) improve reconstruction of complete transcripts. **(c)** Accuracy as a function of the number of training iterations (using 1000 examples). The duality gap was sufficiently small for termiantion after 78 iterations. All results were obtained using unfiltered RNA-Seq alignments for *C. elegans*. Empirical execution times in (a) and (c) were averaged across three HM-SVM trainings.

and transcriptomes have been extensively characterized [177], making it possible to use annotated gene models as a ground truth for evaluating the quality of transcripts reconstructed from RNA-Seq data. Although these genome annotations were neither complete nor free of errors, which only allowed for approximative evaluations, these were nonetheless useful for assessing mTim's transcript reconstruction accuracy relative to other methods.

We evaluated the accuracy of transcripts reconstructed by mTim in a whole-genome comparison to annotated protein-coding genes using cross-validation (see Methods for details). Here we used two popular criteria that evaluate intron and transcript quality respectively. The first is an assessment of the total number of introns that are inferred correctly (with single-nucleotide precision), whereas the second counts the number of gene loci for which at least one transcript isoform has been reconstructed correctly (all introns predicted correctly). Note that both criteria do not evaluate transcript starts and ends at nucleotide resolution, because annotations are generally more uncertain for these than for intron boundaries; in transcript evaluation, however, predicted transcript fusion or split predictions will be regarded as errors.

For both criteria we assessed the sensitivity and precision of predicted transcripts. The former is defined as the proportion of annotated introns (or transcripts) which were inferred correctly, whereas the latter is defined as the proportion of inferred introns (or transcripts) which correctly matched an annotated intron (or transcript). The F-score is an aggregate accuracy measure, defined as the harmonic mean of sensitivity and precision:

$$F = 2 \cdot \frac{\text{sensitivity} \cdot \text{precision}}{\text{sensitivity} + \text{precision}}$$

In initial assessments we verified the effectiveness of mTim's training algorithm and modeling approach. We first evaluated how efficiently the HM-SVM training exploits the available training data. Intron accuracy quickly reached a level where additional training sequences no longer led to substantial improvements: with as little as 80 training examples an intron accuracy (F-score) of 0.75 was exceeded, which was only 6.5% below the maximum of 0.812 (Fig. 5.5a). Transcript reconstruction accuracy continued to improve with additional training examples, although with 250 training sequences transcript accuracy was less than 10% below the maximum of 0.373 (Fig. 5.5a). Second, we assessed the impact of expression-specific submodels (see Fig. 5.3 and Methods) on transcript reconstruction accuracy (Fig. 5.5b). While we

observed little effect on intron reconstruction, we confirmed that submodels were valuable for correctly inferring whole transcripts: with five submodels, transcript accuracy increased by 25% relative to the simple model without submodels (Fig. 5.5b). Since expression-specific submodels provided an effective means to group exons with similar expression levels into one transcript and terminate it when expression changes dramatically, we used five submodels for all subsequent mTim experiments. Third, we assessed convergence speed of mTim's optimization approach. Results obtained for a training set consisting of $1,000$ sequences suggest that after about 80 iterations, completed in < 2 CPU hours, prediction accuracy had converged (Fig. 5.5c).

To benchmark mTim's transcript reconstruction performance in comparison to other methods, we extended our evaluations to include Cufflinks [184], a widely adopted method, applying the same assessment criteria as before. Comparative evaluations revealed that mTim inferred relatively accurate transcript structures, almost always as good as or better than Cufflinks (Fig. 5.4). Notably, mTim's predictions were relatively robust against issues in the underlying read alignments (intron accuracy was unaffected by alignment filtering, and transcript F-score decreased by at most 16%). Cufflinks in contrast was found to be much more sensitive to these issues; without alignment filtering, its intron and transcript accuracy (F-score) dropped by $13 - 35\%$ and $30 - 50\%$, respectively (Fig. 5.4). The quality of transcripts inferred by mTim appeared to be relatively high (Fig. 5.4) and consistently so across the diverse range of input data tested here; in particular mTim maintained high precision (Table 5.1).

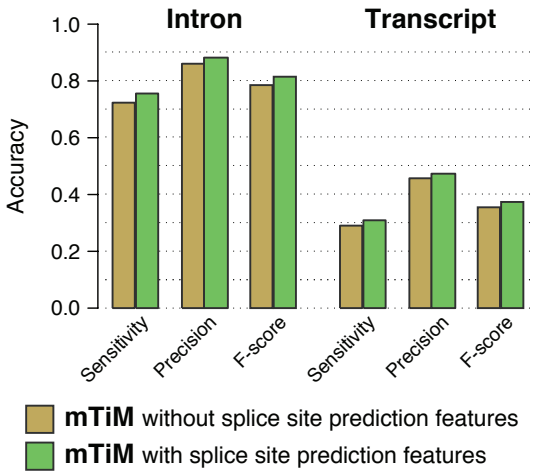

Figure 5.6 – Accuracy of mTim when trained with and without features derived from genomic sequence signals around splice sites. Both mTim instances were trained and evaluated on unfiltered RNA-Seq alignments from *C. elegans*.

Due to its modular architecture and its general machine-learning approach, mTim can easily be tailored to specific application requirements. For instance features corresponding to genomic splice site predictions can be disabled, making mTim rely completely on RNA-Seq alignment features thereby eliminating any potential bias against non-coding transcripts. We assessed the extent to which this affects transcript reconstruction accuracy and found the effect to be minor (Fig. 5.6).

Application Outcome Here, we have introduced mTim, a discriminative machine learning-based method that reconstructs transcripts from RNA-Seq read alignments and splice site predictions. We have shown that it is able to infer transcripts with high accuracy and that it is more robust errors in the underlying read alignments. Pre-trained mTim predictors used for this work are available within the Oqtans Galaxy webserver [2]. Moreover, mTim is open-source software provided via GitHub [3].

[2] http://oqtans.org/
[3] https://github.com/nicococo/mTIM

Table 5.1 – Sensitivity and precision of introns and transcripts reconstructed with mTim or Cufflinks applied to PalMapper alignments.

	Alignm. filtered	**Sensitivity [%]** mTim	Cufflinks	**Precision [%]** mTim	Cufflinks
			Intron evaluation		
C. elegans	NO	**75.4**	58.6	**88.1**	86.4
	YES	**74.0**	71.3	89.5	**91.6**
A. thaliana	NO	**69.4**	30.9	**87.8**	77.9
	YES	**69.1**	53.5	86.5	**95.6**
D. melanogaster	NO	**70.5**	66.6	**82.1**	66.5
	YES	68.1	**70.7**	88.0	**88.6**
			Transcript evaluation		
C. elegans	NO	**30.8**	20.3	**47.3**	33.8
	YES	**31.5**	30.4	**51.2**	45.2
A. thaliana	NO	**24.6**	8.6	**46.2**	17.0
	YES	**23.9**	21.2	**44.2**	25.2
D. melanogaster	NO	**28.0**	24.7	**49.4**	28.7
	YES	32.1	**34.7**	**63.0**	59.8

Accuracy values of the best-performing method in each category are in bold face. See main text for definitions of sensitivity and precision and details on alignment filtering.

5.4.2 Hidden Markov Anomaly Detection

We conducted experiments for the scenario of label sequence learning where we have full access to the ground truth as well as a real-world computational biology scenario. Our interest is to assess the anomaly *detection* performance of our hidden Markov anomaly detection (HMAD) method for groups of measurements. As baseline methods that excel in one-class classification settings, we chose one-class support vector machines (OC-SVM) with appropriate kernels. For initialization, we randomly choose a vector $\mathbf{w}^0$ for each run of our algorithm which is sufficient, since no initialization of structures is needed, as those are deduced from the parameter vector.

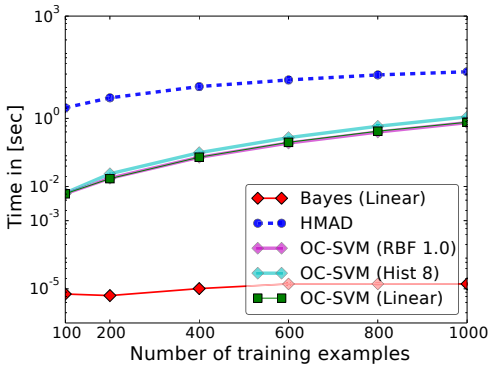

Figure 5.7 – Runtime performance for the controlled experiment. Results are shown for our hidden Markov anomaly detection (HMAD) as well as a set of competitors (using optimal kernel parameters) for an increasing number of training examples.

Controlled Experiment For the controlled experiments, we aim to gain insights into the behavior of our method. We investigate the anomaly detection performance for low to very high (up to 30%) fraction of anomalies. Furthermore, we are interested in the anomaly detection performance for an increasing amount of disorganization in the input sequences. Since HMAD exploits latent structure, it is not clear how it performs when less structure is present. Vanilla OC-SVMs does not exploit latent dependencies and should be unaffected by this. Additionally, we are interested in the runtime behavior for various training set sizes.

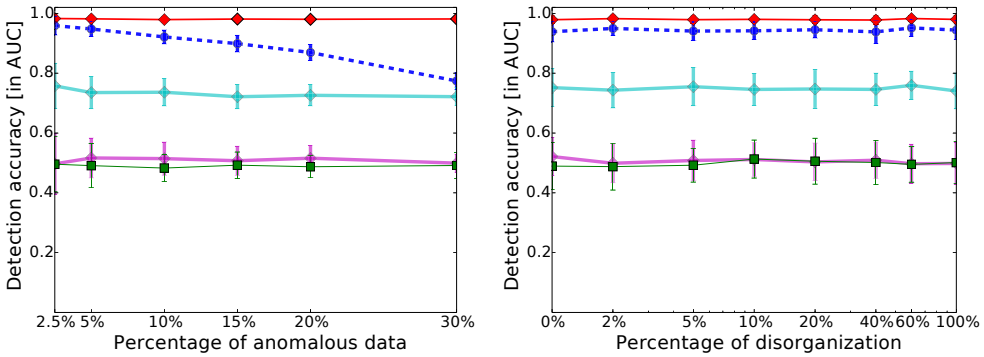

Figure 5.8 – Results for the controlled experiment: (left) anomaly detection performance for various fractions of anomalies in the training set and (right) anomaly detection performance for increasing amount of disorganization. All settings show results for our hidden Markov anomaly detection (HMAD) as well as a set of competitors (using optimal kernel parameters). Noticeable, the detection performance of HMAD is not affected by increasing amounts of disorganization in the input data (right).

We generated Gaussian noise sequences of length 600 with unit variance for the nominal bulk of the data. Non-trivial anomalies (see Fig. 5.9) were induced as blocks of Gaussian noise with non-zero mean and a total, cumulative length of 120 per anomalous example. We vary either the fraction of anomalies in the training data set or the number of blocks, depending on the amount of structure that is modeled into the data (see Figure 5.9: from 120 sub-blocks of length 1 (100% disorganization) to a single block of length 120 (0% disorganization). We employ a binary state model consisting of 2 states and 4 possible transitions with an constant prior $\delta(\cdot)$. We report on the average area under the ROC curve (AUC) for the anomaly detection performance over 50 repetitions of the experiment. Since we know the underlying ground truth we can exactly compute the Bayes classifier,[4] which in our case lies within the set of linear classifiers, and serves as a hypothetical upper performance bound for the maximal achievable detection performance.

We compare the detection performance of our method to the one achieved by OC-SVMs with RBF kernels, histogram kernels, and linear kernels using l_1- and l_2-feature normalization, and optimal kernel parameters (1.0 for the RBF kernel, 8 for the histogram kernel, and l_1 for the linear kernel). The results of the anomaly detection experiment are shown in Figure 5.8 and Figure 5.7. As can be seen in the figure, our method achieves tremendously higher detection rates than the OC-SVMs using linear or RBF kernel, which perform similar bad as random guessing. Most competitive baseline methods are OC-SVMs with histogram kernels and optimal bin size (8 bins). There exists a strong relation between our method HMAD and Fisher kernels [190] in the sense, that the same representation is used. Unlike Fisher kernels, our methodology includes the parameter optimization procedure, and therefore, given the same model parameters both

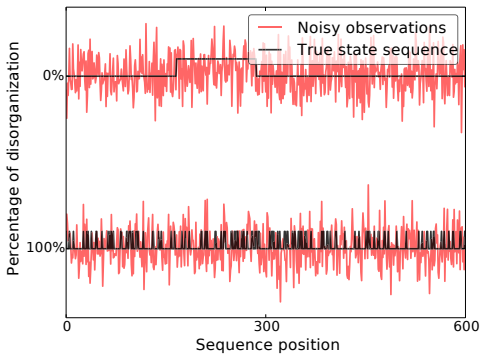

Figure 5.9 – Examples of observation sequences for two extreme cases of our controlled experiments: even in the easy setting (top), the true state sequence is barely visible to the naked eye in the noisy observed sequence, while in the challenging setting (bottom) it is almost impossible for humans to extrapolate the truly underlying state sequence.

[4] For data that is i.i.d. realized from a distribution (which is the case in our synthetic experiment), the Bayes classifier is defined as the classifier achieving the maximal accuracy among all measurable functions.

methods are on par. Remarkably, our method achieves stable on-par performance with the Bayes classifier for all levels of disorganization, even when there is no structure to be exploited in the data (see Figure 5.8 right) and outperforms significantly all competitors for varying fraction of anomalies (see Figure 5.8 left).

As an example, we depict two typical anomalous observation sequences of length 600 and anomalous block length 120 of the experiment in Fig. 5.9 for the 0% (top) and 100% disorganization (bottom) settings. As can be seen, anomalies are not trivially detectable. We also conducted runtime experiments (Fig. 5.8 right) to compare the runtime of our method HMAD against that of the baseline methods. We used the same two-state model as in the previous controlled experiment, but with training set size varying from 100 to 1000 examples. We used a fraction 10% of anomalies to ensure there is a sufficient number of anomalies in the data. As expected, absolute computational runtime is higher than for vanilla OC-SVMs. This is due to the iterative approach that includes Viterbi decoding of the sequences and solving a vanilla OC-SVM in each step. However, computational complexity grows with increasing number of examples comparable to OC-SVM which gives a total complexity of $\mathcal{O}(\text{OC-SVM}) + \mathcal{O}(c)$, where c is a constant.

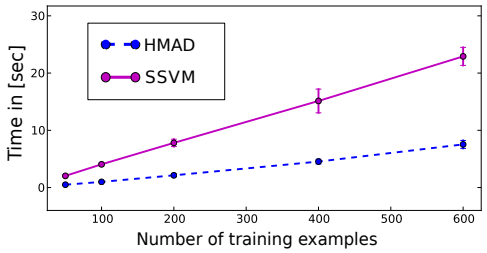

Figure 5.10 – Without the need for constraint generation, our hidden Markov anomaly detection easily outperforms the structured SVM.

We report run time comparisons of the structured output SVM (SSVM) and our hidden Markov anomaly detection in Figure 5.10. Since the HMAD does not need to add constraints in each iteration, it easily outperforms the SSVM. However, it does require multiple iterations that include Viterbi decoding as well as solving a vanilla one-class SVM and therefore is slower than the OC-SVM (for a comparison see Fig. 5.8).

Fisher kernels [190–192] have been proposed as a way of incorporating graphical models into the framework of kernel-based learning [33] and therefore benefit from the vast amount of kernel machines. A practical Fisher kernel is defined as the gradient of the log-likelihood of the probabilistic model with respect to its model parameters.

There is a strong connection of Fisher kernels and our HMAD, in the sense, that we use the same representation of graphical models. However, our method HMAD includes the parameter optimization procedure. Specifically, given the same model parameters learned by our method, the corresponding Fisher kernel employed in an one-class SVM leads to the same solution. Of course, learning the right model parameter is the key to good performance.

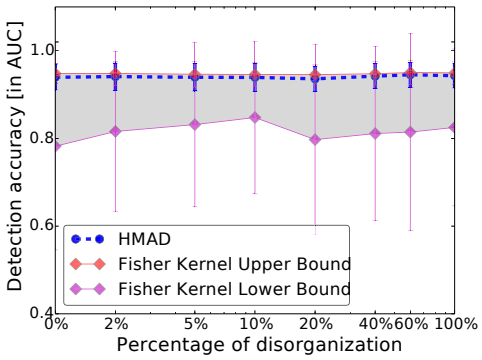

Figure 5.11 – Comparison for an increasing amount of disorganization of our method HMAD (blue) against a variety of Fisher kernels (gray area), including a lower bound (magenta) based on random model parameters and an upper bound (red) that was trained on *ground truth* data.

To cope with a variety of parameter learning settings and hence, have a realistic comparison against multiple parameter estimation methodologies for Fisher kernels and derive an upper and a lower bound for the maximum likelihood estimation for Fisher kernels. Here, a

lower bound can be easily obtained by using random model parameters, whereas an upper bound uses the *ground truth* latent states information for parameter estimation.

The results in Fig. 5.11 and Fig. 5.12 show the range of possible solutions for the Fisher kernel (gray area) with the upper bound (red) and (unsurprisingly unstable) lower bound (magenta), in the same setting as in Section 5.4.2. Moreover, it shows that our method HMAD performs nearly as good as the upper bound in *absence* of any label information.

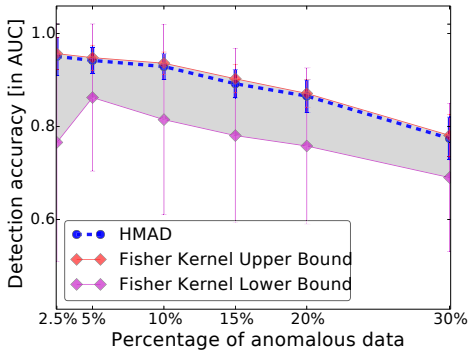

Figure 5.12 – Comparison for an increasing amount of anomalies of our method HMAD (blue) against a variety of Fisher kernels (gray area), including a lower bound (magenta) based on random model parameters and an upper bound (red) that was trained on *ground truth* data.

To assess the stability of the found solution, we did experiments with an increasing number of hidden states for our proposed method HMAD in the same setting as in Section 5.4.2. The results in Fig. 5.13 show, that our method is not sensible to the number of hidden states.

Procaryotic Gene Detection In prokaryotes (mostly bacteria and archaea) gene structures consist of the protein coding region that starts by a start codon (one out of three specific 3-mers in many prokaryotes) followed by a number of codon triplets (of three nucleotides each) and is terminated by a stop codon (one out of five specific 3-mers in many prokaryotes) [193]. Genic regions are first transcribed to RNA and then translated into a protein. Since genes are separated from one another by intergenic regions, the problem of identifying genes can be posed as a label sequence learning task, were one assigns a label (out of intergenic, start, stop, exonic) to each position in the genome [188].

We downloaded the genome of the widely studied *escherichia coli* bacteria, which is publicly available.[5]

Genomic sequences were cut between neighboring genes (splitting intergenic regions equally), such that a minimum distance of 6 nucleotides between genes was maintained. Intergenic regions have a minimum distance of 50 nucleotides to genic regions. Features were derived from the nucleotide sequence by transcoding it to a numerical representation of triplets. All examples have a minimum length of 500 nucleotides and do not exceed 1200 nucleotides.

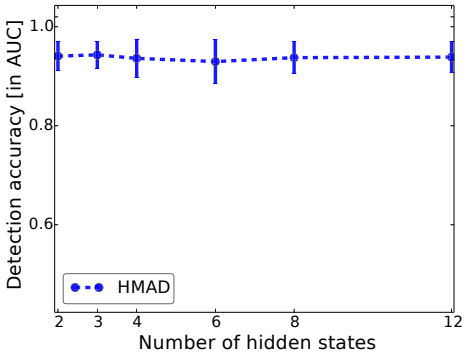

Figure 5.13 – Performance evaluation for an increasing number of hidden states of our method HMAD (blue).

For the OC-SVM we use matching spectrum kernels of order 1,2, and 3 (resp. 64, 4160, and 266.304 dimensions), while the SSVM and HMAD obtain a sequence of binary entries as input data. A description of the used state model, which is based on Görnitz, Widmer, Zeller, Kahles, Sonnenburg, and Rätsch [26], is given in Figure 5.15. Start and stop states use corresponding features that encode start and stop codons. Any other states is using all 64 binary input features. Furthermore, we choose

[5] `http://www.sanger.ac.uk. . .`
`. . . /resources/downloads/bacteria/escherichia-coli.html`

$\delta(z)$ to have a slightly higher probability towards the intergenic state. For a more fair comparison, OC-SVM and HMAD are given the true fraction of anomalies which varies from 2.5% up to 30%. The training set contained 200 examples of intergenic and genic examples with a total length of >170.000 nucleotides, while the testing set contained 350 intergenic and 50 genic examples of length >330.000 nucleotides, rending this a computationally challenging experiment. The experiment was repeated 20 times where training and test set are drawn randomly.

We further employ a simple feature selection procedure where the 8 most distinctive genic- and intergenic features are selected on a comparable labeled procaryote (*e. fergusonii*), which increased performance for OC-SVM by more than 10%. While performance for our HMAD remained unchanged, training and prediction times dropped down to 15% when compared to the full model.

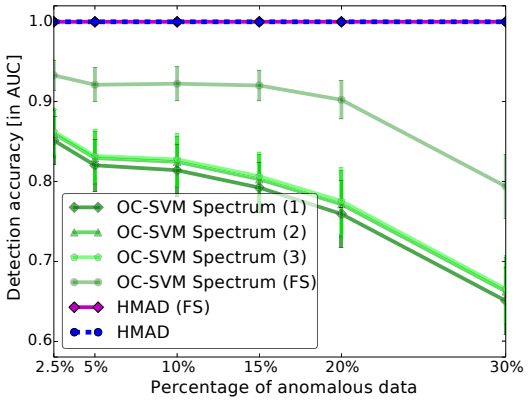

Figure 5.14 – Detection performance for various fractions of outliers in terms of AUC for the procaryotic gene finding experiment. Clearly, the accuracy of our hidden Markov anomaly detection exceeds the vanilla one-class SVM performance even when using higher order (1,2 & 3 codons = 64, 4160 and 266.304 dimensions) spectrum kernels.

The results in Figure 5.14 show a vastly superior performance of our method (HMAD) in terms of the detection accuracy: HMAD achieves a perfect AUC of 1.00 (which means: it exactly identifies every sequence containing a gene with zero error) for all outlier fractions, while the classical one-class SVM shows much worse performance with an AUC of 0.85 at best and 0.66 in the worst case. Using higher order spectrum kernels increases the detection performance only marginally.T h is result is remarkable as it has been reported that string kernels such as spectrum kernel achieve state of the art performance in this application [188].

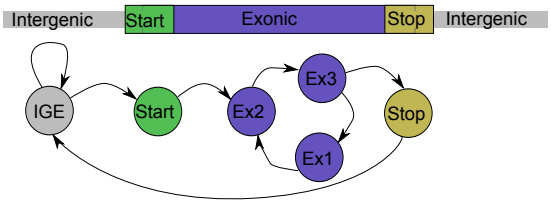

Figure 5.15 – State model of prokayotic gene finding.

Application Outcome Here, we investigated various properties of our proposed method, hidden Markov anomaly detection (HMAD), on a artificially generated data set where we had full control over the ground truth. We further applied our method to real data from procaryotic bacteria and showed that unsupervised detection of genes is possible. In both settings, HMAD performed comparable or better than baseline competitors. Moreover, as no column generation is necessary, the runtime performance is much better than its supervised counterpart.

5.5 Summary and Discussion

The main contributions of this chapter are twofold. In the first part, we took a closer look at the supervised structured output SVM (SSVM) which serves as a foundation for the latter introduced unsupervised anomaly detection method. We re-formulated the margin rescaling variant of the SSVM and incorporate a more complex regularization that enables hierarchical multitask learning. We then derived an optimization method based on bundle methods to

solve the specific problem of SSVM when a fair amount of samples is given and each sample consists of large numbers of measurements. The speed improvements are shown in Fig. 5.2.

In Section 5.3, we derived an anomaly detection method that is based on the supervised structured output principle but does not need label information. Due to its access to the latent dependency structure of a group of measurements, it can be applied as collective anomaly detection method. This is verified in the experimental section that gives positive results when compared against baseline methods.

Limits of Bundle Method Optimization for SSVMs The proposed optimization scheme was designed with the specific application of Section 5.4.1 in mind. This setting includes a low to medium number of samples with a low to medium number of features but where each sample possibly contains thousands and even millions of measurements. On the other hand, when this setting is not given (i.e. many samples with small amount of measurements each), the optimization might take much longer than the standard column generation. Moreover, due to the fine-grained label information needed to train those models, SSVMs are generally not suitable as anomaly detectors even if results tend to be very good.

Limits of Latent Structure Anomaly Detector Even though the proposed unsupervised method is much faster than its supervised (SSVM) counterpart, it still needs various iterations of a standard one-class SVM until convergence and is hence, multiple times slower than the standard formulation. Moreover, compared to the non-structured baseline, we loose the important convexity property and might get stuck in low-quality minima. The possibly biggest drawback, however, is its complexity. The kind of structure is pre-coded into a fairly complicated joint feature map and needs to be chosen carefully when deploying to a specific application. If done properly, the empirical evaluation suggests that it can be beneficial.

Source code and resources for the proposed methods are available on github [a] [b]. Parts of this chapter are based on:

> **Görnitz, N.**, Braun, M., Kloft, M., "Hidden Markov Anomaly Detection", in *International Conference on Machine Learning (ICML)*, 2015, pp. 1833–1842
>
> **Görnitz, N.**, Widmer, C., Zeller, G., Kahles, A., Sonnenburg, S., Rätsch, G., "Hierarchical Multitask Structured Output Learning for Large-scale Sequence Segmentation", in *Advances in Neural Information Processing Systems (NIPS)*, 2011, pp. 2690–2698
>
> Zeller, G., **Görnitz, N.**, Kahles, A., Behr, J., Mudrakarta, P., Sonnenburg, S., Rätsch, G., "mTim: rapid and accurate transcript reconstruction from RNA-Seq data", *ArXiv*, 2013

[a] https://github.com/nicococo/tilitools
[b] https://github.com/nicococo/mtim

IV Contextual Anomalies

Chapter 6

Learning with Latent Class Dependencies

In this chapter, we turn our attention to the contextual anomaly setting. That is, we consider samples to be singletons but with some contextual connection. In this setting, data points will be considered anomalous only if the contextual information admits it. Of course, there are many different forms of contextual information that can be examined. Here, we restrict ourselves to latent class dependencies where the anomaly score for each data points depends on the respective observation and some hidden class information.

According to Chandola, Banerjee, and Kumar [1], contextual anomaly detectors need two kinds of attributes: (i) contextual attributes that embed the corresponding data point into a neighborhood and (ii) behavioral attributes which determine its normality. Common approaches in the literature consider time-series [194] or spatial [195] contexts. In this chapter we consider latent class dependencies.

We start in Section 6.2 by extending the support vector data description (SVDD) to incorporate latent class dependencies and a mechanism to infer the bespoken latent class on-the-fly. This is done employing techniques from structured output prediction (cf. Chapter 5). In Section 6.3, we abandon the flexibility of the joint feature map and make the latent class dependency explicit which allows us to establish a—somewhat surprising—connection to k-means. The third part (Section 6.4), however, starts with a certain regression setup in mind where data points are connected to their respective latent class variables while, at the same time, the latter exhibit connections among themselves. An extension towards contextual one-class classification using our previously developed methods is discussed. Here, the boundary between contextual and collective outlier detection vanishes and this method allows to bind groups of data points.

6.1 Preliminaries

Throughout this chapter, we consider support vector data description (SVDD) as our base model. Here, the data is mapped from the input space into a RKHS feature space $\phi\colon \mathcal{X} \to \mathcal{F}$

that gives rise to a kernel k [33, 196]. The goal is to find a model $f\colon \mathcal{X} \to \mathbb{R}$ and a density level-set $L := \{\mathbf{x}\colon f(\mathbf{x}) \leq R^2\}$ containing most of the regular data points, while for anomalies and outliers $\mathbf{x} \notin L$ holds. In case of the support vector data description (SVDD) method, $f_{\text{SVDD}}(\mathbf{x}) = \|\mathbf{c} - \phi(\mathbf{x})\|^2$ and parameter estimation corresponds to solving a quadratically constrained quadratic program (QCQP),

$$\begin{aligned} &\min_{R,\mathbf{c},\xi\geq 0} \; R^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{s.t.} \quad \|\mathbf{c} - \phi(\mathbf{x}_i)\|^2 \leq R^2 + \xi_i\,, \quad i = 1,\ldots,n\,. \end{aligned} \tag{6.1}$$

That allows for the following simple geometric interpretation: a ball of radius R is computed that comprises most the regular data points, while all points lying outside of the normality radius are declared being anomalous.

An important note on Section 6.4 where techniques for with learning latent class dependency *structure* are developed: the original motivation behind this section is different than contextual anomaly detection and rather focuses on a semi-supervised regression setting. However, due to its modularity, a straightforward extension towards one-class classification based on support vector data descriptions is discussed in detail.

6.2 A Joint Feature Map Formulation

One way of incorporating behavioral $\mathbf{x} \in \mathcal{X}$ and contextual $\mathbf{z} \in \mathcal{Z}$ information blocks into standard methods is by employing *joint feature maps* $\Psi : \mathcal{X} \times \mathcal{Z} \to \mathcal{F}$ to encode the information and embed it into a feature space which then can be plugged in without further changes of the method. As straightforward as it seems, we did not specify the structure of Ψ yet and, most importantly, do not assume that the context $\mathbf{z}$ is given in advanced and instead must be inferred based on the observations $\mathbf{x}$. Hence, the contributions of this section are:

- we extend the support vector data description to handle contextual information based on the notion of joint feature maps
- we present a simple embedding for the problem of latent classes
- a corresponding solver based on difference of convex (DC) functions programming is proposed.

In this section, we extend the classical mapping f_{SVDD} by the inclusion of a latent variable $\mathbf{z} \in \mathcal{Z}$ in a joint feature map $\Psi : \mathcal{X} \times \mathcal{Z} \to \mathcal{F}$. As we try to find the tightest description of the data, it makes sense to define the contextual information to correspond to finding a minimizer of the hyper-sphere description. As a consequence, the resulting model

$$f : \mathcal{X} \to \mathbb{R},\ \mathbf{x} \mapsto \min_{\mathbf{z}\in\mathcal{Z}} \|\mathbf{c} - \Psi(\mathbf{x},\mathbf{z})\|^2 \tag{6.2}$$

becomes more expressive (a similar idea appeared also recently in the context of supervised learning [157]). The latent state variable $\widehat{\mathbf{z}}$ of a given data point $\mathbf{x}$ can be inferred by $\widehat{\mathbf{z}} = \operatorname{argmin}_{\mathbf{z}\in\mathcal{Z}} \|\Psi(\mathbf{x},\mathbf{z})\|^2 - 2\langle \mathbf{c}, \Psi(\mathbf{x},\mathbf{z})\rangle$. The extended model, which we call LatentSVDD, leads to a modified optimization problem:

$$\begin{aligned} &\min_{R,\mathbf{c},\xi\geq 0} \; R^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{s.t.} \quad \min_{\mathbf{z}\in\mathcal{Z}} \|\mathbf{c} - \Psi(\mathbf{x}_i,\mathbf{z})\|^2 \leq R^2 + \xi_i \quad \forall\, i\,. \end{aligned} \tag{LatentSVDD}$$

Because of the $\min$ operator in the constraints, the resulting optimization problem is no longer convex, but we can derive an optimization strategy by decomposing the problem into convex and concave parts and iteratively linearizing the concave part (*DC Programming* [197]). In order to do so, we re-write the above problem in an equivalent, unconstrained fashion as follows:

$$\min_{\mathbf{c},R} \; R^2 + C\sum_{i=1}^{n} \max(0, \min_{\mathbf{z}\in\mathcal{Z}} \|\mathbf{c} - \Psi(\mathbf{x}_i, \mathbf{z})\|^2 - R^2). \tag{6.3}$$

Substituting $\Omega := R^2 - \|\mathbf{c}\|^2$, this is equivalent to

$$\min_{\mathbf{c},\Omega} \|\mathbf{c}\|^2 + \Omega + C\sum_{i=1}^{n} \max\left(0, -\Omega + \min_{\mathbf{z}\in\mathcal{Z}} \|\Psi(\mathbf{x}_i, \mathbf{z})\|^2 - 2\langle \mathbf{c}, \Psi(\mathbf{x}_i, \mathbf{z})\rangle\right)$$

subject to the constraint $\|\mathbf{c}\|^2 + \Omega \geq 0$, which can be dropped as it is not active in the optimal point. Note that, for any i, the function

$$g_i(\mathbf{c}, \Omega) := -\Omega + \min_{\mathbf{z}\in\mathcal{Z}} \|\Psi(\mathbf{x}_i, \mathbf{z})\|^2 - 2\langle \mathbf{c}, \Psi(\mathbf{x}_i, \mathbf{z})\rangle \tag{6.4}$$

is concave, so $-g_i$ is convex. Furthermore, note that for any $t \in \mathbb{R} : \max(0, t) = \max(0, -t) + t$. Thus, we have the decomposition

$$\max(0, g_i(\mathbf{c}, \Omega)) \;=\; \underbrace{\max(0, -g_i(\mathbf{c}, \Omega))}_{\text{convex}} \;+\; \underbrace{g_i(\mathbf{c}, \Omega)}_{\text{concave}},$$

because the maximum of two convex functions is convex and can equivalently re-write the LatentSVDD optimization problem as a sum of a convex and a concave function as follows: given the definition of g_i in Eq. (6.4), solve

(LatentSVDD-DC)

$$\min_{\mathbf{c},\Omega} \underbrace{\|\mathbf{c}\|^2 + \Omega + C\sum_{i=1}^{n} \max(0, -g_i(\mathbf{c}, \Omega))}_{\text{convex}} + \underbrace{C\sum_{i=1}^{n} g_i(\mathbf{c}, \Omega)}_{\text{concave}}$$

The above problem is an instance of the class of DC optimization problems. We propose to solve the above problem with the simplified DC algorithm. That is, alternatingly, the concave part is linearized and the resulting approximate problem solved. The resulting algorithm is shown in Algorithm 8.

The proposed algorithm converges against a local optimum (typically in about 10 iterations, as we found in our experiments). This follows from the following theorem that is taken from [50], which is an extension of the convergence theorem in [198] to non-differentiable objective functions.

Theorem 12 ([50], Theorem 3.3)**.** *Let f, g be convex functions. Let x_0 be any feasible point, and put*

$$\forall t > 0 : \quad x_t := \operatorname*{argmin}_x f(x) - x^\top \nabla g(x_{t-1})\,.$$

If the non-smooth parts of f and g are piecewise-linear and the smooth part of f is strictly convex quadratic, then any limit point of the sequence (x_t) is a stationary point.

Algorithm 8 Optimization Algorithm for LatentSVDD

input data $\mathbf{x}_1, \ldots, \mathbf{x}_N$
initialize $\mathbf{c}^{t=0}$ & $\forall i : \widehat{z}_i^{t=0}$ (e.g., randomly)
repeat
  t:=t+1
  for $i = 1, \ldots, N$ **do**
    $\widehat{\mathbf{z}}_i^t := \operatorname{argmin}_{\mathbf{z}\in\mathcal{Z}} ||\mathbf{c}^{t-1} - \Psi(\mathbf{x}_i, \mathbf{z})||^2$
    overwriting the notation of g_i in (6.4), we define
    $g_i(\mathbf{c}, \Omega) := -\Omega + ||\Psi(\mathbf{x}_i, \widehat{\mathbf{z}}_i^t)||^2 - 2\langle \mathbf{c}, \Psi(\mathbf{x}_i, \widehat{\mathbf{z}}^t)\rangle$
  end for
  let $\mathbf{c}^t$ and Ω^t the optimal arguments when solving Problem (LatentSVDD-DC) with the g_i set as above
until $\forall\, i :\ \widehat{\mathbf{z}}_i^t := \widehat{\mathbf{z}}_i^{t-1}$
return optimal model parameters $\mathbf{c} := \mathbf{c}^t$, $R := \sqrt{||\mathbf{c}^t||^2 + \Omega^t}$, and $\mathbf{z}_i := \widehat{\mathbf{z}}_i^t \quad \forall\, i = 1, \ldots, N$

The proposed algorithm also admits a dual representation via the convex conjugate function $f^*(x) := \sup_y \langle x, y\rangle - f(x)$. The dual of the LatentSVDD-DC problem is given by

$$\min_{\mathbf{c},\Omega} \left(-C \sum_{i=1}^{n} g_i(\mathbf{c}, \Omega) \right)^* - \left(\|\mathbf{c}\|^2 + \Omega + C \sum_{i=1}^{n} \max(0, -g_i(\mathbf{c}, \Omega)) \right)^* .$$

This completes the presentation of the first step in our proposed methodology.

Joint feature map As we have described LatentSVDD in terms of a joint feature map $\Psi(\mathbf{x}, \mathbf{z})$ we want to specifically fix the structure for problems involving latent classes dependencies. Hence, we specifically employ a variant of the joint feature map with latent space $\mathcal{Z} := \{1, \ldots, K\}$ that is similar to the multi-class joint feature map in [157]: let be $\Lambda(\mathbf{z}) = \{\delta(z_1, z), \delta(z_2, z), \ldots, \delta(z_K, z)\} \in \{0, 1\}^K$ and $\mathbf{a} \otimes \mathbf{b}$ be the direct tensor product of vectors a and b. Given a data point $\mathbf{x}$, we define our joint feature map as $\Psi(\mathbf{x}, z) = \phi(\mathbf{x}) \otimes \Lambda(z)$. Let us assume the dimensionality of ϕ is d, then the total dimensionality is $d \cdot K$ with at least $d \cdot (K - 1)$ zero entries.

Theoretical Analysis We conclude with a generalization analysis of this unsupervised learning algorithm and define for any $\lambda > 0$, the following hypothesis class

$$\mathcal{F}_{\text{LatentSVDD}} := \Big\{ f_{\mathbf{c},\Omega,\mathcal{Z}} = \big(\mathbf{x} \mapsto \Omega + \max_{\mathbf{z}\in\mathcal{Z}} 2\langle \mathbf{c}, \Psi(\mathbf{x}, \mathbf{z})\rangle - \|\Psi(\mathbf{x}, \mathbf{z})\|^2\big) : 0 \le \|\mathbf{c}\|^2 + \Omega \le \lambda \Big\},$$

and its corresponding loss class $\mathcal{G}_{\text{LatentSVDD}} := l \circ \mathcal{F}_{\text{LatentSVDD}}$, employing the loss function $l(t) := \max(0, -t)$. It is not difficult to verify that (e.g., [199], Proposition 12), by employing the variable substitution $\Omega := R^2 - \|\mathbf{c}\|^2$, for any $C > 0$ there is an $\lambda > 0$ such that Problem (LatentSVDD) is equivalent to

$$\min_{f\in\mathcal{F}_{\text{LatentSVDD}}} \frac{1}{n} \sum_{i=1}^{n} l(f(x_i)) \;=\; \min_{g\in\mathcal{G}_{\text{LatentSVDD}}} \frac{1}{n} \sum_{i=1}^{n} g(x_i).$$

Hence, we may analyze the proposed LatentSVDD within the proven framework of empirical risk minimization.

Let us first briefly review the classical setup of empirical risk minimization. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be an i.i.d. sample drawn from a probability distribution P over $\mathcal{X}$. Let $\mathcal{F}$ be a class of functions mapping from $\mathcal{X}$ to some set $\mathcal{Y}$, and let $l : \mathcal{Y} \to [0, b]$ be a bounded loss function, for some $b > 0$. The goal is to find a function $f \in \mathcal{F}$ that has a low risk $\mathbb{E}[l(f(x))]$. Denoting the loss class by $\mathcal{G} := l \circ \mathcal{F}$, this is equivalent finding a function g with small $\mathbb{E}[g]$. The best function in $\mathcal{G}$ we can hope to learn is $g^* \in \operatorname{argmin}_{g \in \mathcal{G}} \mathbb{E}[g]$. Since g^* is unknown, we instead compute a minimizer $\widehat{g}_n \in \operatorname{argmin}_{g \in \mathcal{G}} \widehat{\mathbb{E}}[g]$, where $\widehat{\mathbb{E}}[g] := \frac{1}{n} \sum_{i=1}^n g(x_i)$. To compare the prediction accuracies of g^* and $\widehat{g}_n$, it is known [200] that, with probability at least $1 - \delta$ over the draw of the sample,

$$\mathbb{E}[\widehat{g}_n] - \mathbb{E}[g^*] \;\leq\; 4R_n(\mathcal{G}) + b\sqrt{\frac{2\log(2/\delta)}{n}}. \tag{6.5}$$

Here, $R_n(\mathcal{G}) := \mathbb{E} \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i g(x_i)$ is the *Rademacher complexity*, where $\sigma_1, \ldots, \sigma_n$ are i.i.d. Rademacher variables (random signs). Usually $R_n(\mathcal{G})$ is of the order $O(1/\sqrt{n})$, when we employ appropriate regularization, and thus so is (6.5). We will show that also LatentSVDD enjoys this favorable rate:

Theorem 13 (Generalization bound for LatentSVDD)**.** *Let* $g^* \in \operatorname{argmin}_{g \in \mathcal{G}_{\text{LatentSVDD}}} \mathbb{E}[g]$ *and* $\widehat{g}_n \in \operatorname{argmin}_{g \in \mathcal{G}_{\text{LatentSVDD}}} \frac{1}{n} \sum_{i=1}^n g(x_i)$*. Assume there is a real number* $B > 0$ *such that* $\mathbb{P}(\|\Psi(\mathbf{x}_i, \mathbf{z})\| \leq B) = 1$*. Denote the cardinality of* $\mathcal{Z}$ *by* $|\mathcal{Z}|$*. Then, the following generalization bound holds:*

$$\mathbb{E}[\widehat{g}_n] - \mathbb{E}[g^*] \;\leq\; 4|\mathcal{Z}|\frac{\lambda + B\sqrt{\lambda}}{\sqrt{n}} + B\sqrt{\frac{2\log(2/\delta)}{n}}.$$

Sketch of Proof. For the proof, we proceed in three steps: first, we prove a Rademacher bound for the classic SVDD (cf. the lemma below). Next, we use Lemma 8.1 in [201] to conclude a Rademacher bound for LatentSVDD. Finally, we conclude the claimed result by (6.5). □

The complete proof of Theorem 13 is shown in Appendix B.1. It builds on the following generalization bound for the classic SVDD, which is also proved in the supplement.

Lemma 1 (Rademacher bound for SVDD)**.** *Put* $\mathcal{F}_{\text{SVDD}}(\mathbf{z}) := \Big\{ f_{\mathbf{c},\Omega} = \big(\mathbf{x} \mapsto \Omega + 2\langle \mathbf{c}, \Psi(\mathbf{x}_i, \mathbf{z})\rangle - \|\Psi(\mathbf{x}_i, \mathbf{z})\|^2\big) \;:\; 0 \leq \|\mathbf{c}\|^2 + \Omega \leq \lambda \Big\}$ *and* $\mathcal{G}_{\text{SVDD}}(\mathbf{z}) := l \circ \mathcal{F}_{\text{SVDD}}(\mathbf{z})$ *with* $l(t) := \max(0, -t)$*. Assume there is a real number* $B > 0$ *such that* $\mathbb{P}(\|\Psi(\mathbf{x}_i, \mathbf{z})\| \leq B) = 1$*. Then the Rademacher complexity of* $\mathcal{G}_{\text{SVDD}}$ *is bounded as follows:*

$$R(\mathcal{G}_{\text{SVDD}}(\mathbf{z})) \;\leq\; \frac{\lambda + B\sqrt{\lambda}}{\sqrt{n}}.$$

As a result the convergence rate can be slightly or even considerably slower than $O(\sqrt{1/n})$), depending on the "degree of violation" of the independence assumption.

6.3 Direct Formulation includes *k*-means as Special Case

Using joint feature maps increases the flexibility of the model. However, if it is known that latent classes will be used, a more direct approach would have various advantages, i.e. less complex. Moreover, we show in this section that using a direct formulation of the above introduced LatentSVDD includes *k*-means as a special case. Hence, our contributions here are

- we give a comprehensive review including proofs of the properties of support vector data descriptions

- we introduce the direct formulation of the LatentSVDD (which we call *ClusterSVDD*) and show that it contains *k*-means as well as the standard SVDD as a special case
- we propose a corresponding solver for primal and dual formulations.

We start by introducing *k*-means and the re-visit SVDD where we prove most important properties. Finally, we introduce our *ClusterSVDD* and proof that *k*-means and SVDD are contained as a special case. Given a set of input instances $\mathbf{x}_1, \ldots, \mathbf{x}_\ell \in \mathcal{X}$, where $\mathcal{X}$ is an arbitrary set that is commonly assumed to be realized from a sequence of independent and identically distributed (i.i.d) random variables. Furthermore, k denotes the number of clusters and $z_i \in \{1, \ldots, k\}$ the membership of the corresponding input instance $\mathbf{x}_i$. Memberships can be expressed by partition sets $\{S_j\}_{j=1}^k$, where $i \in S_j$ if and only if $z_i = j$. It holds that $S_i \cap S_j = \emptyset$ for $i \neq j$ and $\cup_{j=1}^k S_j = \{1, \ldots, \ell\}$.

Kernel based approaches [33, 35] allow the input instances to be mapped into a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ via a feature map $\phi : \mathcal{X} \to \mathcal{H}$.

***k*-means Clustering** *k*-means clustering [64] (a recent overview is given in [202]) is usually introduced as a (non-convex) optimization problem of finding a partition $\{S_j\}_{j=1}^k$, for a pre-defined k, that minimizes the within cluster sum-of-squares (WCSS),

$$\min_{\{S_j\}_{j=1}^k} \sum_{j=1}^{k} \sum_{i \in S_j} \|\mathbf{x}_i - \mathbf{c}_j\|^2 \,, \tag{6.6}$$

with $\{\mathbf{c}_j \in \mathcal{X}\}_{j=1}^k$ being the means of the corresponding clusters. Solving this problem (at least locally optimal) consists of three simple steps:

(1) Initialize the cluster centers $\{\mathbf{c}_j\}_{j=1}^k$ and repeat step (2) & (3) until no changes occur.

(2) Update the partitions $\{S_j\}_{j=1}^k$ by identifying the nearest cluster given the intermediate cluster centers $\mathbf{c}_\cdot$, $z_i = \operatorname{argmin}_{\hat{z} \in \{1,\ldots,k\}} \|\mathbf{c}_{\hat{z}} - \mathbf{x}_i\|^2$.

(3) Update the cluster centers $\mathbf{c}_j = 1/|S_j| \sum_{i \in S_j} \mathbf{x}_i$, $\forall j = 1, \ldots, k$.

Since its introduction, efforts have been made to increase the flexibility of the description [35, 203, 204], i.e. through the use of kernels ([33–35] for an introduction to kernel methods), and to increase the robustness of the method [205–207] against outliers and the curse of dimensionality. In this work, we tackle all of the above mentioned into a single framework.

Another line of research, which we do not investigate further in this work, deals with the inference of the correct number of partitions k [208–212].

Let us first begin by introducing an alternative way of stating the optimization problem of *k*-means. Instead of stating that the cluster centers $\{\mathbf{c}_j\}_{j=1}^k$ should be the means of input instances corresponding to the cluster, we can re-write OP (6.6) more concisely as

$$\min_{\{S_j\}_{j=1}^k} \sum_{j=1}^{k} \min_{\mathbf{c}_j} \sum_{i \in S_j} \|\mathbf{c}_j - \mathbf{x}_i\|^2 \,.$$

This yields the same solution since it is a convex problem w.r.t. $\mathbf{c}_j$ (fixing the partitions), and we can analytically derive the optimal solution by $\partial \left(\sum_{i \in S_j} \|\mathbf{c}_j - \mathbf{x}_i\|^2 \right) / \partial \mathbf{c}_j = 0$, therefore $\mathbf{c}_j^* = 1/|S_j| \sum_{i \in S_j} \mathbf{x}_i$. We can now define an equivalent constrained formulation of OP (6.6).

Definition 10 (*k*-means Constrained Problem)**.** *The constrained optimization problem for* k*-means is given by*

$$\sum_{j=1}^{k} \min_{\mathbf{c}_j} \sum_{i \in S_j} \|\mathbf{c}_j - \mathbf{x}_i\|^2 \tag{6.7}$$
$$\textit{subject to } z_i = \operatorname*{argmin}_{\hat{z} \in \{1,\ldots,k\}} \|\mathbf{c}_{\hat{z}} - \mathbf{x}_i\|^2 \ ,$$

where $i \in S_j$*, if and only if* $z_i = j,\ \forall\ i = 1, \ldots, \ell$ *and* $\forall\ j = 1, \ldots, k$*.*

Revisiting SVDD As noted in the literature [213–215], there are some issues with the original formulation of the SVDD as defined in Problem 6.1. First, the formulation is not convex due to R^2 in the constraints and second, the primal-dual relation breaks down for $0 < C < 1/\ell$. However, this can be fixed and we derive here a rigorous formulation of the SVDD based on the work of Chang et al. [213].

Definition 11 (Primal Constrained Problem)**.** *The primal SVDD optimization problem as a quadratically constrained linear program (QCLP) is given by:*

$$\min_{\mathbf{c}, T \geq 0, \xi \geq 0} T + \frac{1}{\ell\nu} \sum_{i=1}^{\ell} \xi_i \tag{6.8}$$
$$\textit{subject to} \quad \|\mathbf{c} - \phi(\mathbf{x}_i)\|^2 \leq T + \xi_i \quad \forall\ i = 1, \ldots, \ell$$

for all $0 < \nu < 1$ *the constraint* $T \geq 0$ *is dispensable (cf. Lemma* (3)*). We will denote the OP* (6.8) *as Svdd*$(\nu, \{\mathbf{x}_i\}_{i=1}^{\ell})$*.*

Note that ξ_i in OP (6.8) can be substituted, which allows for an unconstrained formulation of the SVDD.

Definition 12 (Primal Unconstrained Problem)**.** *The primal convex, non-smooth, and unconstrained SVDD optimization problem is given by:*

$$\min_{\mathbf{c}, T \geq 0} T + \frac{1}{\ell\nu} \sum_{i=1}^{\ell} \max(0, \|\mathbf{c} - \phi(\mathbf{x}_i)\|^2 - T) \ . \tag{6.9}$$

This definition comes in handy, as solving SVDD in the primal form is sufficient and sub-gradient based solver can be applied.

Deriving a linearly constrained quadratic program (QP) allows to pin the relation between SVDDs and OC-SVMs.

Theorem 14 (Quadratic Program Formulation and Equivalence to One-class SVM)**.** *The SVDD primal optimization problem, given by OP* (6.8)*, can be transformed into the following equivalent linearly constrained quadratic program (QP):*

$$\min_{\mathbf{w}, \rho, \xi \geq 0} \frac{1}{2}\|\mathbf{w}\|^2 - \rho + \frac{1}{\ell\nu} \sum_{i=1}^{\ell} \xi_i \tag{6.10}$$
$$\textit{subject to} \quad \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle \geq \rho + \frac{1}{2}\|\phi(\mathbf{x}_i)\|^2 - \xi_i \ ,\ \forall\ i = 1, \ldots, \ell \ ,$$

i.e. for L_2*-normalized feature vectors* $\|\phi(\mathbf{x})\| = const$*, the above formulation reduces to the one-class SVM formulation as given in Chapter 3.*

Proof. Starting from the formulation of the primal SVDD in OP (6.8), we first extend the constraints from $\|\mathbf{c} - \phi(\mathbf{x}_i)\|^2 \leq T + \xi_i$ to $\|\mathbf{c}\|^2 - 2\langle \mathbf{c}, \phi(\mathbf{x}_i)\rangle + \|\phi(\mathbf{x}_i)\|^2 \leq T + \xi_i$. Second, we re-arrange terms and arrive at $\frac{\|\mathbf{c}\|^2 - T}{2} + \frac{\|\phi(\mathbf{x}_i)\|^2}{2} - \frac{\xi_i}{2} \leq \langle \mathbf{c}, \phi(\mathbf{x}_i)\rangle$. In a third step, we substitute $\rho = \frac{\|\mathbf{c}\|^2 - T}{2} \in \mathbb{R}$, $\zeta_i = \frac{\xi_i}{2} \in \mathbb{R}^+$, and $\mathbf{c} = \mathbf{w} \in \mathcal{H}$, which changes the objective function $T + \frac{1}{\ell\nu}\sum_{i=1}^{\ell} \xi_i$ towards $\|\mathbf{w}\|^2 - 2\rho + \frac{1}{\ell\nu}\sum_{i=1}^{\ell} 2\zeta_i$. Without changing the minimizer, we can multiply the objective by $\frac{1}{2}$ and arrive at the one-class SVM objective $\frac{1}{2}\|\mathbf{w}\|^2 - \rho + \frac{1}{\ell\nu}\sum_{i=1}^{\ell} \zeta_i$ with corresponding constraints $\langle \mathbf{w}, \phi(\mathbf{x}_i)\rangle \geq \rho + \frac{1}{2}\|\phi(\mathbf{x}_i)\|^2 - \zeta_i$. Which proves the first part of the theorem. For the second part, a simple substitution $\hat{\rho} = \rho + \frac{1}{2}\|\phi(\mathbf{x}_i)\|^2 = \rho + \frac{const^2}{2} \in \mathbb{R}$ leads to the desired outcome. □

Lemma 2. *Assume $\nu \leq 1/\ell$ is given, then OP* (6.8) *reduces to the minimum enclosing ball (MEB) problem, i.e. it holds that $\{\xi_i\}_{i=1}^{\ell} = 0$ (hard margin).*

Proof. Assume an optimal solution of OP (6.8) is given by $(T^*, \mathbf{c}^*, \{\xi_i^*\}_{i=1}^{\ell})$. Assume another solution $(T^* + \xi_m^*, \mathbf{c}^*, \{0\}_{i=1}^{\ell})$, where $\xi_m^* = \max_{i\in\{1,\ldots,\ell\}} \xi_i^*$, which is a feasible solution, Therefore,

$$(T^* + \xi_m^*) + \frac{1}{\ell\nu}\sum_{i=1}^{\ell} 0 = T^* + \xi_m^* \leq T^* + \frac{1}{\ell\nu}\sum_{i=1}^{\ell} \xi_i^* \Rightarrow \nu\xi_m^* \leq \frac{1}{\ell}\sum_{i=1}^{\ell} \xi_i^* = \frac{1}{\ell}\left(\sum_{i\setminus m}^{\ell} \xi_i^* + \xi_m^*\right),$$

is strictly fulfilled for $\nu < 1/\ell$ and hence, any optimal solution must include $\{\xi_i^*\}_{i=1}^{\ell} = \{0\}_{i=1}^{\ell}$ and for $\nu = 1/\ell$, the set of optimal solutions does include $\{\xi_i^*\}_{i=1}^{\ell} = \{0\}_{i=1}^{\ell}$. □

Lemma 3. *Assume $0 < \nu < 1$ is given, then the non-negativity constraint in OP* (6.8)*, $T \geq 0$, can be omitted.*

Proof. (According to [213], Theorem 3, Proof in Appendix A) Assume an optimal solution of OP (6.8) is given by $(T^*, \mathbf{c}^*, \{\xi_i^*\}_{i=1}^{\ell})$. Further, assume that $T^* = -|T^*|$ and another feasible solution that does not change the constraints is given by $(0, \mathbf{c}^*, \{\xi_i^* - |T^*|\}_{i=1}^{\ell})$, i.e. $0 \leq \|\mathbf{c}^* - \phi(\mathbf{x}_i)\|^2 \leq -|T^*| + \xi_i^*$. It holds that

$$-|T^*| + \frac{1}{\ell\nu}\sum_{i=1}^{\ell} \xi_i^* \geq 0 + \frac{1}{\ell\nu}\sum_{i=1}^{\ell}(\xi_i^* - |T^*|) = \frac{-|T^*|}{\nu} + \frac{1}{\ell\nu}\sum_{i=1}^{\ell} \xi_i^*$$

is true for $\nu < 1$ and hence, is a contradiction to the assumption that $-|T^*| = T^*$. □

Lemma 4. *Assume $\nu \geq 1$ is given, then due to the non-negativity constraint in OP* (6.8)*, $T \geq 0$, the optimal solution must have $T^* = 0$.*

Proof. (According to [213], Theorem 3, Proof in Appendix A) Assume an optimal solution of OP (6.8) is given by $(T^*, \mathbf{c}^*, \{\xi_i^*\}_{i=1}^{\ell})$ and another feasible solution, that does not change the constraints, is given by $(0, \mathbf{c}^*, \{\xi_i^* + T^*\}_{i=1}^{\ell})$. It holds that

$$T^* + \frac{1}{\ell\nu}\sum_{i=1}^{\ell} \xi_i^* \geq 0 + \frac{1}{\ell\nu}\sum_{i=1}^{\ell}(\xi_i^* + T^*) = \frac{T^*}{\nu} + \frac{1}{\ell\nu}\sum_{i=1}^{\ell} \xi_i^* ,$$

is true for $\nu \geq 1$ and hence, the optimal solution must have $T^* = 0$. □

Therefore, we can now state precise primal and dual optimization problems.

Theorem 15 (Primal Problem and Solution for $\nu \geq 1$)**.** *If $\nu \geq 1$ the primal optimization problem reduces to*

$$\min_{\mathbf{c}} \sum_{i=1}^{\ell} \|\mathbf{c} - \phi(\mathbf{x}_i)\|^2, \tag{6.11}$$

and the optimal solution is given by $\mathbf{c} = 1/\ell \sum_{i=1}^{\ell} \phi(\mathbf{x}_i)$.

Proof. According to Lemma 4, $T = 0$ and $\frac{1}{\ell\nu} > 0$ can be discarded, hence we arrive at

$$\begin{aligned} &\min_{\mathbf{c},\xi \geq 0} \sum_{i=1}^{\ell} \xi_i \\ &\text{subject to} \quad \|\mathbf{c} - \phi(\mathbf{x}_i)\|^2 \leq \xi_i \quad \forall\, i = 1, \ldots, \ell. \end{aligned}$$

Further, $\xi_i \geq 0$ is due to the 2-norm always fulfilled and minimization yields the smallest possible $\xi_i = \|\mathbf{c} - \phi(\mathbf{x}_i)\|^2$ which reads unconstrained

$$\min_{\mathbf{c}} L(\mathbf{c}) = \sum_{i=1}^{\ell} \|\mathbf{c} - \phi(\mathbf{x}_i)\|^2 \ .$$

This quadratic form has a unique optimum at $\partial L(\mathbf{c})/\partial \mathbf{c} = 0$, which is $\mathbf{c} = 1/\ell \sum_{i=1}^{\ell} \phi(\mathbf{x}_i)$.

For $\nu \geq 1$ the dual problem can be solved analytically by $\alpha = 1/\ell$. □

Theorem 16 (Dual Problem)**.** *For $0 < \nu \leq 1$ and appropriately defined Mercer-kernel $\mathsf{k} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}, \mathsf{k}(\mathbf{x}, \mathbf{y}) \mapsto \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$, the dual problem is given by*

$$\begin{aligned} &\max_{0 \leq \alpha \leq \frac{1}{\ell\nu}} \sum_{i=1}^{\ell} \alpha_i \mathsf{k}(\mathbf{x}_i, \mathbf{x}_i) - \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \alpha_i \alpha_j \mathsf{k}(\mathbf{x}_i, \mathbf{x}_j) \\ &\textit{subject to} \quad \sum_{i=1}^{\ell} \alpha_i = 1 \end{aligned} \tag{6.12}$$

with expansions $\mathbf{c} = \sum_{i=1}^{\ell} \alpha_i \phi(\mathbf{x}_i)$.

Proof. Due to Lemma 3, we can skip the non-negativity constraint $T \geq 0$ of the convex OP (6.8). The resulting Lagrangian arrives at

$$L(\alpha, \beta, \mathbf{c}, T, \xi) = T + \frac{1}{\ell\nu} \sum_{i=1}^{\ell} \xi_i + \sum_{i=1}^{\ell} \alpha_i (\|\mathbf{c} - \phi(\mathbf{x}_i)\|^2 - T - \xi_i) - \sum_{i=1}^{\ell} \beta_i \xi_i \ ,$$

and solving for the Lagrange dual function $g(\alpha, \beta)$ (with $\alpha \geq 0, \beta \geq 0$ and $g(\alpha, \beta) = \min_{\mathbf{c},T,\xi} L(\alpha, \beta, \mathbf{c}, T, \xi)$) yields

(1) $\frac{1}{\ell\nu} - \beta_i - \alpha_i = 0$ and hence, $0 \leq \alpha \leq \frac{1}{\ell\nu}$

(2) the expansion $\mathbf{c} = \sum_{i=1}^{\ell} \alpha_i \phi(\mathbf{x}_i)$

(3) the equality constraint $\sum_{i=1}^{\ell} \alpha_i = 1$.

Substitution and re-arrangement then gives us the dual optimization problem in OP (6.12). In order for strong duality to hold, some constraint qualifications, such as *Slater's condition*, must be fulfilled (which holds trivially, cf. [213] Section 3.1). For any primal $(\mathbf{c}^*, T^*, \xi^*)$

and dual optimal solution (α^*, β^*), the complementary slackness constraints are given by $\alpha_i^*(\|\mathbf{c}^* - \phi(\mathbf{x}_i)\|^2 - T^* - \xi_i^*) = 0$ and $\beta_i^* \xi_i^* = 0$. □

Interestingly, the above formulation reduces to the dual one-class SVM optimization problem [7], if $\mathsf{k}(\mathbf{x}, \mathbf{y})$ is a constant for $\mathbf{x} = \mathbf{y}$. Also, the dual formulation allows for a neat interpretation of the ν parameter.

Theorem 17. *Given $0 < \nu \leq 1$, then $\lceil \ell\nu \rceil$ is a lower bound on the number of support vectors and an upper bound on the number of outliers.*

Proof. Due to the complementary slackness constraints (Thm. 16, cf. [213], Eq. (12,17)), we know that constraints in Eq. (6.8) that are not strictly fulfilled yield $\alpha^* = 1/\ell\nu$ ($\xi_i^* > 0 \Rightarrow \beta_i^* = 0$ and $\frac{1}{\ell\nu} - \beta_i^* - \alpha^* = 0$ must hold), whereas constraints that are strictly fulfilled receive $\alpha^* = 0$ ($\xi_i^* = 0$, $\|\mathbf{c}^* - \phi(\mathbf{x}_i)\|^2 < T^*$ and complementary slackness must hold). For data points lying exactly on the border, it holds that $0 \leq \alpha^* \leq 1/\ell\nu$ ($\xi_i^* = 0$ and $\|\mathbf{c}^* - \phi(\mathbf{x}_i)\|^2 = T^*$). Therefore, in order to fulfill the equality constraint in Problem (6.12), at most $\lceil \ell\nu \rceil$ data points can strictly lie outside and there must be at least that much support vectors. □

It therefore makes sense, to restrict ν to be in range $]0, 1]$.

ClusterSVDD In this section, we introduce our unifying formulation ClusterSVDD and prove that *k*-means and SVDD can be recovered as special cases. The core idea of is displayed in Figure 6.1.

Definition 13 (Primal Problem)**.** *Primal non-convex ClusterSVDD optimization problem (again $0 < \nu \leq 1$):*

$$\begin{aligned} \min_{\{\mathbf{c}_j\}_{j=1}^k, \mathbf{T} \geq 0, \xi \geq 0} & \sum_{j=1}^{k} T_j + \sum_{j=1}^{k} \sum_{i=1}^{\ell} \frac{\mathbf{1}[z_i = j]}{\sum_l \mathbf{1}[z_l = j]\nu} \xi_i \\ \textit{subject to} \quad & \|\mathbf{c}_{z_i} - \phi(\mathbf{x}_i)\|_2^2 \leq T_{z_i} + \xi_i, \ \forall\, i = 1, \ldots, \ell \\ \textit{with} \quad & z_i = \operatorname*{argmin}_{\hat{z} \in \{1, \ldots, k\}} \|\mathbf{c}_{\hat{z}} - \phi(\mathbf{x}_i)\|^2 - T_{\hat{z}} \end{aligned} \tag{6.13}$$

Theorem 18 (Decomposability)**.** *The Problem* (6.13) *is decomposable into k sub-problems with k disjunct sets of hypersphere constraints and ℓ global cluster membership constraints.*

Proof. Notice that the data can be partitioned, that is, each datum x_i can only belong to a single set S_j at any given time, where $i \in S_j$ for $j \in 1, \ldots, k$ iff $z_i = j$. It follows that $S_i \cap S_j = \emptyset$ for $i \neq j$ and $\cup_{j=1}^k S_j = \{1, \ldots, \ell\}$. Re-writing $\sum_{i=1}^{\ell} \frac{\mathbf{1}[z_i = j]}{\sum_l \mathbf{1}[z_l = j]\nu} \xi_i =$

$\frac{1}{|S_j|\nu}\sum_{i\in S_j}\xi_i$ (in Problem (6.13)) and arranging terms accordingly achieves

$$\begin{aligned}
&\min_{\{\mathbf{c}_j\}_{j=1}^k,\mathbf{T}\geq 0,\xi\geq 0}\sum_{j=1}^{k}\left(T_j+\frac{1}{|S_j|\nu}\sum_{i\in S_j}\xi_i\right)\\
&\qquad=\sum_{j=1}^{k}\min_{\mathbf{c}_j,T_j\geq 0,\xi\geq 0} T_j+\frac{1}{|S_j|\nu}\sum_{i\in S_j}\xi_i \qquad (6.14)\\
&\text{subject to}\quad \|\mathbf{c}_j-\phi(\mathbf{x}_i)\|^2\leq T_j+\xi_i,\ \forall\, i\in S_j, j=1\\
&\qquad\qquad\vdots\qquad\qquad\vdots\qquad\qquad\vdots\\
&\qquad\qquad \|\mathbf{c}_j-\phi(\mathbf{x}_i)\|^2\leq T_j+\xi_i,\ \forall\, i\in S_j, j=k\\
&\text{with}\quad z_i=\operatorname*{argmin}_{\hat{z}\in\{1,\dots,k\}}\|\mathbf{c}_{\hat{z}}-\phi(\mathbf{x}_i)\|^2-T_{\hat{z}}\,,\ \forall\, i=1,\dots,\ell.
\end{aligned}$$

The above optimization problem is now decomposed into k distinct SVDD optimization problems that are coupled solely through the global cluster assignment constraint. Hence, by applying the notation introduced in Def. 11, the ClusterSvdd optimization problem OP (6.14) can be written as

$$\begin{aligned}
&\sum_{j=1}^{k}\mathrm{Svdd}(\nu,\{\mathbf{x}_i\}_{i\in S_j}) \qquad (6.15)\\
&\text{subject to}\quad z_i=\operatorname*{argmin}_{\hat{z}\in\{1,\dots,k\}}\|\mathbf{c}_{\hat{z}}-\phi(\mathbf{x}_i)\|^2-T_{\hat{z}}\,,\ \forall\, i=1,\dots,\ell.
\end{aligned}$$

□

This is also an interesting result for the optimization in Section 6.3. Notably, given the partitions $\{S_j\}_{j=1}^k$, OP (6.15) is just a sum of convex optimization problems, which itself is a convex optimization problem [43]. Because the problem decomposes neatly into SVDD sub-problems (with exact primal-dual relations where strong duality holds), using kernels is straightforward by simply solving the dual SVDD Problem (6.12) instead of the primal version. We now proceed further and show the equivalence to *k*-means when $\nu \geq 1$.

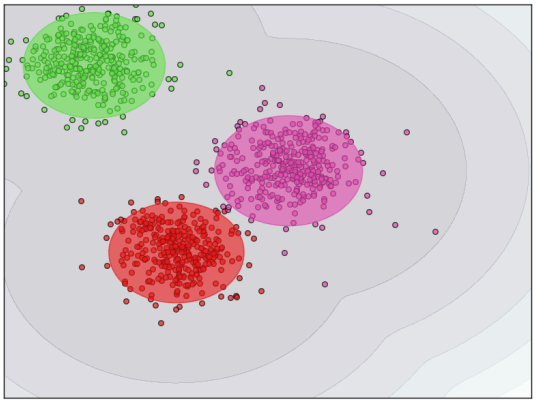

Figure 6.1 – The model: Fitting multiple hyperspheres simultaneously with a pre-defined outlier fraction is the core idea of our proposed method ClusterSVDD.

Theorem 19 (Equivalence I). *Assume* $\nu \geq 1$ *given and* $\phi : \mathcal{X} \to \mathcal{X},\ \mathbf{x} \mapsto \mathbf{x}$ *being the identity function* $id_{\mathcal{X}}$*, then* ClusterSVDD *optimization problem is identical to the* k*-means optimization problem: OP* (6.13) = *OP* (6.7).

Proof. Since the OP (6.8) can be decomposed into OP (6.15) and Thm. 15 holds for each sub-SVDD,

$$\sum_{j=1}^{k} \text{Svdd}(\nu \geq 1, \{\mathbf{x}_i\}_{i \in S_j}) = \sum_{j=1}^{k} \min_{\mathbf{c}_j} \sum_{i \in S_j} \|\mathbf{c}_j - \phi(\mathbf{x}_i)\|^2$$
$$\text{subject to } z_i = \underset{\hat{z} \in \{1,\dots,k\}}{\text{argmin}} \|\mathbf{c}_{\hat{z}} - \phi(\mathbf{x}_i)\|^2 \,, \; \forall \, i = 1, \dots, \ell$$

which is identical to the *k*-means OP (6.7). □

Theorem 20 (Equivalence II)**.** *Assume* $k = 1$ *given, then* ClusterSVDD *optimization problem is identical to the SVDD optimization problem: OP* (6.13) = *OP* (6.8)*.*

Proof. Since the OP (6.8) can be decomposed into OP (6.15), the sum can be ommitted as well as the cluster membership constraints, as they always deliver $1 = z_i = \text{argmin}_{\hat{z} \in \{1,\dots,k\}} \|\mathbf{c}_{\hat{z}} - \phi(\mathbf{x}_i)\|^2 \,, \; \forall \, i = 1, \dots, \ell$. The resulting optimization problem is SVDD$(\nu, \{\mathbf{x}_i\}_{i=1}^{\ell})$, which is in fact the original SVDD formulation as defined in Def. 11. □

Relation to Kernel *k*-means and Spectral Clustering From the decomposability theorem (Thm. 18) a kernelized version of our ClusterSVDD can be derived using the dual of the SVDD as given in Thm. 16. Due to the expansion of $\mathbf{c} = \sum_{i=1}^{\ell} \alpha_i \phi(\mathbf{x}_i)$ of a single SVDD, we can equivalently rewrite the global cluster membership constraint of OP. (6.13) as

$$z_i := \underset{j \in \{1,\dots,k\}}{\text{argmin}} \sum_{m,n \in S_j} \alpha_m \alpha_n \mathsf{k}_j(\mathbf{x}_m, \mathbf{x}_n) - 2 \sum_{m \in S_j} \alpha_m \mathsf{k}_j(\mathbf{x}_m, \mathbf{x}_i) + \mathsf{k}_j(\mathbf{x}_i, \mathbf{x}_i) - T_j \,.$$

Moreover, a proper dual version of *k*-means can be derived as a special case due to Thm. 19 which ensures the equivalence to kernel *k*-means [204]. Interestingly, Dhillon et al. [203] showed that an explicit theoretical connection between kernel *k*-means and spectral clustering [216] can be drawn under certain conditions. In return, there is also a connection between our ClusterSVDD and spectral clustering with kernel *k*-means being the link.

Algorithm 9 ClusterSVDD

input data $x_1, \dots, x_\ell$ and outlier fraction $\nu > 0$
put $t = 0$
choose $z_i \in \{1, \dots, k\}$ $\forall i \in \{1, \dots, \ell\}$ (e.g. randomly)
let $(\mathbf{c}_j^t, T_j^t)$ be the optimal arguments when solving the SVDD optimization problem OP (6.8) with subset $\mathbf{x}_i, \; i \in S_j, \forall \, j = 1, \dots, k$.
repeat
  t:=t+1
  for $i = 1, \dots, \ell$ **do**
    $z_i^t := \text{argmin}_{j \in \{1,\dots,k\}} ||\mathbf{c}_j^{t-1} - \phi(\mathbf{x}_i)||^2 - T_j^{t-1}$
  end for
  let $(\mathbf{c}_j^t, T_j^t)$ be the optimal arguments when solving the SVDD optimization problem OP (6.8) with subset $\mathbf{x}_i, \; i \in S_j, \forall \, j = 1, \dots, k$.
until $\forall \, i = 1, \dots, \ell : \; z_i^t = z_i^{t-1}$
Return optimal model parameters $\mathbf{c}_\cdot := \mathbf{c}_\cdot^t$, $T_\cdot = T_\cdot^t$, the cluster memberships $z_i := z_i^t \quad \forall \, i = 1, \dots, \ell$, and the anomaly scores $s_i := ||\mathbf{c}_{z_i}^t - \phi(\mathbf{x}_i)||^2 - T_{z_i}^t$

Optimization Following the ideas of CCCP [197] (concave-convex procedure), a variant of DC-programming [52] (difference of convex functions), which itself is a special instance of MM (majorization-minimization), we separate the problem into two sub-problems:

(1) inferring the partition, and

(2) calculating the new hypersphere centers and radii.

This approach does not guarantee the globally optimal solution (except for $k = 1$), but will provide locally optimal solutions. Due to Theorem (18), the optimization is similar to the original *k*-means optimization, where the first step also considers kernels, and the second step can be solved using existing SVDD implementations. The resulting optimization algorithm is described in Algorithm 9 for its primal form and in Algorithm 10 for its kernelized counterpart. Despite the non-convex nature of the optimization problem, we found in our experiments that the algorithm tends to converge fast.

Algorithm 10 Kernel ClusterSVDD

input data $x_1, \ldots, x_\ell$ and outlier fraction $\nu > 0$
put $t = 0$
choose $z_i \in \{1, \ldots, k\}\ \forall i \in \{1, \ldots, \ell\}$ (e.g. randomly) and therefore fixing $S_j^t, \forall j = 1, \ldots, k$
let (α_j^t, T_j^t) be the optimal arguments when solving the SVDD dual optimization problem OP (6.12) with subset $\mathbf{x}_i,\ i \in S_j^t, \forall\, j = 1, \ldots, k$ for the corresponding kernel k_j.
repeat
  t:=t+1
  for $i = 1, \ldots, \ell$ **do**
    $z_i^t := \operatorname{argmin}_{j \in \{1,\ldots,k\}} \sum_{m,n \in S_j^{t-1}} \alpha_m^{t-1} \alpha_n^{t-1} \mathsf{k}_j(\mathbf{x}_m, \mathbf{x}_n)$
      $-2 \sum_{m \in S_j^{t-1}} \alpha_m^{t-1} \mathsf{k}_j(\mathbf{x}_m, \mathbf{x}_i) + \mathsf{k}_j(\mathbf{x}_i, \mathbf{x}_i) - T_j^{t-1}$
    Note: $z_i^t, \forall i = 1, \ldots, \ell$ implies $S_j^t, \forall\, j = 1, \ldots, k$
  end for
  let (α_j^t, T_j^t) be the optimal arguments when solving the SVDD dual optimization problem OP (6.12) with subset $\mathbf{x}_i,\ i \in S_j^t, \forall\, j = 1, \ldots, k$ for the corresponding kernel k_j.
until $\forall\, i = 1, \ldots, \ell :\ z_i^t = z_i^{t-1}$
Return optimal model parameters $\alpha_\cdot := \alpha_\cdot^t$, $T_\cdot = T_\cdot^t$, the cluster memberships $z_i := z_i^t \quad \forall\, i = 1, \ldots, \ell$, and the anomaly scores $s_i := \sum_{m,n \in S_{z_i}^t} \alpha_m^t \alpha_n^t \mathsf{k}_j(\mathbf{x}_m, \mathbf{x}_n) - 2 \sum_{m \in S_{z_i}^t} \alpha_m^t \mathsf{k}_j(\mathbf{x}_m, \mathbf{x}_i) + \mathsf{k}_j(\mathbf{x}_i, \mathbf{x}_i) - T_j^t$

6.4 Extension to Non-independent Samples

Inferring latent classes from observations does pose an restriction on our model that might not be wanted: similar observations will be always grouped together. While this is reasonable, there are settings where such properties are undesired. However, to overcome this restriction more information is necessary, e.g. dependency structure or additional label information.

In detail, the requirements for our model to cope with this setting are the following: (i) we are only interested in predicting the data points that we already have (transductive setting); (ii) of those, none or only few carry actual labels that we are interested in (scarce labeled data); (iii) data points with distinct labels can have the same behavioral values (overlapping clusters); (iv) selection of cluster based on structure (inference of structures in latent

space); (v) labels are based on the inferred structured latent states as well as their respective behavioral values. The resulting model is displayed in Figure 6.2.

We designed this method with a specific application in mind (cf. Section 6.5.2) which is, rather suprisingly, a regression setting and the resulting method is called transductive conditional random field regression (TCRFR). However, due to the sketched properties and its modularity, an extension towards one-class classification using methods derived earlier this chapter (LatentSVDD, Section 6.2), will be discussed in detail.

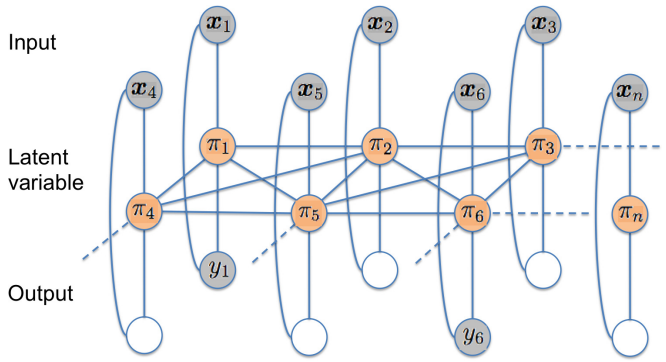

Figure 6.2 – The model: data points are connected through latent variables. Observations $\mathbf{x}$ are given but none or only few corresponding target values y are known.

Hence, our main contributions in this section are

- we derive a transductive regression method that leverages latent class dependency structure (transductive conditional random field regression, TCRFR);
- we present a corresponding solver based on loopy belief propagation and linear program approximations;
- we extend the methodology to contextual one-class classification based on the LatentSVDD (cf. Section 6.2).

As we shortly leave the beaten track of one-class classification, we start by reviewing the work that is related to the regression setting that we bear in mind.

Related Work According to the described properties, we grouped related work into 3 distinct classes:

Methods in group one consists can best be described as *general purpose*. These are algorithms that are fast, easy to apply and make only a few assumptions about the data. This, however, comes at the price of not leveraging all the information available and, therefore, creates less accurate predictions.

Methods in the second group are technically most closely related to our method and can be described as methods dealing with structured data. However, interestingly, none of these methods can be applied to our setting. That is, each and every method assumes IID training data. Further, from 7 methods, only 4 are regression methods [217–219] and only 3 consider continuous labels[217, 219]. All of these remaining 3 methods assume completely known latent states for training, which our setting does not provide. Structure has been modeled by conditional random fields (CRFs) [158], and extensions thereof comprise diverse continuous methods [217, 220]. For kernel machines, the classical structured output support vector machines (SSVM) [157], allows to learn on joint feature maps; for extensions to regression, see [218, 219, 221]. Extensions to semi-supervised settings have been developed [222].

Finally, methods in the third group do make many more assumptions on the data to extract more information for higher prediction accuracies. Mostly, these methods are specialized, advanced versions of their *general purpose* counterparts in the first group. Transductive Regression [223, 224] copes with the semi-supervised setting by inferring virtual labels for unlabeled examples by superposition of information of labeled examples [223]. Here, interactions between examples are imposed implicitly by choosing an appropriate metric. However,

those methods do not take latent dependency structure into account. Another line of research is a mixture of experts model [20, 225, 226], where multiple regression models (experts) are trained, and one (or a weighted sum) of thereof is used to predict the output label of new samples. Laplacian regularized learning machines [227–229] assume that data lies on a manifold in transductive or semi-supervised settings. We will apply this technique to kernelized support vector regression, which itself includes the function class of (kernel) ridge regression.

Transductive Conditional Random Field Regression (TCRFR) Given a labeled sample set $\mathcal{S} = \{(\mathbf{x}_i, y_i) \in \mathbb{R}^D \times \mathbb{R}\}_{i=1}^n$, and an unlabeled sample set $\mathcal{U} = \{\mathbf{x}_i \in \mathbb{R}^D\}_{i=n+1}^{n+m}$, consider a regression model with Gaussian noise:

$$y = f(\mathbf{x}; \mathbf{w}) + \epsilon, \quad \epsilon = y - f(\mathbf{x}; \mathbf{w}) \sim \mathcal{N}(0, \sigma^2),$$
$$p(y|\mathbf{x}, \mathbf{w}) \propto \exp(-\frac{1}{2\sigma^2}|y - f(\mathbf{x}; \mathbf{w})|^2) \ ,$$

where $\mathbf{x} \in \mathbb{R}^D$ and $y \in \mathbb{R}$ are input and output variables, respectively, and $f(\mathbf{x}; \mathbf{w}) = \langle \mathbf{w}, \mathbf{x} \rangle$ is a linear regression function with an unknown parameter $\mathbf{w} \in \mathbb{R}^D$. σ^2 denotes the noise variance. We assume the Gaussian prior for $\mathbf{w}$: $p(\mathbf{w}) \propto \exp\left(-\frac{\lambda'}{2}\|\mathbf{w}\|^2\right)$. Then, the *maximum a posteriori (MAP)* estimator is obtained by maximizing the joint distribution of $\{y_i\}_{i=1}^n$ and $\mathbf{w}$ (assuming IID data):

$$\max_{\mathbf{w} \in \mathbb{R}^D} p(\{y_i\}_{i=1}^n | \{\mathbf{x}_i\}_{i=1}^n, \mathbf{w}) p(\mathbf{w}) = \prod_{i=1}^n p(y_i|\mathbf{x}_i, \mathbf{w}) p(\mathbf{w}) \ , \tag{6.16}$$

or, equivalently, minimizing the negative logarithm of the joint distribution $\min_{\mathbf{w} \in \mathbb{R}^D} \mathcal{L}_0(\mathbf{w})$, where

$$\mathcal{L}_0(\mathbf{w}) = \lambda' \|\mathbf{w}\|_2^2 + \sum_i \frac{|y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle|^2}{\sigma^2} \ . \tag{6.17}$$

This is the standard ridge regression setting, which we extend threefold: First, in the spirit of kernel ridge regression, we introduce feature functions ϕ for the input data $\phi : \mathbb{R} \to \mathcal{X}$. Moreover, we explicitly model the dependency of the regression function $f(\mathbf{x})$ on a latent variable $\pi \in \mathcal{Z}$ using *local* joint feature maps $\Phi : \mathcal{X} \times \mathcal{Z} \to \mathcal{H}_1$ on the labeled sample set $\mathcal{S}$, Second, we focus on predicting labels on the unlabeled data set $\mathcal{U}$ only, and, finally, we respect the dependency of the inputs of the labeled and unlabeled sample sets $\mathcal{S}$ and $\mathcal{U}$ that can be exploited, e.g. they share spatial relations that can be modeled by conditional random fields (CRF) using a *global* joint feature map $\Psi : \bigotimes_{i=1}^{n+m} \mathcal{X} \times \bigotimes_{i=1}^{n+m} \mathcal{Z} \to \mathcal{H}_2$. Note that both *local* Ψ and *global* Φ features map the samples into reproducing kernel Hilbert spaces $\mathcal{H}$. that correspond to kernel functions [33]. This is a principled way of approaching the encoding problem for arbitrary dependencies between $\mathbf{x}$ and π as it is common in the structured output literature [157].

With these extensions, we tackle the problem of inferring latent variables under spatio-temporal structure from few precise output measurements and many noisy input measurements.

We propose transductive conditional random field regression (TCRFR, cf. Fig 6.2), which consists mainly of two parts: (a) a least-squares regression part with parameter $\mathbf{u}$, conditioned on the latent states and input instances, and (b) a conditional random field part with parameter $\mathbf{v}$ that explicitly models the dependencies of the latent variables and is conditioned on the input instances only. Both parts receive a Gaussian prior for stabilization and

we are only interested in *maximum a posteriori* estimates (starting from the ridge regression likelihood, cf. Eq. (6.16)):

$$\begin{aligned}
\max_{\mathbf{u}} p(\{y_i\}_{i=1}^n|\{\mathbf{x}_i\}_{i=1}^{n+m},\mathbf{u})p(\mathbf{u}) &\geq \max_{\mathbf{u},\mathbf{v},\{\pi_i\}_{i=1}^{n+m}} p(\{y_i\}_{i=1}^n,\{\pi_i\}_{i=1}^{n+m},\mathbf{v}|\{\mathbf{x}_i\}_{i=1}^{n+m},\mathbf{u})p(\mathbf{u}) \\
&= \max_{\mathbf{u},\mathbf{v},\{\pi_i\}_{i=1}^{n+m}} p(\{y_i\}_{i=1}^n,\{\pi_i\}_{i=1}^{n+m}|\{\mathbf{x}_i\}_{i=1}^{n+m},\mathbf{u},\mathbf{v})p(\mathbf{u})p(\mathbf{v}) \\
&= \max_{\mathbf{u},\mathbf{v},\{\pi_i\}_{i=1}^{n+m}} p(\{y_i\}_{i=1}^n|\{\pi_i\}_{i=1}^n,\{\mathbf{x}_i\}_{i=1}^n,\mathbf{u})p(\mathbf{u}) \\
&\qquad p(\{\pi_i\}_{i=1}^{n+m}|\{\mathbf{x}_i\}_{i=1}^{n+m},\mathbf{v})p(\mathbf{v}) \\
&= \max_{\mathbf{u},\mathbf{v},\{\pi_i\}_{i=1}^{n+m}} \prod_{i=1}^n p(y_i|\pi_i,\mathbf{x}_i,\mathbf{u})p(\mathbf{u}) \\
&\qquad p(\{\pi_i\}_{i=1}^{n+m}|\{\mathbf{x}_i\}_{i=1}^{n+m},\mathbf{v})p(\mathbf{v}). \qquad (6.18)
\end{aligned}$$

The probabilities are defined accordingly:

$$p(y|\pi,\mathbf{x},\mathbf{u}) \propto \exp\left(-\tfrac{|y-\langle\mathbf{u},\Phi(\mathbf{x},\pi)\rangle|^2}{2\sigma^2}\right), \qquad (6.19)$$

$$p(\mathbf{u}) \propto \exp\left(-\tfrac{\lambda'}{2}\|\mathbf{u}\|^2\right), \qquad (6.20)$$

$$p(\{\pi\}_{i=1}^{n+m}|\{\mathbf{x}_i\}_{i=1}^{n+m},\mathbf{v}) = \tfrac{1}{Z(\{\mathbf{x}_i\}_{i=1}^{n+m},\mathbf{v})}\exp\left(\langle\mathbf{v},\Psi(\{\mathbf{x}_i\}_{i=1}^{n+m},\{\pi_i\}_{i=1}^{n+m})\rangle\right), \qquad (6.21)$$

$$p(\mathbf{v}) \propto \exp\left(-\tfrac{1}{2}\mathbf{v}^\top\Gamma\mathbf{v}\right), \qquad (6.22)$$

where λ' and $\Gamma \in \mathcal{S}_+^{dim\mathcal{H}_2}$ (positive semi-definite matrix) are regularization constants and $Z(\{\mathbf{x}_i\}_{i=1}^{n+m},\mathbf{v}) = \sum_{\hat{\Pi}\in\bigotimes_{i=1}^{n+m}\mathcal{Z}}\exp\left(\langle\mathbf{v},\Psi(\{\mathbf{x}_i\}_{i=1}^{n+m},\hat{\Pi})\rangle\right)$ is the partition function. Thus, the MAP estimator for all unknown variables, including the model parameters $\mathbf{u}\in\mathcal{H}_1$ and $\mathbf{v}\in\mathcal{H}_2$, and the latent variables $\{\pi_i\}_{i=1}^{n+m}$, can be obtained by solving the following problem:

$$\min_{\mathbf{u}\in\mathcal{H}_1,\mathbf{v}\in\mathcal{H}_2,\{\pi_i\in\mathcal{Z}\}_{i=1}^{n+m}} \mathcal{L}(\mathbf{u},\mathbf{v},\{\pi_i\}_{i=1}^{n+m}), \qquad (6.23)$$

where $\mathcal{L}(\mathbf{u},\mathbf{v},\{\pi_i\}_{i=1}^{n+m})$ is a convex combination of the objectives of the regression model and the conditional random field:

$$\mathcal{L}(\mathbf{u},\mathbf{v},\{\pi_i\}_{i=1}^{n+m}) = \theta\mathcal{L}_{\text{rr}}(\mathbf{u},\{\pi_i\}_{i=1}^n) + (1-\theta)\mathcal{L}_{\text{crf}}(\mathbf{v},\{\pi_i\}_{i=1}^{n+m}), \qquad (6.24)$$

where

$$\mathcal{L}_{\text{rr}}(\mathbf{u},\{\pi_i\}_{i=1}^n) = \frac{\lambda}{2}\|\mathbf{u}\|_2^2 + \frac{1}{2}\sum_{i=1}^n|y_i - \langle\mathbf{u},\Phi(\mathbf{x}_i,\pi_i)\rangle|^2, \qquad (6.25)$$

$$\mathcal{L}_{\text{crf}}(\mathbf{v},\{\pi_i\}_{i=1}^{n+m}) = \frac{1}{2}\|\mathbf{v}\Gamma^{\frac{1}{2}}\|_2^2 - \langle\mathbf{v},\Psi(\{\mathbf{x}_i\}_{i=1}^{n+m},\{\pi_i\}_{i=1}^{n+m})\rangle + \log Z(\{\mathbf{x}_i\}_{i=1}^{n+m},\mathbf{v}). \qquad (6.26)$$

Here we re-parameterize the noise and the regularization parameters $\sigma^2 \to (1-\theta)\theta^{-1}$, $\sigma^2\lambda' \to \lambda$ for the regression part, so that the trade-off between the regression loss and the latent structure loss is explicit. From Eqn (6.24), it is apparent that $0 \leq \theta \leq 1$ is the parameter of a convex combination that weighs the CRF and the ridge regression objective functions. Setting $\theta = 1$ therefore assigns 100% weight to the ridge regression part, omitting any CRF input. Practically, θ will have to lie in the range $0 < \theta < 1$. Also, in most applications, labeled data will be sparse and hence, in those cases it is expected that $\theta > 0.5$ will prove preferable.

Algorithm 11 Transductive Conditional Random Field Regression (TCRFR)

input data $\mathcal{S}$ and $\mathcal{U}$
put $t = 0$ and initialize $\mathbf{u}^t$ and $\mathbf{v}^t$ (e.g., randomly)
repeat
t:=t+1
Update $\{\pi_i^t\}_{i=1}^{n+m}$ by Eq. (6.27) using the intermediate solutions $\mathbf{u}^{t-1}$ and $\mathbf{v}^{t-1}$
Update $\mathbf{u}^t$ by Eq.(6.28) and $\{\pi_i^t\}_{i=1}^{n+m}$
Update $\mathbf{v}^t$ by Eq.(6.29) and $\{\pi_i^t\}_{i=1}^{n+m}$
until $\forall\, i = 1, \ldots, N:\ \pi_i^t = \pi_i^{t-1}$
Predict unlabeled examples $\mathcal{U}$ using the inferred states $\{\pi_i^t\}_{i=n+1}^{n+m}$ and regression parameter $\mathbf{u}^t$: $y_i = \langle \mathbf{u}^t, \Phi(x_i, \pi_i^t)\rangle$

To solve the problem (6.23), we adopt a CCCP-style scheme [51, 197], a kind of majorization-minimization scheme, which has been successfully used in structured output settings with latent variables [230]. In each (t-th) iteration, we infer the most likely configuration $\{\pi_i\}$, given $\mathbf{u}$ and $\mathbf{v}$, for all training examples,

$$\begin{aligned}\{\hat{\pi}_i\}_{i=1}^{n+m} &= \operatorname*{argmin}_{\{\pi_i \in \mathcal{Z}\}_{i=1}^{n+m}} \mathcal{L}(\mathbf{u}, \mathbf{v}, \{\pi_i\}_{i=1}^{n+m}) \\ &= \operatorname*{argmin}_{\{\pi_i \in \mathcal{Z}\}_{i=1}^{n+m}} \frac{\theta}{2} \sum_{i=1}^{n} |y_i - \langle \mathbf{u}, \Phi(\mathbf{x}_i, \pi_i)\rangle|^2 - (1-\theta)\langle \mathbf{v}, \Psi(\{\mathbf{x}_i\}_{i=1}^{n+m}, \{\pi_i\}_{i=1}^{n+m})\rangle,\end{aligned} \tag{6.27}$$

and then update the ridge regression parameter $\mathbf{u}$ and the CRF parameter $\mathbf{v}$ respectively,

$$\hat{\mathbf{u}} = \operatorname*{argmin}_{\mathbf{u} \in \mathcal{H}_1} \mathcal{L}(\mathbf{u}, \mathbf{v}, \{\pi_i\}_{i=1}^{n}) = \operatorname*{argmin}_{\mathbf{u} \in \mathcal{H}_1} \mathcal{L}_{\text{rr}}(\mathbf{u}, \{\pi_i\}_{i=1}^{n}), \tag{6.28}$$

$$\hat{\mathbf{v}} = \operatorname*{argmin}_{\mathbf{v} \in \mathcal{H}_2} \mathcal{L}(\mathbf{u}, \mathbf{v}, \{\pi_i\}_{i=1}^{n+m}) = \operatorname*{argmin}_{\mathbf{v} \in \mathcal{H}_2} \mathcal{L}_{\text{crf}}(\mathbf{v}, \{\pi_i\}_{i=1}^{n+m}). \tag{6.29}$$

Below (cf. Algorithm 11), we detail how to perform each of the steps (6.27)–(6.29).

For Algorithm 11, we can show that each iteration monotonically decreases the objective if certain assumptions are met.

Theorem 21 (Monotonicity of Convergence for Algorithm 11)**.** *Given a minimizer for the inference problem in Eq.* (6.27) *that suffices*

$$\mathcal{L}(\mathbf{u}^t, \mathbf{v}^t, \{\pi_i^{t+1}\}_{i=1}^{n+m}) \leq \mathcal{L}(\mathbf{u}^t, \mathbf{v}^t, \{\pi_i^t\}_{i=1}^{n+m}), \tag{6.30}$$

then the log*-likelihood in Eq.* (6.23) *is monotonically decreasing for increasing number of iterations t, i.e. $\mathcal{L}(\mathbf{u}^t, \mathbf{v}^t, \{\pi^t\}_{i=1}^{n+m}) \leq \mathcal{L}(\mathbf{u}^{t-1}, \mathbf{v}^{t-1}, \{\pi^{t-1}\}_{i=1}^{n+m})$.*

Proof. $\mathcal{L}(\mathbf{u}^{t+1}, \mathbf{v}^{t+1}, \{\pi_i^{t+1}\}_{i=1}^{n+m}) \leq \min_{\{\mathbf{u}\}} \mathcal{L}(\mathbf{u}, \mathbf{v}^{t+1}, \{\pi_i^{t+1}\}_{i=1}^{n+m})$
$\leq \min_{\{\mathbf{v}\}} \mathcal{L}(\mathbf{u}^t, \mathbf{v}, \{\pi_i^{t+1}\}_{i=1}^{n+m}) \leq \mathcal{L}(\mathbf{u}^t, \mathbf{v}^t, \{\pi_i^{t+1}\}_{i=1}^{n+m}) \leq \mathcal{L}(\mathbf{u}^t, \mathbf{v}^t, \{\pi_i^t\}_{i=1}^{n+m})$ since Assumption (6.30) must hold, and due to the convexity of $\mathcal{L}_{\text{rr}}$ and $\mathcal{L}_{\text{crf}}$ (for fixed π). □

Choice of Joint Feature Maps Given an undirected graph $G = (V, E)$ with binary edges E and vertices V, where each vertex corresponds to a sample and the state space is $S = \mathcal{Z}$,

$$\Psi(\{\mathbf{x}_i\}_{i=1}^{n+m}, \{\pi_i\}_{i=1}^{n+m}) = \begin{pmatrix} (\sum_{(e_1,e_2)\in E} \mathbf{1}[\pi_{e_1} = s_1 \wedge \pi_{e_2} = s_2])_{(s_1,s_2)\in S,} \\ (\sum_{v \in V} \mathbf{1}[\pi_v = s]\, \phi(x_v))_{s \in S} \end{pmatrix}. \tag{6.31}$$

The graph-model consists basically of two parts: a transition part and a emission part.

Accordingly, we fix the regression joint feature map to be

$$\Phi(\mathbf{x}, \pi) = \phi(\mathbf{x}) \otimes \Lambda(\pi), \tag{6.32}$$

where $\Lambda(\pi) \in \{0,1\}^K$ with entries $(\Lambda(\pi))_k = 1$ if $\pi = k$ and 0 otherwise. $K \in \mathbb{N}^+$ is the number of hidden states and ϕ the feature function $\phi : \mathbb{R}^D \to \mathcal{X}$. Basically, the regression map is a K times replicated feature vector where all parts that do not correspond to the current active state π are set to zero. For further information and examples of joint feature maps, we refer to [157].

Latent State Inference (Eq.(6.27)) Latent state inference is computationally hard in general. While for tree-like structures efficient global inference schemes exist, this does not hold true for settings with loops. Since we are focusing on the latter, we rely on one of the two approximation methods:

We first discuss an approach based on linear program approximation. However, in this case, an extension to quadratic programs is necessary and hence, we call the resulting approach quadratic program approximation (QPA). This approach is inspired from the idea of linear program approximations and marginal polytopes [231]. Therefore, instead of using the explicit relation of the parameter vector with the joint feature map, we need to model the explicit relation between the latent variables:

$$\begin{aligned}
\mathcal{L}(\mathbf{u}, \mathbf{v}, \{\pi_i\}_{i=1}^{n+m}) &= \frac{\theta}{2} \sum_{i=1}^{n} |y_i - \langle \mathbf{u}, \Phi(\mathbf{x}_i, \pi_i)\rangle|^2 - (1-\theta)\langle \mathbf{v}, \Psi(\{\mathbf{x}\}_{i=1}^{n+m}, \{\pi\}_{i=1}^{n+m})\rangle \\
&= \frac{\theta}{2} \sum_{i=1}^{n} (\langle \mathbf{u}, \Phi(\mathbf{x}_i, \pi_i)\rangle)^2 - \theta \sum_{i=1}^{n} y_i \langle \mathbf{u}, \Phi(\mathbf{x}_i, \pi_i)\rangle \\
&\quad - (1-\theta)\langle \mathbf{v}, \Psi(\{\mathbf{x}\}_{i=1}^{n+m}, \{\pi\}_{i=1}^{n+m})\rangle + \text{const.} \\
&= \frac{\theta}{2} \mu_l^\top B(\mathbf{u}, \{\mathbf{x}_i\}_{i=1}^{n}) \mu_l - \theta \mu_l^\top c(\mathbf{u}, \{\mathbf{x}_i\}_{i=1}^{n}, \{y_i\}_{i=1}^{n}) \\
&\quad - (1-\theta)\big(\mu^\top d(\mathbf{v}, \{\mathbf{x}\}_{i=1}^{n+m}) + \sigma^\top e(\mathbf{v})\big) + \text{const.},
\end{aligned}$$

with the variables B, c, d, e defined accordingly:

$$\begin{aligned}
\forall_{i=1}^{n} :&\quad B_{i|S|:(i+1)|S|-1, i|S|:(i+1)|S|-1}(\mathbf{u}, \{\mathbf{x}_i\}_{i=1}^{n}) = (\langle \mathbf{u}_{s_1}, \phi(\mathbf{x}_i)\rangle \langle \mathbf{u}_{s_2}, \phi(\mathbf{x}_i)\rangle)_{(s_1,s_2)\in S}, \\
\forall_{i=1}^{n} :&\quad c_{i|S|:(i+1)|S|-1}(\mathbf{u}, \{\mathbf{x}_i\}_{i=1}^{n}, \{y_i\}_{i=1}^{n}) = (y_i \langle \mathbf{u}_s, \phi(\mathbf{x}_i)\rangle)_{s\in S}, \\
\forall_{i=1}^{n+m} :&\quad d_{i|S|:(i+1)|S|-1}(\mathbf{v}, \{\mathbf{x}\}_{i=1}^{n+m}) = (\langle \mathbf{v}_s, \phi(\mathbf{x}_i)\rangle)_{s\in S}, \\
\forall (e_1, e_2) \in E :&\quad e_{e_1,e_2}(\mathbf{v}) = \big(\mathbf{v}_{e_1,e_2}\big).
\end{aligned}$$

Here, $\mu_i^s = \Lambda(\pi_i)$ correspond to the relaxed unary term, $\sigma_{(e_1,e_2)}^{(s_1,s_2)} = [\pi_{e_1} = s_1 \wedge \pi_{e_2} = s_2]$ corresponding to the relaxed pairwise term, which must satisfy the marginal (i.e. local) polytope constraints [232] $\mathcal{P}(G)$. $\mu_l \subseteq \mu$ selector for labeled data points within all data points μ, $B(\mathbf{u}, \{\mathbf{x}_i\}_{i=1}^{n})$ is a matrix with $|S| \times |S|$ sub-matrices on its diagonal, $c(\mathbf{u}, \{\mathbf{x}_i\}_{i=1}^{n}, \{y_i\}_{i=1}^{n})$ is the linear part of the quadratic regression model containing the labels; $d(\mathbf{v}, \{\mathbf{x}\}_{i=1}^{n+m})$ contains the score of the parameter vector and the features for each vertex and each state; $e(\mathbf{v})$ is the vector of pairwise connection weights.

In our case, the regression term for labeled examples can be expressed as a positive semi-definite matrix which consequently leads to the following quadratic program formulation:

$$\{\mu_i^*\}_{i=1}^{n+m} = \underset{\mu,\sigma:\mathcal{P}(G)}{\operatorname{argmin}} \frac{\theta}{2}\mu_l^\top B(\mathbf{u}, \{\mathbf{x}_i\}_{i=1}^n)\mu_l - \theta\mu_l^\top c(\mathbf{u}, \{\mathbf{x}_i\}_{i=1}^n, \{y_i\}_{i=1}^n) - (1-\theta)\big(\mu^\top d(\mathbf{v}, \{\mathbf{x}_i\}_{i=1}^{n+m}) + \sigma^\top e(\mathbf{v})\big).$$

While we found empirically that this approach is more reliable and stable than loopy belief propagation, it is also computationally demanding and does not scale well, i.e. with the number of edges. Furthermore, we cannot ensure that Assumption (6.30) holds.

Another approach is based on loopy belief propagation approximation [233], where each $\hat{\pi}_i$ is sequentially updated given the states of its neighbors. This approach is proven to monotonically decrease the objective for each iteration and therefore Assumption (6.30) holds even in the presence of loops. Moreover, in case of tree-like structures, LBPA does converge to the global solution. The algorithm works by iteratively sending messages $M_{ij}(s)$ from node i to node j (in state s) in the proximity of its location:

$$M_{ij}(s) \leftarrow \varepsilon + \max_t \iota_{ij}(s,t) + \vartheta_i(t) + \sum_{k\in N(i)/j} M_{ki}(t)\ ,$$

where ε is some normalization constant, $N(i)$ denotes the set of neighboring nodes of node i and

$$\begin{aligned}\iota_{ij}(s,t) &= (1-\theta)\mathbf{v}_{st},\\ \vartheta_i(t) &= (1-\theta)\langle \mathbf{v}_t, \phi(\mathbf{x}_i)\rangle - \underbrace{\mathbf{1}[i\le n]\frac{\theta}{2}|y_i - \langle \mathbf{u}, \Phi(\mathbf{x}_i, t)\rangle|^2}_{\text{regression part}}\ .\end{aligned} \tag{6.33}$$

After convergence, max-marginals $\mu_i(s)$ can be computed as follows,

$$\mu_i(s) \leftarrow \varepsilon + \max_t \vartheta_i(t) + \sum_{k\in N(i)} M_{ki}(t)\ .$$

Finally, backtracking using the max-marginals reveals the latent states per node. We empirically found that the quadratic approximation performs similar, but it is time-consuming, while the LBP approximation gives a reasonable performance and is scalable.

Parameter Estimation (Eq. (6.29) and Eq. (6.28)) This optimization problem for $\mathbf{v}$ (Eq.(6.29)) is convex and therefore we apply a gradient-based solver with L-BFGS, the method of choice for parameter estimation of CRFs. To perform the gradient descent, we need to compute the objective $\mathcal{L}_{\text{crf}}$, and its gradient with respect to $\mathbf{v}$, which is written as

$$\begin{aligned}\nabla_{\mathbf{v}}\mathcal{L}_{\text{crf}}(\mathbf{v}, \{\pi_i\}_{i=1}^{n+m}) &= \Gamma\mathbf{v} - \Psi(\{\mathbf{x}_i\}_{i=1}^{n+m}, \{\pi_i\}_{i=1}^{n+m})\\ &+\mathbb{E}_{\hat{\pi}\sim p(\{\hat{\pi}_i\}_{i=1}^{n+m}|\{\mathbf{x}_i\}_{i=1}^{n+m},\mathbf{v})}[\Psi(\{\mathbf{x}_i\}_{i=1}^{n+m}, \{\hat{\pi}_i\}_{i=1}^{n+m})].\end{aligned} \tag{6.34}$$

The objective (6.24) contains the partition function $\log Z(\{\mathbf{x}\}_{i=1}^{n+m}, \mathbf{v})$, and the gradient (6.34) involves the expectation

$$\mathbb{E}_{\hat{\pi}\sim p(\{\hat{\pi}_i\}_{i=1}^{n+m}|\{\mathbf{x}_i\}_{i=1}^{n+m},\mathbf{v})}[\Psi(\{\mathbf{x}_i\}_{i=1}^{n+m}, \{\hat{\pi}_i\}_{i=1}^{n+m})].$$

Computation of partition function with pairwise interaction is known to be hard. Therefore, we approximate it with the pseudo-likelihood [234].

The estimation of $\mathbf{u}$, Eq. (6.28), is simply a ridge regression problem, of which the solution is available analytically:

$$\frac{\partial \mathcal{L}_{rr}(\mathbf{u}, \{\pi_i\}_{i=1}^n)}{\partial \mathbf{u}} = 0 \Rightarrow \mathbf{u} = (\lambda I + \Phi\Phi^\top)^{-1}\Phi\mathbf{y} \ ,$$

with $I \in \{0,1\}^{dim\mathcal{H}_1 \times dim\mathcal{H}_1}$ being the identity matrix, $\Phi \in \mathbb{R}^{dim\mathcal{H}_1 \times n}$ the design matrix of only the labeled samples, and $\Phi\Phi^\top$ the corresponding covariance matrix.

One fundamental assumption in our application setting is the linearity of the regression model within each latent state. For this setting, the above regression model is sufficient. It is, however, quite easy to extend to non-linear settings. For that, kernel ridge regression can be applied and solved analytically. Notably, *maximum a posteriori* estimation of the latent states needs to be changed if no expansion of $\mathbf{u}$ can be provided.

Towards One-class Classification Here, we extend the idea to one-class classification. The resulting method, ContextualSVDD, employs results from Section 6.2 and the loopy belief propagation derivations as given in Eq. (6.33). Instead of optimizing the ridge regression problem, we replace it by an slightly modified version of the unconstrained LatentSVDD as given in Eq. (6.3). All we need to do, is to substitute the *local* latent variable $\mathbf{z}$ with our *global* variable π_i for each data point and remove the corresponding minimization,

$$\mathcal{L}_{\text{ContextualSVDD}}(\mathbf{c}, R, \{\pi_i\}_{i=1}^{n+m}) := \ R^2 + C\sum_{i=1}^{n} \max(0, \|\mathbf{c} - \Psi(\mathbf{x}_i, \pi_i)\|^2 - R^2).$$

When latent variables are fixed, the resulting optimization problem is convex. However, this formulation does no longer represent the negative logarithm of some joint distribution and hence, we loose the probability interpretation. To have a fully functional version, inference methods need to be adjusted as well which we do here for the more scalable loopy belief propagation:

$$\vartheta_i(t) = (1-\theta)\langle \mathbf{v}_t, \phi(\mathbf{x}_i)\rangle - \underbrace{\theta C \max(0, \|\mathbf{c} - \Psi(\mathbf{x}_i, \pi_i)\|^2 - R^2)}_{\text{ContextualSVDD part}} \ . \tag{6.35}$$

As can be seen, all examples are considered as long as they receive slack, i.e. lie outside of the hypersphere. Moreover, if labels some are available a semi-supervised extension of ContextualSVDD can be derived using ideas from [12].

6.5 Evaluation and Applications

We test our derived methods (TCRFR and LatentSVDD) on challenging applications from BCI and geoscience.

6.5.1 Extracting Latent Brain States

In many real-world applications, the simplified assumption of independent and identically distributed noise breaks down, and labels can have structured, systematic noise. For example, in brain-computer interface applications, training data is often the result of lengthy experimental sessions, where the attention levels of participants can change over the course of the experiment. In such application cases, structured label noise will cause problems because most ma- chine learning methods assume independent and identically distributed label noise. In this paper, we present a novel methodology for learning and evaluation in presence of systematic label noise.

We are given a data set $\mathcal{D}$ consisting of N data points $\mathbf{x}_1, \ldots, \mathbf{x}_N$, lying in some input space $\mathcal{X}$, and labels $y_1, \ldots, y_N \in \mathcal{Y}$. As mentioned above, we consider a learning scenario where we have varying confidence in the labels (some y_i are more trustworthy than others). To this end, we propose a methodology for learning with non-i.i.d. label noise that consists of four steps.

As a result we obtain a learning methodology that outputs, for a training set $\mathcal{D}$, an inductive rule

$$g_{\mathcal{D}} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}\,,$$

that lets us assign to any pair $(\mathbf{x}, y)$ a denoised label $\widehat{y} := g_{\mathcal{D}}(y)$, which is our guess for the true underlying label.

The various steps of the above methodology are detailed below.

Pipeline The first step is an application of our proposed method, LatentSVDD in Section 6.2.

To remove outliers in the second step, we divide the data set $\mathcal{D}$ into two disjoint sets $L_- := \{\mathbf{x}\colon f(\mathbf{x}) \leq \rho\}$, containing most of the regular data, and $L_+ := \{\mathbf{x}\colon f(\mathbf{x}) > \rho\}$, consisting of the anomalies. Here f is defined as in Eq. (6.2). LatentSVDD provides us with a natural choice of a threshold $\rho = R^2$, but usually we employ a small and thus conservative radius $R << \|\psi(\mathbf{x}, \mathbf{z})\|_\infty$, so that choosing $\rho = R^2$ would be too aggressive (too many anomalies removed). As a remedy, we apply the following procedure to determine a good threshold ρ. Set $f_i := f(\mathbf{x}_i)$ and arrange the f_i in non-decreasing order, $f_{(1)} \leq \ldots \leq f_{(n)}$. Put

$$\rho := \max\left(R^2, \max_{i=1,\ldots,N-1} f_{(i+1)} - f_{(i)}\right).$$

Thus intuitively we determine the threshold where the anomaly score $f(\mathbf{x})$ has the steepest slope. The motivation of which is that regular data is quite densely sampled and thus has a rather smooth increase of anomaly scores, so that choosing an area with steep slope of anomaly scores corresponds to an anomalous region in input space. Indeed we have observed that this heuristic often leads to good results in practice. Finally we output $\mathcal{W} := L_-$ as our (sanitized) working training set.

In the third step, we aim at assigning a label $\widehat{\mathbf{y}}$ for each data point $\mathbf{x}$ using the information from the latent variable $\mathbf{z} \in \mathcal{Z}$, as computed by LatentSVDD. We start by partitioning the working data set $\mathcal{W}$ is into m smaller sets $\mathcal{W}_1, \ldots, \mathcal{W}_m$, where $m := |\mathcal{Z}|$ denotes the cardinality of the latent state space, by grouping all data points that have the same latent state in the LatentSVDD model.

Then, we wish to flip the labels of data points such that the data within each group $\mathcal{W}_i$ has identical labels. To this end, we could simply perform a majority vote within each group. We follow a different, more sophisticated approach here: we determine each group's joint label by choosing the labels such that the working set's kernel-target-alignment (KTA) score is maximized after label assignment.

Kernel target alignment (KTA) [235, 236] is a method that measures the fit between the Gram matrix $K = (\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle)_{1 \leq, i, j \leq n}$ and the label vector $\mathbf{y} = (y_1, \ldots, y_n)$ as follows:

$$\mathrm{KTA}(K, \mathbf{y}) = \frac{\langle K, \mathbf{y}\mathbf{y}^\top \rangle_F}{\|K\|_F \, \|\mathbf{y}\mathbf{y}^\top\|_F}$$

Here, $\langle A, B \rangle_F := \sum_{i,j=1}^n a_{ij} b_{ij}$ denotes the Frobenius inner product and $\|A\|_F := \langle A, A \rangle_F^{1/2}$ denotes its induced norm. This measure has been utilized for optimizing kernels or feature representations [235, 237]. In this paper, we reverse the perspective: instead of optimizing a kernel to match the labels, we optimize the labels to match the kernel.

Let $\mathcal{W} = \mathcal{W}_1 \cup \ldots \cup \mathcal{W}_m$ be the partition of the working training set $\mathcal{W}$ into disjoint sets $\mathcal{W}_i$ such that examples having the same latent state are grouped within the same $\mathcal{W}_i$. Then we compute the denoised label vector $\widehat{\mathbf{y}}$ as

$$\begin{aligned} \widehat{\mathbf{y}} := \underset{\mathbf{y} \in \{+1,-1\}^N}{\operatorname{argmax}} \; & \operatorname{KTA}(K, \mathbf{y}) \\ \text{s.t.} \quad & \forall i, j, k : \mathbf{x}_i, \mathbf{x}_j \in \mathcal{W}_k \Rightarrow y_i = y_j. \end{aligned}$$

Here, the constraints require that all data points within a group $\mathcal{W}_i$ are assigned with the same label. This ensures that we only have to optimize over a few possible label combinations, e.g., over $2^5 = 32$ instead of 2^N, if we have $m = 5$ groups. This renders the optimization problem feasible.

Fair evaluation of learning algorithms for label denoising is a major challenge and the final step in our pipeline: while we cannot trust the observed labels, we usually cannot access the underlying ground truth of an experiment.

When evaluating our experiments on real-world data, we employ three indicators for the prediction accuracy of an algorithm. First, note that it is our intrinsic interest that the accuracy of a classifier increases after denoising the labels. For this purpose we measure the classification performance in terms of the area under the ROC curve (AUC) [238] before and after denoising, and take the difference as an indicator for a algorithm's performance: a good denoising algorithm should yield a substantial higher classification accuracy after denoising. Second, we use kernel-target-alignment scores as an indicator for the fit between labels and data before and after denoising. KTA scores are complementary to AUCs in the sense that capture how well the separability of the data correlates with the labels. Third, we invoke expert opinions to ensure the quality of the delivered solution. This has the advantage that we do not rely on labels in this case, but the disadvantage that the expert opinion is subjective and might be biased. In summary, the combined application of the above described measures lets us obtain a guess for the true performance of a denoising algorithm.

Motivation & Neuroscientific Background We evaluated our proposed learning methodology on the data of an EEG-BCI experiment, for which we recorded 20 participants. The results are presented in this section.

In our EEG experiment, we address the question of whether or not the brain of a participant processed a response error. Conventionally, the EEG data would be analyzed based on the *behavioral* response of the participant, grouping all trials together where the behavioral response is de facto correct or wrong (= behavioral labels). However, having committed a mistake *behaviorally* does not equate having processed it *neurally* [239]. While the *neural* processing is what we are really interested in, these neural labels are unknown, as no ground truth is available. We used LatentSVDD for finding these neural labels in a data-driven way, with the goal of dividing the EEG trials: those where an error was processed neurally, and those where none was processed.

When participants recognize having committed a response error, two specific components are evoked in the event-related potential (ERP) of the EEG signal: an error negativity (N_e) and an error positivity (P_e). Out of these, only the P_e has been attributed to error or post-error processing itself [240]. Therefore, we focus on the P_e in the following, which is characterized by a centro-parietal maximum 200–500ms after feedback [241–244].

Paradigm & Methods In our experiment, 20 participants were asked to perform a fast-paced d2 test [245], a common test of visual selective attention. In this test, participants are presented two types of visual stimuli and are asked to distinguish between these two stimuli

by pressing the corresponding button: the right hand should be used for the target stimulus (20% of trials), the left hand for the non-target stimulus (80% of trials). In total, each participant assessed 300 stimuli under time pressure. Feedback was given 500 ms after each response, both on reaction time and correctness. Brain activity was recorded with multichannel EEG amplifiers (BrainAmp DC by Brain Products, Munich, Germany) with 119 Ag/AgCl electrodes placed according to an extended international 10-10 system, sampled at 1000 Hz and band-pass filtered between 0.05 Hz and 200 Hz.

We examined the neural response that was elicited by receiving feedback. For this, the EEG data was divided into epochs of 500 ms, starting from the onset of feedback. These epochs were baseline corrected (based on the 200 ms interval prior to feedback) and artifact rejection was performed. As features for LatentSVDD and classification, we calculated 9 features per epoch. For this purpose, the interval [0 500 ms] was divided in 10 non-overlapping intervals of 50 ms length. We then calculated the mean signal in each of these intervals and subsequently, the gradient between these means. In order to test class separability, we classified the EEG data using shrinkage LDA, sampling 30 times from the data set and dividing the data set into 75% training data and 25% test data. Classification was run using (a) behavioral labels, (b) the 'neural' labels suggested by LatentSVDD, and, for comparison, those derived by SVDD, LP and RDE. We expect the 'neural' classes to be better separable than before (higher AUC values) and to have a better matching of labels and data (higher KTA scores), compared to using behavioral labels (correct vs. incorrect responses).

Class Re-Assignment and Anomalous Trials On average, LatentSVDD flipped the labels for 35.94% of all trials. This resulted in a *neural* error rate of 31.18%, compared the lower *behavioral* error rate (18.05%). Based on the anomaly score that LatentSVDD returns for each trial, we rejected a small percentage of trials for each participant (cf section 2.2.). For the majority of participants, there are only few trials with high anomaly scores, with a steep drop-off compared to the other trials (cf Figure 6.3). Visual inspection revealed that the results also make sense neuroscientifically: the rejected trials show typical artifacts (eye blinks, voltage drifts with respect to all electrodes or a single electrode) that have escaped the conventional artifact rejection run prior to applying LatentSVDD, as well as trials with unusually high amplitudes.

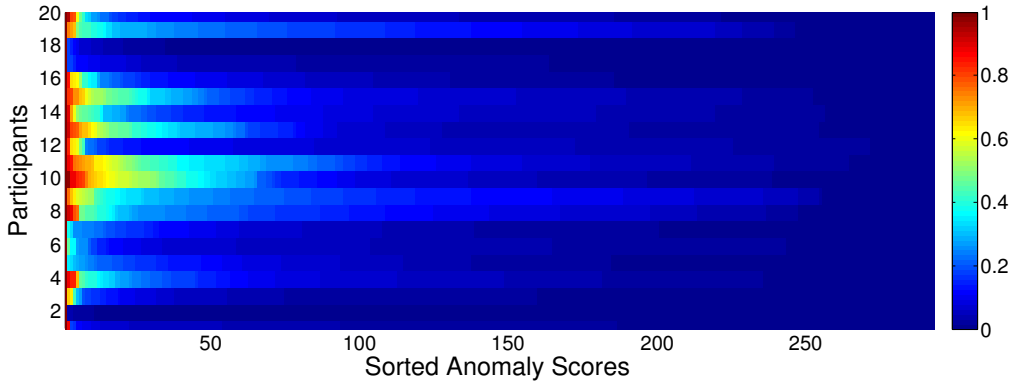

Figure 6.3 – Sorted anomaly scores for each data point of each participant.

Quantitative Assessment We quantified the benefits of LatentSVDD using KTA scores and linear classification (LDA). Both measures confirm that the labels assigned by LatentSVDD allow a much better separation of the data than behavioral labels for all 20 participants. As

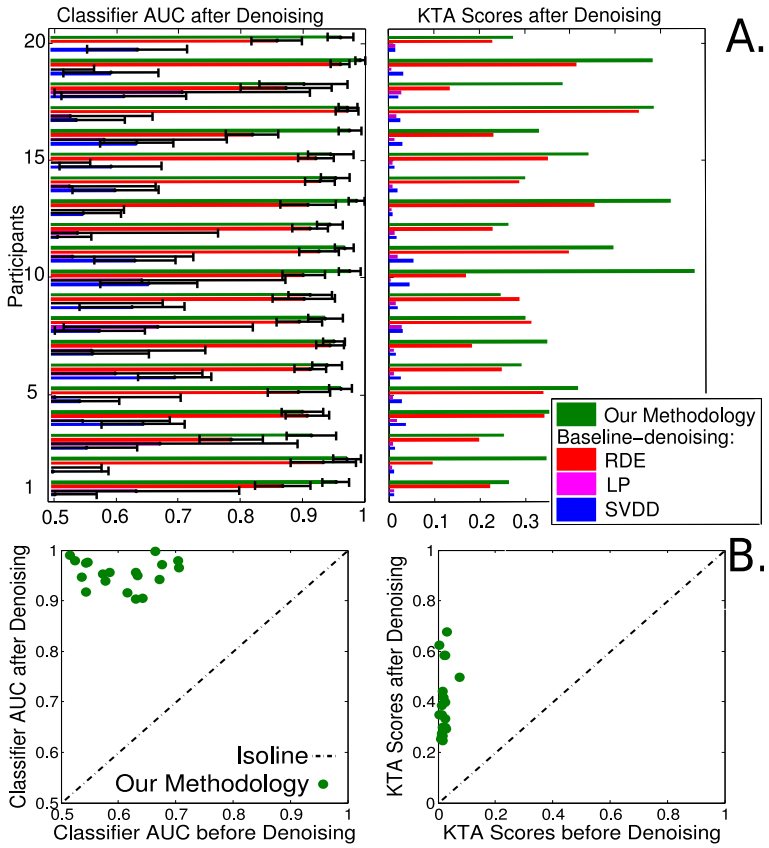

Figure 6.4 – AUC and KTA results for all participants of the experiment.

can be seen in Figure 6.4.B, LatentSVDD renders the classes clearly more distinct from each other, reflected in higher AUC values (0.95 ± 0.02 versus 0.60 ± 0.08). This is accompanied by substantially higher KTA score for all participants. As can be seen in Figure 6.4.A, LatentSVDD is also superior compared to other denoising methods (SVDD, LP, RDE). SVDD and LP lag far behind, both in AUC and KTA scores. In fact, applying these methods even makes separability of classes worse than before (no method: 0.60 ± 0.08, SVDD: 0.59 ± 0.07, LP: 0.54 ± 0.17). In contrast, RDE proves to be a close competitor to LatentSVDD. However, our approach shows better results for this EEG experiment, with a mean AUC score of 0.95 ± 0.02 (RDE: 0.90 ± 0.04) and a mean KTA score of 0.3911 (RDE: 0.2842).

Neuroscientific Assessment While AUC and KTA scores help quantify the positive effect of LatentSVDD, the results are also neurophysiologically sound. In the following, we discuss this for our methodology at the example of participant 5. The different steps of our methodology are visualized in Figure 6.5. Each plot shows the same data (time course at electrode Cz), yet grouped in different classes. The conventional approach is shown on the far left (a), the superior results retained by LatentSVDD on the far right (d), with classes that are clearly better separable. Initially (Figure 6.5(a), classes show great similarity (correct responses in green, erroneous responses in red). Our methodology reveals four latent brain states (Figure 6.5(b)). The state with the highest amplitude (purple) corresponds to typical error processing, with a clear positive component P_e. A clear positivity also occurs in the blue and pink state, yet less pronounced and with different latencies. In contrast, no error has been processed in the black state. Based on the latent variable, a subset of trials is then re-assigned (Figure 6.5(c)). Red and green indicate labels that are retained, orange and light

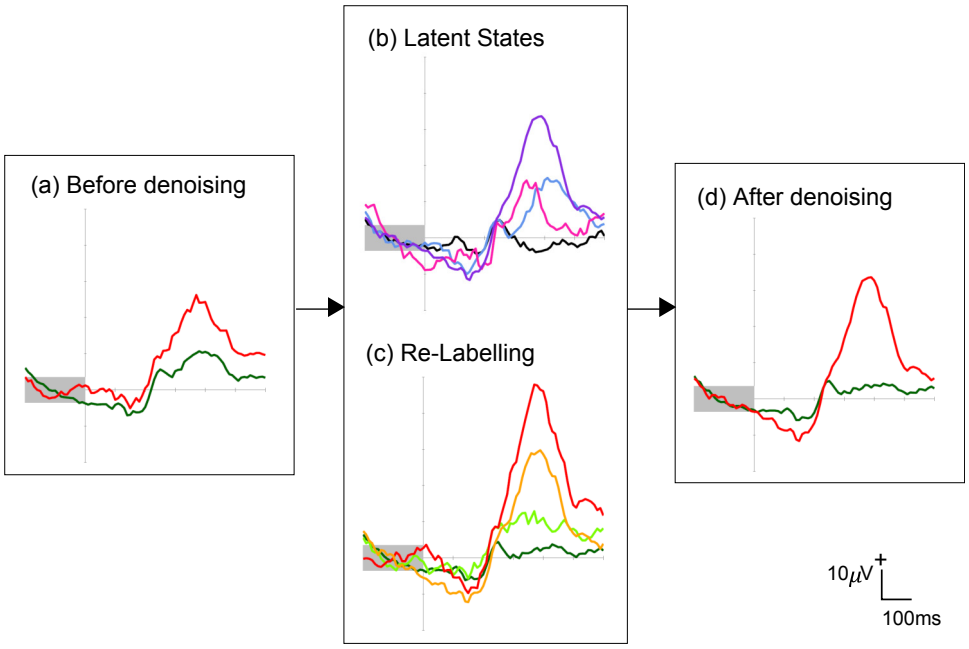

Figure 6.5 – Time course at electrode Cz: (a) before denoising (behavioral labels), (b) latent brain states revealed by LatentSVDD, (c) resulting re-assignment of labels, (d) after denoising.

green signify trials where the labels were switched (orange to red, light green to green). As can be seen, the re-assignment makes sense intuitively. Finally, Figure 6.5(d) shows the denoised data, which reveals a more pronounced error positivity P_e (red) than before. While the latent states themselves are highly subject-specific, we find similar results, i.e. the recovery of a stronger P_e component than before, for all other participants.

Application Outcome Finding the true label for data with systematic, non-i.d.d. label noise is a common challenge in experimental disciplines such as the neurosciences. We proposed a 4-step methodology for learning and evaluation in presence of non-i.i.d. label noise, in the heart of which lies our novel learning algorithm—LATENTSVDD—that allows to capture the hidden state of the label noise. We demonstrate in an extensive case study of EEG-BCI data recorded during an attention test, where we observed that the labels denoised by the proposed methodology lead to substantial better separability of the data (assessed with linear classification; rise in the mean AUC from 0.60 to 0.95 for EEG data). Visual inspection of the data by a domain expert shows that the class assignments output are neurophysiologically plausible, leading to more easily interpretable brain states that subsequently allow for a better and more meaningful experimental evaluation.

6.5.2 Porosity Estimation

Here, we will empirically evaluate our proposed method from Section 6.4, transductive conditional random field regression (TCRFR). First, we will verify various properties using artificially generated data. In a second step, we will apply our method to realistically simulated impedance data where ground truth porosity values are known as well as real data from a Brazilian offshore area.

Controlled Experiment In this section, we assess the various properties of our proposed TCRFR model and compare it against baseline methods in a controlled environment. In all

experiments, we applied cross validation and hyper-parameter tuning on the training samples for all methods with 20 repetitions. The search range for each parameter is shown in Table 6.1.

We evaluate the performance with different criteria: for prediction, we show the mean absolute error (MAE), the mean square error (MSE), the root mean squared error (RMSE), the median absolute error (MDAE), and the R^2-score; for clustering (latent variable estimation) accuracy, we show the adjusted rand score (ARS).

We chose the following as the competitors: ridge regression (RR); support vector regression (SVR) with linear and RBF kernel (SVR RBF) and a Laplacian regularized transductive SVR (with RBF kernel); a RC (Regression and Clustering) approach for assessing latent states by applying k-means and using ridge regression within each cluster (k-means+RR); a mixture of experts approach (MoE) [225, 226];[1] and the transductive regression (TR) [223]. We also plot the lower bounds of the errors, which are the prediction errors under the assumption that the latent variable is known for all the test samples.

Method	Parameter	Range
SVR	C	0.1, 1.0, 10.0, 100.0
	ϵ	1E-0, ..., 1E-5
SVR (RBF)	C	0.1, 1.0, 10.0, 100.0
	ϵ	1E-0, ..., 1E-5
	σ^2	1.0, 0.1, 0.01
LapSVR (RBF)	C	0.1, 1.0, 10.0, 100.0
	ϵ	1E-0, ..., 1E-5
	σ^2	1.0, 0.1, 0.01
	γ_I/γ_A	0.01
RR	σ	1E-6, 1E-5, ..., 1E-1
TR	ϵ	1E-6, 1E-5, ..., 1E-1
	C	0.1, 1.0, 10.0, 100.0
	C'	0.1, 1.0, 10.0, 100.0
MoE (FlexMix)	iter.	2000
	tol.	1E-4, 1E-3, ..., 0.1
k-means RR	ϵ	1E-5, 1E-4, 1E-3
TCRFR	θ	0.75, 0.85, 0.95
	λ	1E-4, 1E-3, 1E-2
	γ	0.1, 1., 10.

Table 6.1 – Optimized hyper-parameters for support vector regression (SVR), Laplacian SVR (LapSVR), ridge regression (RR), transductive regression (TR), mixture of experts (MoE), and our proposed method (TCRFR).

Synthetic structured data was created according to the sequence model illustrated in Fig. 6.6. From 2 latent states with heavily overlapping inputs and additional Gaussian noise, 800 data points were generated. The data was randomly split into training, validation and test data, and the experiment was repeated 20 times. We tested our approach against the baseline methods.

We often face cases where the number of labeled samples is extremely small. In such a case, we found that enhancing the propagation of information from the labeled samples to the unlabeled samples significantly boosts the performance, as shown below. For this purpose, we treat the labeled and the unlabeled samples asymmetrically: the labeled samples are connected to the neighbors lying in a radius-R-near ball, while the unlabeled samples are connected only to the 2 nearest neighbors in the sequence (4 nearest neighbors in the lattice grid). We set $R \sim P^{-1/2}$, where P is the proportion of the labeled samples. This way, we can fix the ratio between the total number of edges between labeled and unlabeled samples and the total number of edges between unlabeled samples. Also, we encourage the ferromagnetic interactions (that is, we favor same states for neighboring latent variables) by reducing the relative regularization parameter for $\mathbf{v}$, i.e., Γ in Eq.(6.22) is diagonal with its elements equal to γ, except the ones corresponding to the ferromagnetic pairwise terms that are equal to 0.01γ. Fig.6.7 shows the performance of TCRFR(R). We can clearly see that optimizing R significantly improves the performance, which supports our strategy.

[1] We use the FlexMix software package.

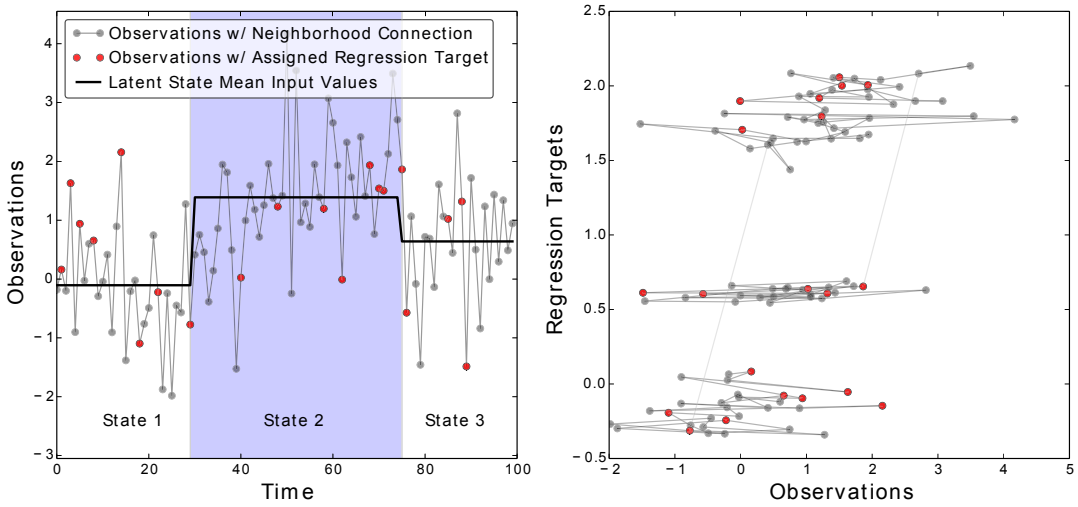

Figure 6.6 – Toy example for structured linear regression problem with few labeled (red) and many unlabeled (gray) data. Left: sequence data (structure: temporal) was generated from three latent states. Right: Considering the input observations (horizontal axis) only, clustering or inferring latent states is futile, whereas harvesting label information (vertical axis), which are available for the red dots, and temporal structure (edges) allows unique clustering.

We compare our two inference schemes, TCRFR-QPA and TCRFR-LBPA, to assess the difference in performance and, specifically, runtime. To achieve a fair comparison (in runtime), both methods share the same parameters and no model selection was done. Fig.6.8 compares the accuracy criteria and the runtime of the two methods for different numbers of samples. Although Fig. 6.9 hints that TCRFR-QPA performs better in settings with low fractions of labeled samples, TCRFR-LBPA has a big advantage in computation time. Therefore, for settings with a small to medium number of data points, and especially a small fraction of labeled data points, TCRFR-QPA should be preferred. However, due to the much larger number of data points in subsequent experiments, we adopt TCRFR-LBPA as our method of choice.

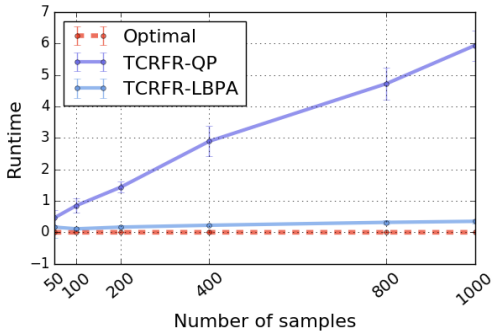

Figure 6.8 – Runtime comparison of our two proposed approximate inference schemes, TCRFR-QPA and TCRFR-LBPA. LBPA proves superior, although the accuracy performance (cf. Fig 6.9) gives a slight advantage to the QPA, which is the better candidate for a smaller number of data points.

Now we assess the accuracy when changing the fraction of labeled data. Figure 6.9 compares the performance of our proposed TCRFR with the baseline methods. We clearly see that our method outperforms all baseline methods under all accuracy criteria, and gives close performance to the lower bound optimal strategy in some cases. Note that the ARS criterion for latent variable estimation is reported only for TCRFR, k-means+RR, and MoE, since the other methods do not provide a latent variable estimator.

RR, SVR, and SVR (RBF) do not consider the dependence of the regression model on the latent state. TR and Laplacian SVR (RBF) consider the transductive setting, but also do not have a latent variable. For this reason, those three methods cannot accurately predict the

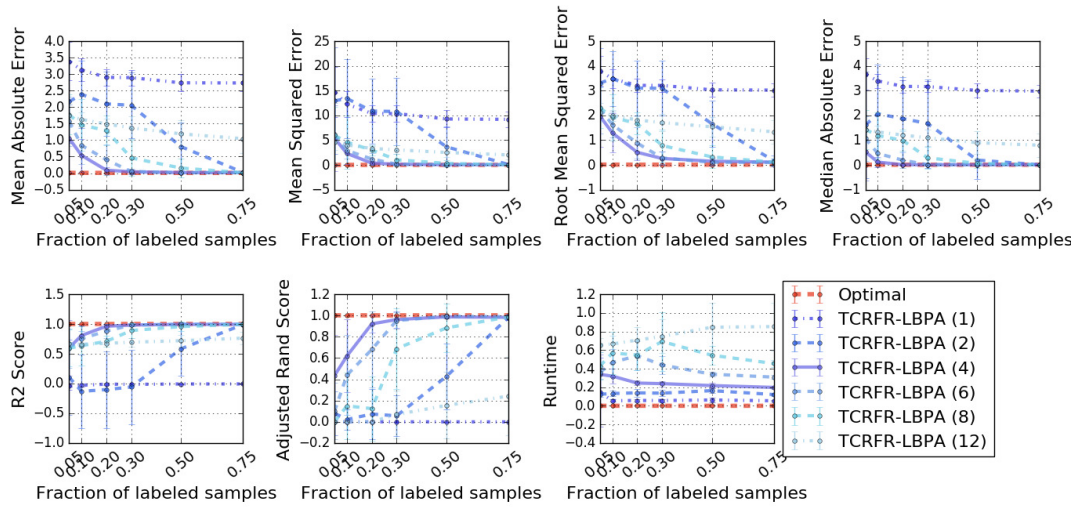

Figure 6.7 – Performance on synthetic structured data. MAE, MSE, RMSE, MDAE, R^2-score, ARS, and the runtime for different fractions of the labeled samples are shown. Here, we assess the quality of the specific grid construction.

labels of unlabeled samples when generated from multiple regression models dependent on the latent state. k-means+RR and MoE consider multiple regression models, depending on the latent variable. However, it doesn't take the structure, i.e., interaction between neighbors, into account. For this reason, they tend to fail to infer the latent states of the unlabeled data, which also results in poor label prediction performance.

Our TCRFR, which performs significantly better than the others, is the only method that identifies multiple regression models from a limited number of labeled samples, and appropriately propagates the label information to the unlabeled data, by capturing the structure of latent variables.

The number of assumed latent states is a crucial parameter. Here, we examine the impact of choosing latent states differing from the ground truth for all methods that are sensitive to this parameter (our TCRFR, K-means+ridge regression, and the Mixture-of-Experts) and ridge regression as a "calibration". Figure 6.10 shows the dependence of the performance on the assumed number of latent states. We can see that the performance of TCRFR is not very sensitive to the assumed number of latent states, as long as it is larger than the true number of latent states (2 in this dataset). This is because the redundant components tend to be discarded if the regularization coefficient γ is optimized.

Porosity Prediction Porosity estimation is a crucial step in the analysis of petroleum reservoirs for the oil industry. Although estimating porosity from seismic impedance is less accurate than from drilled wells [19], plenty of measurements are available, typically on a 3D grid covering over tens of square kilometers. The left panel of Figure 6.11 shows an example of a seismic impedance data horizontal slice [246].

As stated earlier, the correlation between seismic impedance and porosity depends on bodies (or units) of rock known as *facies* [247]. The segmentation of the reservoir into facies allows local heterogeneity and strong contrasts in rock properties to be preserved between different geological layers [248]. The middle panel in Figure 6.11 shows the facies pattern of the same data, adapted from [246]. In many cases, facies classification is carried out by hand, based on the data available from seismic surveys, well logs, and collected core samples. For automation, cell-based geostatistical modeling, object-based stochastic modeling [247], *k*-means [64] or Mixture of Gaussians [249] are often applied. Porosity estimation is then

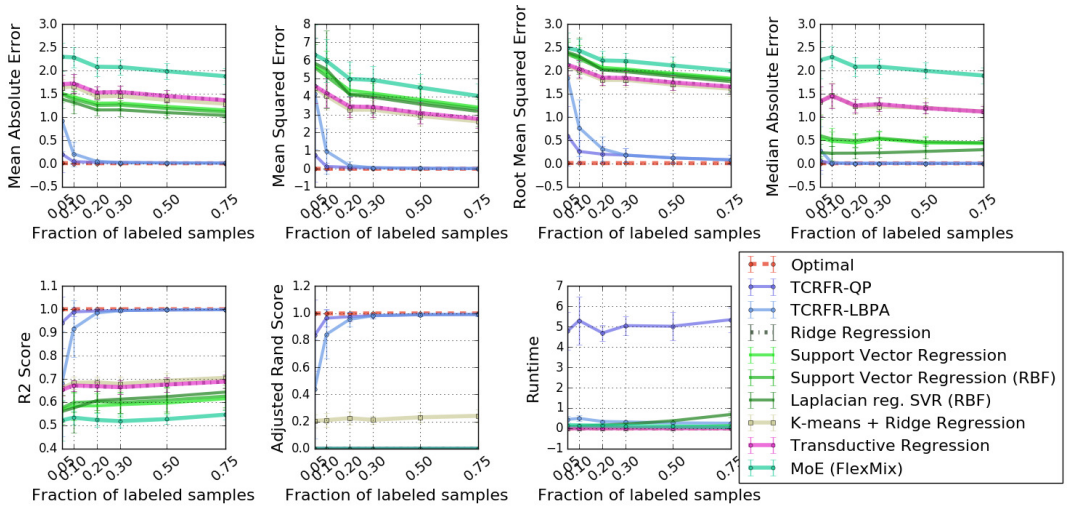

Figure 6.9 – Performance on synthetic structured data for varying fractions of the labeled samples.

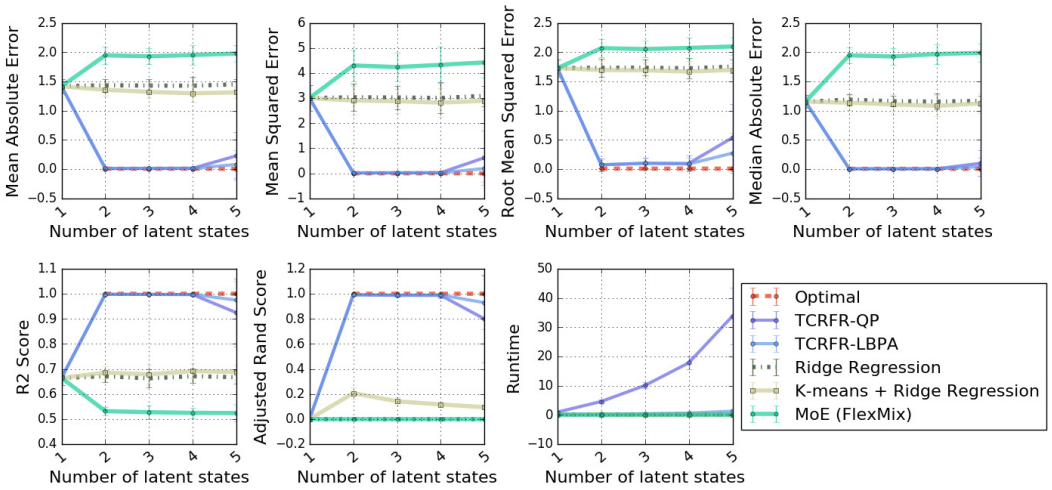

Figure 6.10 – Performance on synthetic structured data for a varying number of maximum latent states.

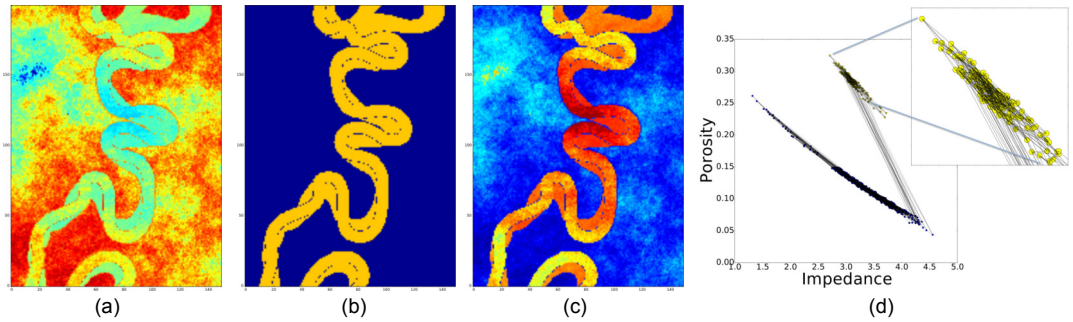

Figure 6.11 – Porosity prediction problem. The goal is to estimate (c) porosity (unknown at most of the locations) from (a) impedance (known) by using a linear relationship between them. However, this relationship depends on the (b) facies (unknown), and accurate facies estimation requires porosity measurements because of the overlapped marginal distribution of the impedance (d).

Method	MAE	MSE	RMSE	MDAE	R2
MoE	2.38477	8.44310	2.90562	1.57930	0.47237
k-means+RR	2.08030	6.27532	2.50489	1.93901	0.61407
SVR	1.84235	11.37484	3.37256	0.24478	0.28910
RR	2.05989	6.19819	2.48950	1.89004	0.61271
TR	2.05993	6.19791	2.48944	1.89106	0.61273
TCRFR	**0.69878**	**3.55215**	**1.88422**	**0.14865**	**0.77804**
L. bound	0.15237	0.03567	0.18885	0.13740	0.99777

Table 6.3 – Performance on synthetic seismic data for 5% of labeled data.

performed within each facies usually by *kriging* [250], an interpolation method for spatial data based on Gaussian processes commonly used in geostatistics [251]. This whole process is extremely time-consuming and requires the specialized knowledge of a geologist (see Appendix B).

In the following subsections, we show the performance of TCRFR and the baseline competitors on synthetic and real data. In all experiments, we applied 3-fold cross validation on the training samples to tune the hyper-parameters for all methods. The search range for each parameter is shown in Table 6.2.

Method	Parameter	Range
MoE	iter.	300, 400, ..., 800
	tol.	1E-4, 1E-3, ..., 0.1
KMRR	ϵ	1E-5, 1E-4, 1E-3
SVR	C	1E-3, 1E-2, ..., 1.
	ϵ	0.1, 1., 10.
	kernel	linear
RR	tol.	1E-6, 1E-5, ..., 0.1
TR	ϵ	1E-6, 1E-5, 1E-4
	C	10., 100., ..., 1E4
	C'	0.001, 0.01, ..., 1
TCRFR	R	3, 4, ..., 8
	θ	0.7, 0.75, ..., 1.0
	λ	1E-4, 1E-3, 1E-2
	γ	0.1, 1., 10.

Table 6.2 – Optimized hyperparameters in the porosity prediction experiment.

Synthetic Seismic Data We use the synthetic 3D reservoir benchmark data set [246] ($150 \times 200 \times 40$ voxels), which was created through realistic geological modeling. Figure 6.11 shows one horizontal slice of the data with 150×200 voxels. There are two facies, the sand channels (yellow in Fig.6.11 (b)) and the background shale (blue). From the data, we observe the following trend: the sand channels have higher porosity (Fig.6.11 (c)) than the background shale, and the impedance (Fig.6.11 (a)) has a negative correlation with porosity (see also Fig.6.11 (d)). Due to the vertical low resolution during the seismic acquisition process [247], we simplify our setting by only considering connections in the horizontal slices. So, from each of those volumes, we extract 150×200 horizontal slices, and assume that the whole impedance data and part of the porosity data are available as the input and the regression label (output), respectively. Our goal is to infer the latent structure (facies), and to predict the porosities at the unlabeled samples.

Among the $150 \times 200 = 30,000$ pixels, we randomly choose 5% of them as labeled samples, and the others are treated as unlabeled samples. We iterate this process 10 times and report the average performance.

Table 6.3 summarizes the performance of TCRFR and the baseline methods. A clear advantage of TCRFR is found. To discuss the reason of the success of TCRFR, we show the

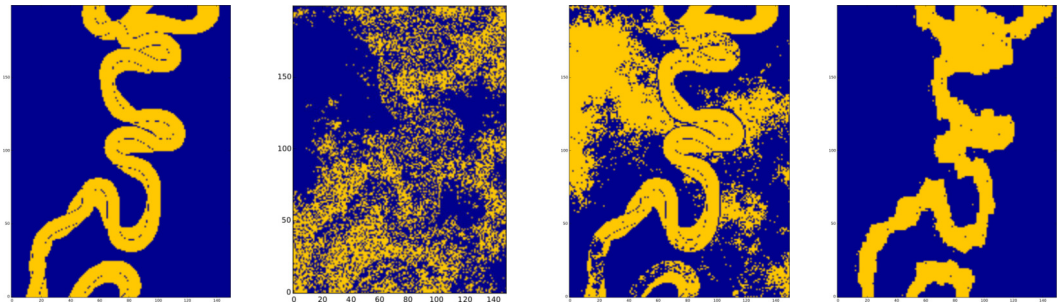

Figure 6.12 – Facies estimation results for 5% of labeled examples.

estimated facies and the predicted porosity for a single trial in Figure 6.12 and Figure 6.13, respectively.

Figure 6.12 implies that TCRFR successfully recovers the facies structure, while MoE and k-means+RR fail. The excellent facies estimation by TCRFR, despite the small fraction of labeled data, is because it acquires the facies structure with adequate strength of correlation between neighbors, through the learning process of conditional random field. This enables appropriate propagation of the label information, which is necessary for good facies estimation from only 5% of labeled samples. On the other hand, MoE and k-means+RR are not capable of taking the structure of facies into account. Therefore, although equipped with multiple regression models for each facies, they fail to identify the facies of the unlabeled samples, because no information is propagated from labeled samples. In fact, we found that the facies estimation by MoE is accurate on the labeled samples, and the bad performance is only on the unlabeled samples.

Thanks to the high quality of facies estimation, TCRFR provides a significantly better porosity estimation result, as shown in Figure 6.13. SVR, RR, and TR are not capable of dealing with multiple regression models and, therefore, do not perform as well as our TCRFR method. Also note that they do not provide facies estimation results.

Figure 6.14 shows MAE, RMSE, and MDAE for a range of labeled samples fractions. For any fraction in this range, our TCRFR outperforms all state-of-the-art competitors, which again proves a clear advantage of our approach.

Last, Figure 6.15 shows the facies estimation results (top) and the porosity prediction results (bottom) by TCRFR for different fractions of labeled samples. Notably, although degradation is observed to some extent, TCRFR still provides reasonable facies estimation and porosity prediction, even if only $1 \sim 2\%$ of labeled samples are available. In fact, $1 \sim 2\%$ is still high for the porosity prediction application—we should assume an extremely small number of labeled samples available only at the drilled wells. Nevertheless, we see the current research as a good starting point, and will further improve our method by using domain knowledge and other heuristics to cope with fewer labeled samples.

Real Data Experiment We apply our TCRFR to a real petroleum reservoir, located in the offshore coast of Brazil. It covers an area of approximately 100 square kilometers, with 460 meters in depth. The data in this region comprises a 3D volume with $313 \times 549 \times 74$ voxels containing acoustic impedance samples. This data contains *truly* labeled data from only three wells, with which no *general-purpose* machine learning method can cope. Accordingly, we use additional labeled samples, which were created by geoscientists through a handcrafted procedure (see Appendix B.2 for details).

Table 6.4 shows the performance of TCRFR and the baseline methods on the real data for 5% of labeled samples (including additional handcrafted labels). Similarly to the experiment

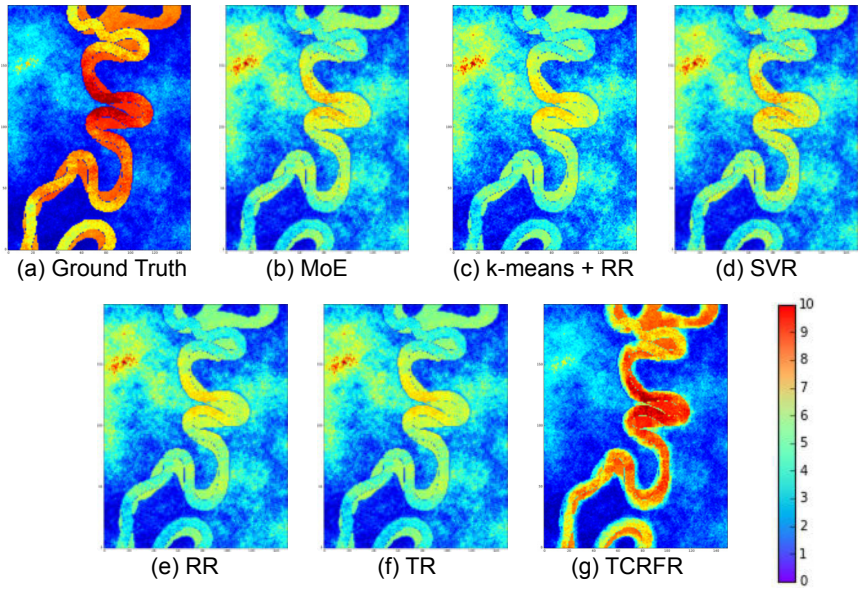

Figure 6.13 – Porosity prediction results for 5% of labeled data.

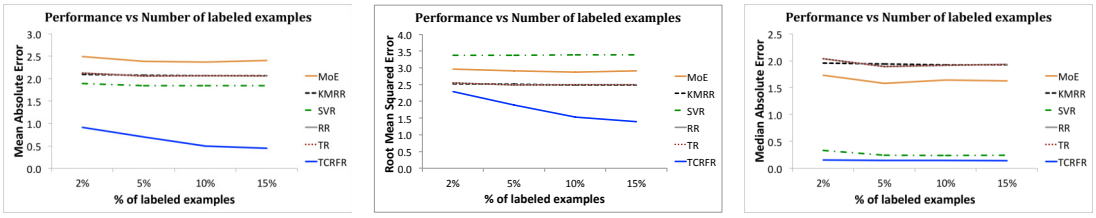

Figure 6.14 – MAE, RMSE, and MDAE on synthetic seismic data for a range of labeled data fractions.

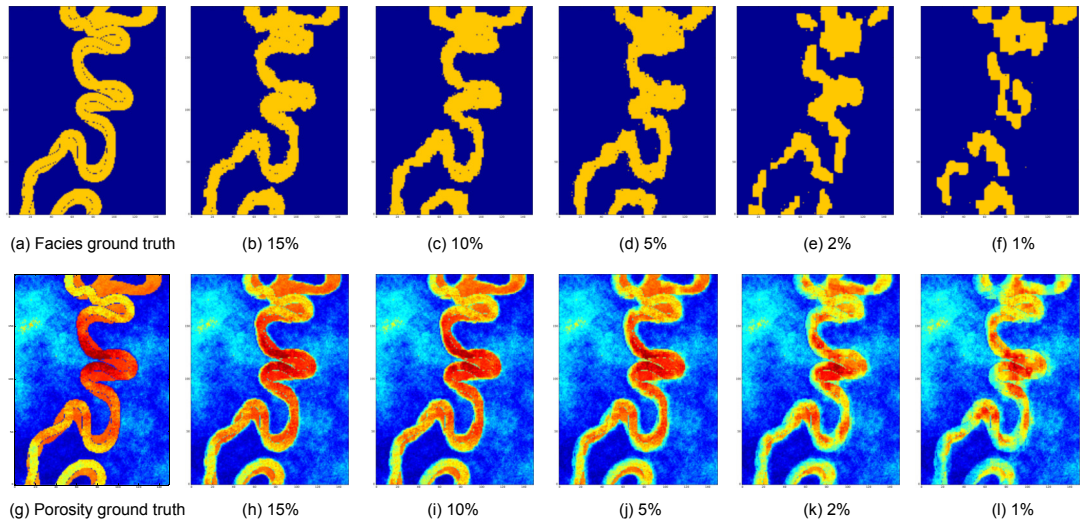

Figure 6.15 – Estimated facies and the predicted porosity by TCRFR for different fractions of labeled samples.

Method	MAE	MSE	RMSE	MDAE	R2
MoE	0.42502	0.55195	0.74268	0.22591	0.88991
k-means+RR	0.45002	0.44259	0.66513	0.28474	0.90910
SVR	0.48028	0.46350	0.68055	0.35463	0.90757
RR	0.45716	0.45581	0.67490	0.28999	0.90909
TR	0.45717	0.45581	0.67490	0.29000	0.90909
TCRFR	**0.24225**	**0.13712**	**0.37001**	**0.14571**	**0.97264**

Table 6.4 – Porosity prediction performance on the real data with 5% of labeled examples.

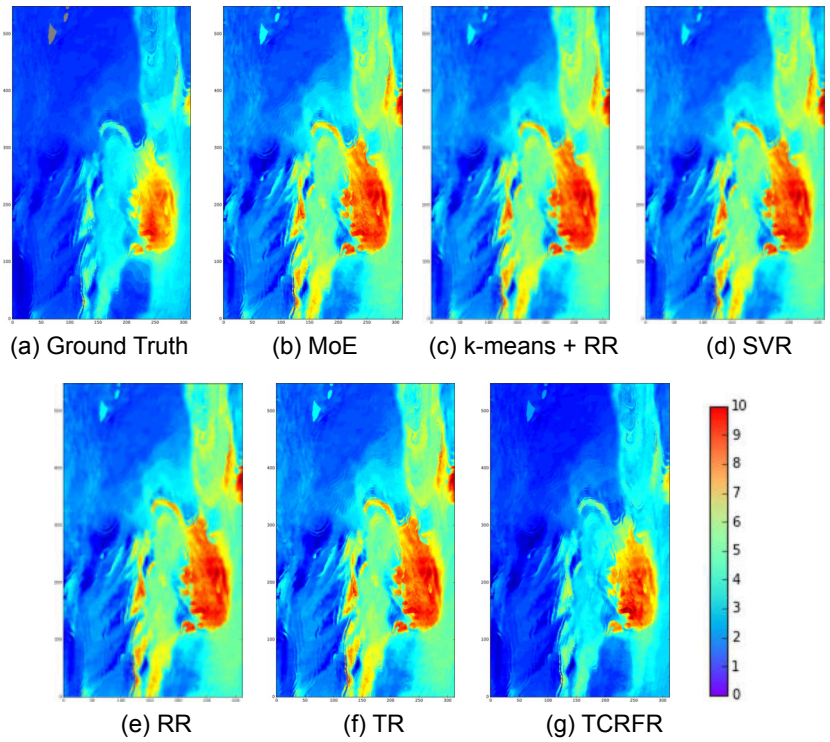

Figure 6.16 – Predicted porosity on the real data.

on synthetic data in the previous subsection, our TCRFR compares highly favorably with the baselines.

Figure 6.16 shows the predicted porosity by TCRFR and the baseline methods. Note that the ground truth here is not the *true* labels available only at the wells, but the *additional* handcrafted labels. Again, TCRFR provides excellent results and allows a first and useful assessment of geologically attractive regions for oil exploration (red and yellow regions).

Application Outcome Handling data under spatio-temporal structure with limited labels and, therefore, a combination of certainty and vast uncertainty requires novel robust modeling strategies. We tackled this challenging problem in time-series analysis and porosity prediction for the oil industry. Experiments on toy time-series data and synthetic porosity prediction data clearly showed successful inference. Finally, we have studied real world data from an offshore oil field and could show remarkable performance of our new model, which compares very favorably with the state-of-the-art competitors.

6.6 Summary and Discussion

In this chapter, we proposed extensions of the support vector data description to cope with contextual anomalies. We focused on the specific setting of latent class dependencies. Our first contribution—LatentSVDD—leveraged joint feature maps to incorporate latent classes into the objective function. While the notion of joint feature maps allows great flexibility beyond latent classes (cf. Chapter 5), by deriving a restricted version—ClusterSVDD—and reviewing rigorously its properties, we were able to show that k-means is contained as a special case. Finally, we imposed structural dependencies among the latent variables itself, effectively relating the input samples. Although the proposed method—TCRFR—was derived with a regression setting in mind, we discussed an extension to one-class classification—ContextualSVDD. Extensive applications from neurosciences and geosciences display the benefits of the proposed solutions.

Although the applications draw a positive picture of our developed methods, various limitations exist. Besides the fact that all of the presented methods are non-convex extensions of convex base models, all of them are also computationally much more demanding.

Limits of LatentSVDD LatentSVDD with the proposed joint feature map will tend to join cluster with small sample sizes which is the reason why some cluster remain empty our application in Section 6.5.1. While this might seem like a huge benefit, there is unfortunately no straightforward way of controlling this behavior. Furthermore while the joint feature map adds a lot of flexibility to encode latent feature space and input space, it still remains an open question how to leverage this in real-world applications. Finally, the proposed formulation can not leverage the flexibility of kernels.

Limits of ClusterSVDD ClusterSVDD on the other hand does not show such behavior as the concentration of cluster and can be seen as a k-means variant with inherent anomaly detection. However, there is no clear rationale that each cluster should assume a fraction of ν outliers and hence, the resulting clustering and anomaly detection becomes less interpretable. Moreover, if setting $\nu = 1$ the k-means algorithm is recovered and optimal solutions can be calculated analytically while for any other setting a much more computationally demanding quadratic problem needs to be solved.

Limits of TCRFR The most severe limitation that our TCRFR method faces is of computational nature. Especially demanding are the inference steps for large number of nodes and edges and, i.e. the calculation of the partition function. Even though fast (and crude) approximations of the partition function are employed, due to the pure number of function calls, calculations will take the bulk of the time necessary to converge. The impact of the approximation quality on the overall solution needs to be examined more closely.

Source code and resources for the proposed methods are available on github [a] [b]. Parts of this chapter are based on:

Görnitz, N., Porbadnigk, A. K., Kloft, M., Binder, A., Sannelli, C., Braun, M., Müller, K.-R., "When brain and behavior disagree: A novel ML approach for handling systematic label noise in EEG data", in *Machine Learning and Interpretation in Neuroimaging Workshop (MLINI)*, 2013

Görnitz, N., Porbadnigk, A. K., Binder, A., Sanelli, C., Braun, M., Müller, K.-R., Kloft, M., "Learning and Evaluation in Presence of Non-i.i.d. Label Noise", in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 33, 2014, pp. 293–302

Porbadnigk, A. K., **Görnitz, N.**, Sannelli, C., Binder, A., Braun, M., Kloft, M., Müller, K.-R., "When Brain and Behavior Disagree: Tackling systematic label noise in EEG data with Machine Learning", in *IEEE International Winter Workshop on Brain-Computer Interface (BCI)*, 2014

Porbadnigk, A. K., **Görnitz, N.**, Sannelli, C., Binder, A., Braun, M., Kloft, M., Müller, K.-R., "Extracting latent brain states — Towards true labels in cognitive neuroscience experiments", *NeuroImage*, vol. 120, pp. 225–253, 2015

Görnitz, N., Lima, L. A., Varella, L. E., Müller, K.-R., Nakajima, S., "Transductive Regression for Data with Latent Dependency Structure", *IEEE Transactions on Neural Networks and Learning (TNNLS)*, 2017

Görnitz, N., Lima, L. A., Müller, K.-R., Kloft, M., Nakajima, S., "Support vector data descriptions and k-means clustering: one class?", *IEEE Transactions on Neural Networks and Learning (TNNLS)*, 2017

Lima, L. A., **Görnitz, N.**, Varella, L. E., Vellasco, M., Müller, K.-R., Nakajima, S., "Porosity Estimation by Semi-supervised Learning with Sparsely Available Labeled Samples", *Computers & Geosciences*, vol. 106, pp. 33–48, 2017

[a] https://github.com/nicococo/tilitools
[b] https://github.com/nicococo/niidbox

Chapter 7
Conclusions

> The world exploded into a whirling network of kinships, where everything pointed to everything else, everything explained everything else.
>
> Umberto Eco (Foucault's Pendulum)

We have addressed the central research question of how to tie together various information, such as labels, dependency structure, sparseness, to obtain better anomaly detection models for the three different classes of anomalies. To that end, we have presented extensions to the one-class SVM as well as the related support vector data description (SVDD) to incorporate various kinds of side information, i.e. dependency structure. We showed for artificially generated data as well as for a variety of real-world applications that incorporating side information does help to increase detection performance when compared with the respective base models. In detail, we presented the following extensions in the spirit of the one-class classification paradigm:

Point Anomalies Assuming that anomalies are scarce and occur independently of each other, methods for controlling the sparsity of the found solutions in terms of single independent features (Semi-supervised ℓ_p-norm regularized one-class SVM) and groups of features (Semi-supervised ℓ_p-norm regularized multiple kernel learning one-class SVM) have been derived.

Collective Anomalies In this scenario anomalies are assumed to appear as groups of measurements instead of single entries. Techniques from structured output learning have been (i) extended to cope with large-scale problems (Bundle Methods Optimization for SSVM), (ii) employed to derive an unsupervised anomaly detector (Latent Structure Anomaly Detector) for groups of measurements that exhibit a latent dependency structure.

Contextual Anomalies Anomalies appear only in specific contexts and are supposed to carry two signals that contain behavioral and contextual information. Contributions in this scenario consider latent class dependencies and are threefold: (i) we derived a method capable of detecting contextual anomalies (LatentSVDD), (ii) theoretical insight reveal *k*-means as a special case (ClusterSVDD), and (iii) a method for learning with latent class dependencies when an additional structure is imposed on the latent variables (TCRFR for regression and ContextualSVDD for anomaly detection).

However, we would like to emphasize that extending the basic model to incorporate side information such as data dependency structure is no silver bullet, it comes with higher complexity and thus, more possibilities to fail. In many cases we had to give up on desired properties such as convexity in order to derive a solution.

A research direction of growing importance, not only for anomaly detection, will be the interpretability of complex methods [252–258]. Explanations of single decisions, models, or data sets greatly helps the acceptance of those models in application domains. Furthermore, they can reveal, in a way a human can understand, important information that otherwise might have stayed cloaked.

In a broader perspective, solving complicated real-world problem requires taking every single bit of information into account and no other technique than deep learning has been more successful at this and transformed machine learning more in the recent years. Now, in order to be successful those methods need large scale data and corresponding labels. However, there have important attempts towards unsupervised and semi-supervised extensions. Most importantly, generative adversarial networks (GANs) [259] and (variational) auto encoders. Unsupervised learning of concise, meaningful and interpretable feature descriptions is the holy grail of anomaly detection. There have been attempts at combining techniques for deep learning and one-class classification [91] with promising results. However, there are a number of complex technical details that need to be solved in order to avoid trivial solutions, i.e. manifold collapse.

Appendix A

Learning with Structured Data

A.1 Proofs of Results in Section 5.3

We show the equivalence of (5.10) and (P) for loss $l(t) = \max(0, t)$.

Proof of equivalence of (5.10) *and* (P) *for* $l(t) = \max(0, t)$. First note that for loss $l(t) = \max(0, t)$ the problem (5.10) becomes the structured one-class SVM problem (P′) from Section 5.3. To see that (5.10) is equivalent to (P′), we employ a variable substitution $\tilde{\mathbf{w}} := \mathbf{w}/\rho^*$ in (5.10). This yields

$$\begin{aligned} \text{Eq. (P}') \;=\; -\rho^* + \rho^* \min_{\tilde{\mathbf{w}} \in \mathcal{H}} \Bigg(\frac{1}{2} \|\tilde{\mathbf{w}}\|^2 \\ + \frac{1}{\nu n} \sum_{i=1}^{n} \max \left(0, 1 - \max_{z \in \mathcal{Z}} \langle \tilde{\mathbf{w}}, \Psi(x_i, z) \rangle + \delta(z)' \right) \Bigg), \end{aligned} \tag{A.1}$$

where $\delta(z)' = \delta(z)/\rho^*$ and ρ^* is optimal in (P′). Thus, in order to solve (A.1) (and thus (P′)), it is sufficient to solve

$$\begin{aligned} \min_{\mathbf{w} \in \mathcal{H}} \; \frac{1}{2} \|\mathbf{w}\|^2 + \frac{1}{\nu n} \sum_{i=1}^{n} \max \Big(0, 1 \\ - \max_{z \in \mathcal{Z}} \langle \mathbf{w}, \Psi(x_i, z) \rangle + \delta(z) \Big). \end{aligned} \tag{A.2}$$

By Lemma 5 below, for each $\nu \in]0, 1]$, there exists a $C > 0$ such that (A.2) is, indeed, equivalent to (5.10). □

Lemma 5. *Let* $D \subset \mathbb{R}^d$ *be a set, let* $f, g : D \to \mathbb{R}$ *be arbitrary functions. Consider the optimization tasks*

$$\min_{\mathbf{x} \in D} \quad f(\mathbf{x}) + \sigma g(\mathbf{x}), \tag{A.3}$$

$$\min_{\mathbf{x} \in D : g(\mathbf{x}) \leq \tau} \quad f(\mathbf{x}). \tag{A.4}$$

Assume that the minima exist. Then we have that for each $\sigma > 0$ *there exists* $\tau > 0$ *such that OP* (A.3) *is equivalent to OP* (A.4)*, that is, each optimal solution* x^* *of one is an optimal solution of the other, and vice versa.*

Proof. The proof is similar to the one of Proposition 12 in [105]. Let be $\sigma > 0$ and $\mathbf{x}^*$ be the optimal of (A.3). We have to show that there exists a $\tau > 0$ such that $\mathbf{x}^*$ is optimal in (A.4).

We set $\tau = g(\mathbf{x}^*)$. Suppose $\mathbf{x}^*$ is not optimal in (A.4), that is, it exists $\tilde{\mathbf{x}} \in D : g(\tilde{\mathbf{x}}) \leq \tau$ such that $f(\tilde{\mathbf{x}}) < f(\mathbf{x}^*)$. Then we have

$$f(\tilde{\mathbf{x}}) + \sigma g(\tilde{\mathbf{x}}) < f(\mathbf{x}^*) + \sigma\tau,$$

which by $\tau = g(\mathbf{x}^*)$ translates to

$$f(\tilde{\mathbf{x}}) + \sigma g(\tilde{\mathbf{x}}) < f(\mathbf{x}^*) + \sigma g(\mathbf{x}^*).$$

This contradicts the optimality of $\mathbf{x}^*$ in (A.3), and hence shows that $\mathbf{x}^*$ is optimal in (A.4), which was to be shown.

□

Proof of Theorem 8. By [200] we have that, if l is L-Lipschitz and ranges in $[0, D]$, with probability at least $1 - \epsilon$ over the draw of the sample,

$$E\, l(\hat{f}) - E\, l(f^*) \leq 8LR_n(\mathcal{F}) + \frac{l(0)}{n} + D\sqrt{\frac{2\log(2/\epsilon)}{n}}, \tag{A.5}$$

where $R_n(\mathcal{F}) := \mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)$ is the *Rademacher complexity* of the class $\mathcal{F}$ and $\sigma_1, \ldots, \sigma_n$ denote i.i.d. Rademacher variables (random signs). For many learning algorithms $R_n(\mathcal{G})$ is of the order $O(1/\sqrt{n})$, when employing appropriate regularization, and thus so is (A.5). We will show that also the latent anomaly detection method of (5.10) enjoys this favorable rate, too: By definition of the Rademacher complexity of $\mathcal{F}$,

$$\begin{aligned}
R_n(\mathcal{F}) \;&=\; E \max_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i) \\
&=\; E \max_{\mathbf{w}\in\mathcal{H}:\|\mathbf{w}\|\leq C} \frac{1}{n}\sum_{i=1}^n \sigma_i \Big(1 \\
&\qquad - \max_{z\in\mathcal{Z}} \big(\langle \mathbf{w}, \Psi(X_i, z)\rangle + \delta(z)\big)\Big) \\
&=\; \underbrace{\Big(1 + \max_{z\in\mathcal{Z}} |\delta(z)|\Big)\, E\left[\left|\frac{1}{n}\sum_{i=1}^n \sigma_i\right|\right]}_{(*)} \\
&\quad +\; \underbrace{E \max_{\mathbf{w}\in\mathcal{H}:\|\mathbf{w}\|\leq C} \frac{1}{n}\sum_{i=1}^n \sigma_i \max_{z\in\mathcal{Z}} \langle \mathbf{w}, \Psi(X_i, z)\rangle}_{(**)}
\end{aligned}$$

We bound the two summands in the above expression separately: on one hand, by Jensen's inequality, $E\big|\frac{1}{n}\sum_{i=1}^n \sigma_i\big| \leq \sqrt{E\frac{1}{n^2}\sum_{i,j=1}^n \sigma_i\sigma_j} = \frac{1}{\sqrt{n}}$ because $E\sigma_i\sigma_j = 0$ when $i \neq j$, which shows $(*) \leq \frac{1+A}{\sqrt{n}}$. To bound the second summand, note that $(**) \leq R_n(\mathcal{F}')$ with $\mathcal{F}'$ defined as $\mathcal{F}' := \big\{f_{\mathbf{w}} = \big(x \mapsto \max_{z\in\mathcal{Z}} \langle \mathbf{w}, \Psi(x, z)\rangle\big) : \|\mathbf{w}\| \leq C\big\}$. Furthermore put $\mathcal{F}'' := \big\{f_{\mathbf{w}} = \big(x \mapsto \max_{z\in\mathcal{Z}} f_z\big) : f_z \in \mathcal{F}_z, z \in \mathcal{Z}\big\}$ and $\mathcal{F}_z := \big\{f_{\mathbf{w}} = \big(x \mapsto \langle \mathbf{w}, \Psi(x, z)\rangle\big) : \|\mathbf{w}\| \leq C\big\}$. Clearly, $\mathcal{F}' \subset \mathcal{F}''$ and thus $R_n(\mathcal{F}') \leq R_n(\mathcal{F}'')$. By Lemma 8 in the supplemental material, $R_n(\mathcal{F}'')$ is itself bounded by $R_n(\mathcal{F}'') \leq \sum_{z\in\mathcal{Z}} R_n(\mathcal{F}_z)$, and the terms $R_n(\mathcal{F}_z)$, for each $z \in \mathcal{Z}$ are known from [200] to be bounded as $R_n(\mathcal{F}_z) \leq \frac{B}{\sqrt{n}}$.[1] This shows $(**) \leq \frac{BC|\mathcal{Z}|}{\sqrt{n}}$. The result is then obtained from (A.5) by noting, that D can be chosen as $D := L(1+A+BC)$. □

[1] Again this quickly follows from Jensen's inequality because $E\sigma_i\sigma_j = 0$ when $i \neq j$.

In the proof of Theorem 8 above, we use the following result.

Lemma 6 (Lemma 8.1 in [201]). *Let $\mathcal{F}_1, \ldots, \mathcal{F}_l$ be sets of functions $f : \mathcal{X} \to \mathbb{R}$, and let $\mathcal{F} := \{\max(f_1, \ldots, f_l\} : f_i \in \mathcal{F}_i, i \in \{1, \ldots, l\}\}$. Then,*

$$R_n(\mathcal{F}) \leq \sum_{j=1}^{l} R_n(\mathcal{F}_j).$$

Sketch of proof [201]. The idea of the proof is to write $\max(h_1, h_2) = \frac{1}{2}(h_1 + h_2 + |h_1 - h_2|)$, and then to show that

$$\mathbb{E}\left[\sup_{h_1 \in \mathcal{F}_1, h_2 \in \mathcal{F}_2} \frac{1}{n} \sum_{i=1}^{n} |h_1(x_i) - h_2(x_i)|\right] \leq R_n(\mathcal{F}_1) + R_n(\mathcal{F}_2).$$

This proof technique also generalizes to $l > 2$. For the complete proof see Section 8 in [201]. □

A.2 Proofs of Results in Section 5.3 II

Proof of Theorem 11. First observe that it holds $\alpha_i^* \max(0, f(x_i)) = 0$ for all $i = 1, \ldots, n$ in the optimal point of the Lagrangian saddle point problem.[2] This implies that we have $f(x_i) \leq 0$ if x_i is a *support vector* (that is, $\alpha_i^* > 0$) [33, 196]. Since $\sum_{i=1}^{n} \alpha_i^* = 1$ and $\alpha_i^* \leq \frac{1}{\nu n}$ there must at least $\lceil \nu n \rceil$ many such points (the function $\lceil \cdot \rceil$ rounds a real number up to the next large integer). Hence there can be no more than $n - \lfloor \nu n \rfloor$ many points with $f(x_i) > 0$, which corresponds to a fraction of $\frac{n - \lfloor \nu n \rfloor}{n} \leq 1 - \nu$, and thus shows the assertion (b). Next observe that if we have $f(x_i) < 0$ then $\alpha_i^* = \frac{1}{\nu n}$ (to see this, note that if $\alpha_i^* < \frac{1}{\nu n}$ we could increase the objective of the Lagrangian by increasing α_i^*, which would contradict the optimality of α_i^*). Since $\sum_{i=1}^{n} \alpha_i^* = 1$ there can be no more than $\lfloor \nu n \rfloor$ many such points, which corresponds to a fraction of $\frac{\lfloor \nu n \rfloor}{n} \leq \nu$, thus showing the assertion (a). □

[2] For convex problems, this statement is known as the KKT condition *complementary slackness*. The argument holds, however, for the solution of the Lagrangian saddle point problem, regardless of whether or not the problem is convex, and for arbitrary objective and constraint functions.

Appendix B

Learning with Latent Class Dependencies

B.1 Analysis of LatentSVDD

When bounding the Rademacher complexity for Lipschitz continuous loss classes (such as the hinge loss or the squared loss), the following lemma is often very helpful.

Lemma 7 (Talagrand's lemma [260]). *Let $l : \mathbb{R} \to \mathbb{R}$ be a loss function that is L-Lipschitz continuous and $l(0) = 0$. Let $\mathcal{F}$ be a hypothesis class of real-valued functions and denote its loss class by $\mathcal{G} := l \circ \mathcal{F}$. Then the following inequality holds:*

$$R_n(\mathcal{G}) \leq 2LR_n(\mathcal{F}).$$

We can use the above result to prove Lemma 1.

Proof of Lemma 1. Since the LatentSVDD loss function is 1-Lipschitz with $l(0) = 0$, by Lemma 7, it is sufficient to bound $R(\mathcal{F}_{\text{SVDD}}(\mathbf{z}))$. To this end, it holds

$$\begin{aligned} R(\mathcal{F}_{\text{SVDD}}(\mathbf{z})) &\overset{\text{def.}}{=} \mathbb{E}\Big[\sup_{\mathbf{c},\Omega:0\leq\|\mathbf{c}\|^2+\Omega\leq\lambda} \frac{1}{n}\sum_{i=1}^{n}\sigma_i\big(\Omega + 2\langle\mathbf{c}, \Psi(\mathbf{x}_i,\mathbf{z})\rangle - \|\Psi(\mathbf{x}_i,\mathbf{z})\|^2\big)\Big] \\ &\leq \mathbb{E}\left[\sup_{\Omega:-\lambda\leq\Omega\leq\lambda} \frac{1}{n}\sum_{i=1}^{n}\sigma_i\Omega\right] + 2\mathbb{E}\left[\sup_{\mathbf{c}:\|\mathbf{c}\|^2\leq\lambda} \frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(\langle\mathbf{c}, \Psi(\mathbf{x}_i,\mathbf{z})\rangle\right)\right] \\ &\quad + \underbrace{\mathbb{E}\left[-\frac{1}{n}\sum_{i=1}^{n}\sigma_i\|\Psi(\mathbf{x}_i,\mathbf{z})\|^2\right]}_{=0 \text{ (by symmetry of } \sigma_i)}. \end{aligned} \tag{A.1.1}$$

Note that the term to the right is zero because the Rademacher variables are random signs, independent of $\mathbf{x}_1, \ldots, \mathbf{x}_n$. The term to the left can be bounded as follows:

$$\mathbb{E}\left[\sup_{\Omega:-\lambda\leq\Omega\leq\lambda} \frac{1}{n}\sum_{i=1}^{n}\sigma_i\Omega\right] = \lambda\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\right|\right] \overset{(*)}{\leq} \lambda\sqrt{\mathbb{E}\left[\frac{1}{n^2}\sum_{i,j=1}^{n}\sigma_i\sigma_j\right]} = \frac{\lambda}{\sqrt{n}}. \tag{A.1.2}$$

where for $(*)$ we employ Jensen's inequality. Moreover, applying the Cauchy-Schwarz inequality and Jensen's inequality, respectively, we obtain

$$\begin{aligned}\mathbb{E}\left[\sup_{\mathbf{c}:\|\mathbf{c}\|^2\leq\lambda}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(\langle\mathbf{c},\Psi(\mathbf{x}_i,\mathbf{z})\rangle\right)\right] &\overset{\text{C.-S.}}{\leq} \mathbb{E}\left[\sup_{\mathbf{c}:\|\mathbf{c}\|^2\leq\lambda}\|\mathbf{c}\|\left\|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\Psi(\mathbf{x}_i,\mathbf{z})\right\|\right] \\ &\overset{\text{Jensen}}{\leq} \sqrt{\lambda\mathbb{E}\left[\frac{1}{n^2}\sum_{i,j=1}^{n}\sigma_i\sigma_j\langle\Psi(\mathbf{x}_i,\mathbf{z}),\Psi(\mathbf{x}_j,\mathbf{z})\rangle\right]} \\ &= \sqrt{\lambda\frac{1}{n^2}\sum_{i=1}^{n}\|\Psi(\mathbf{x}_i,\mathbf{z})\|^2} \leq B\sqrt{\frac{\lambda}{n}} \qquad \text{(A.1.3)}\end{aligned}$$

because $\mathbb{P}(\|\Psi(\mathbf{x}_i,\mathbf{z})\|\leq B)=1$. Hence, inserting the results (A.1.2) and (A.1.3) into (A.1.1), yields the claimed result, that is,

$$R(\mathcal{G}_{\text{SVDD}}(\mathbf{z})) \overset{\text{Lemma 7}}{\leq} R(\mathcal{F}_{\text{SVDD}}(\mathbf{z})) \leq \frac{\lambda}{\sqrt{n}}+B\sqrt{\frac{\lambda}{n}} = \frac{\lambda+B\sqrt{\lambda}}{\sqrt{n}}. \qquad \text{(A.1.4)}$$

□

Next, we invoke the following result, taken from [201] (Lemma 8.1).

Lemma 8. *Let $\mathcal{F}_1,\ldots,\mathcal{F}_l$ be hypothesis sets in $\mathbb{R}^{\mathbf{x}}$, and let $\mathcal{F}:=\{\max(f_1,\ldots,f_l\}:f_i\in\mathcal{F}_i, i\in\{1,\ldots,l\}\}$. Then,*

$$R_n(\mathcal{F}) \leq \sum_{j=1}^{l}R_n(\mathcal{F}_j).$$

Sketch of proof [201]. The idea of the proof is to write $\max(h_1,h_2)=\frac{1}{2}(h_1+h_2+|h_1-h_2|)$, and then to show that

$$\mathbb{E}\left[\sup_{h_1\in\mathcal{F}_1,h_2\in\mathcal{F}_2}\frac{1}{n}\sum_{i=1}^{n}|h_1(x_i)-h_2(x_i)|\right] \leq R_n(\mathcal{F}_1)+R_n(\mathcal{F}_2).$$

This proof technique also generalizes to $l>2$. □

We can use Lemma 8 and Lemma 1, to conclude the main theorem of this paper, that is, Theorem 13, which establishes generalization guarantees of the usual order $O(1/\sqrt{n})$ for the proposed LatentSVDD method.

Proof of Theorem 13. First observe that, because l is 1-Lipschitz,

$$R_n(\mathcal{G}_{\text{LatentSVDD}}) \leq R_n(\mathcal{F}_{\text{LatentSVDD}}).$$

Next, note that we can write

$$R_n(\mathcal{F}_{\text{LatentSVDD}}) = \left\{\max_{\mathbf{z}\in\mathbf{z}}(f_{\mathbf{z}}) : f_{\mathbf{z}}\in\mathcal{F}_{\text{SVDD}}(\mathbf{z})\right\}.$$

Thus, by Lemma 2 and Lemma 4,

$$R_n(\mathcal{F}_{\text{LatentSVDD}}) \leq |\mathbf{z}|\max_{z\in\mathbf{z}}R_n(\mathcal{F}_{\text{SVDD}}(\mathbf{z})) \leq |\mathbf{z}|\frac{\lambda+B\sqrt{\lambda}}{\sqrt{n}}.$$

Moreover, observe that the loss function in the definition of $\mathcal{G}_{\text{LatentSVDD}}$ can only range in the interval $[0, B]$. Thus, Theorem 13 in the main paper gives the claimed result, that is,

$$\mathbb{E}[\widehat{g}_n] - \mathbb{E}[g^*] \leq 4R_n(\mathcal{G}_{\text{LatentSVDD}}) + B\sqrt{\frac{2\log(2/\delta)}{n}} \leq 4|\mathbf{z}|\frac{\lambda + B\sqrt{\lambda}}{\sqrt{n}} + B\sqrt{\frac{2\log(2/\delta)}{n}}.$$

□

B.2 Handcrafted procedure for porosity estimation

The common procedure for porosity estimation involves many intermediate domain knowledge decisions and it relies upon the interpolation method known as kriging [250]. The following are the main steps [247]:

1. First, the volume needs to be segmented into facies, which is usually accomplished by applying a combination of semi-automatic clustering methods and domain knowledge. The result is a facies model. All the following steps need then to be executed within each facies;

2. Determine the degree of correlation between the porosity, sampled in the drilled wells, and the seismic data, available at every node of the volume. Calibrate the seismic data to porosity from the well data samples;

3. Define a function describing the degree of spatial dependence of the seismic-derived porosity. This function is known as a *variogram*, and it is defined as the variance of the difference between a property value at two different locations in the reservoir. Three variograms must be created, one for each of the x, y, and z directions, due to the anisotropy usually present in the reservoir;

4. Create a variogram model consistent with the data as a result of the previous step and fit the model parameters;

5. Choose a kriging method (simple, ordinary, anisotropic, universal, etc.), passing the variogram model created in the previous step, and interpolate the data, generating the porosity volume.

The procedure described above demands a great amount of specialized human effort, usually taking days or even weeks to be accomplished.

Bibliography

[1] Chandola, V., Banerjee, A., Kumar, V., "Anomaly detection: A survey", *ACM Computing Surveys (CSUR)*, vol. 41, no. 3, 1–58, 2009.

[2] Aggarwal, C. C., *Outlier Analysis*. Springer, 2013.

[3] Harmeling, S., Dornhege, G., Tax, D., Meinecke, F., Müller, K.-R., "From outliers to prototypes: Ordering data", *Neurocomputing*, vol. 69, no. 13-15, pp. 1608–1618, 2006.

[4] Moya, M. M., "A constrained second-order network with mean square error minimization and boundary size minimization for one-class classification", Sandia National Labs., Albuquerque, NM (United States), Tech. Rep., 1993.

[5] Moya, M. M., Hush, D. R., "Network constraints and multi-objective optimization for one-class classification", eng, *Neural networks*, vol. 9, no. 3, pp. 463–474, 1996.

[6] Laskov, P., Schäfer, C., Kotenko, I., Müller, K.-R., "Intrusion detection in Unlabeled Data with Quarter-sphere Support Vector Machines", *Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA)*, vol. 27, pp. 71–82, 2004.

[7] Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., Williamson, R. C., "Estimating the Support of a High-dimensional Distribution", *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[8] Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J., Platt, J., "Support Vector Method for Novelty Detection", in *Advances in Neural and Information Processing Systems (NIPS)*, 2000, pp. 582–588.

[9] Tax, D., Duin, R., "Support Vector Data Description", *Machine Learning*, vol. 54, pp. 45–66, 2004.

[10] **Görnitz, N.**, Kloft, M., Rieck, K., Brefeld, U., "Active learning for network intrusion detection", in *ACM Workshop on Artificial Intelligence and Security (AISec)*, 2009, pp. 47–54.

[11] **Görnitz, N.**, Kloft, M., Brefeld, U., "Active and semi-supervised data domain description", in *European Conference on Machine Learning & Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, Springer, 2009, 407–422.

[12] **Görnitz, N.**, Kloft, M., Rieck, K., Brefeld, U., "Toward Supervised Anomaly Detection", *Journal of Artificial Intelligence Research (JAIR)*, vol. 46, pp. 235–262, 2013.

[13] Banerjee, A., Burlina, P., Diehl, C., "A support vector method for anomaly detection in hyperspectral imagery", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 8, pp. 2282–2291, 2006.

[14] Schölkopf, B., Giesen, J., Spalinger, S., "Kernel methods for implicit surface modeling", in *Advances in Neural and Information Processing Systems (NIPS)*, 2004, pp. 1193–1200.

[15] **Görnitz, N.**, Porbadnigk, A. K., Binder, A., Sanelli, C., Braun, M., Müller, K.-R., Kloft, M., "Learning and Evaluation in Presence of Non-i.i.d. Label Noise", in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 33, 2014, pp. 293–302.

[16] **Görnitz, N.**, Porbadnigk, A. K., Kloft, M., Binder, A., Sannelli, C., Braun, M., Müller, K.-R., "When brain and behavior disagree: A novel ML approach for handling systematic label noise in EEG data", in *Machine Learning and Interpretation in Neuroimaging Workshop (MLINI)*, 2013.

[17] Gao, H., Tang, J, Liu, H, "Mobile location prediction in spatio-temporal context", *Nokia mobile data challenge workshop*, no. 2, pp. 1–4, 2012.

[18] Kanamori, H., "Earthquake prediction: An overview", *International Handbook of Earthquake and Engineering Seismology*, vol. 81, pp. 1205–1216, 2003.

[19] Xu, W., Tran, T. T., Srivastava, R. M., Journel, A. G., "Integrating Seismic Data in Reservoir Modeling: The Collocated Cokriging Alternative", in *Annual Technical Conference and Exhibition of the Society of Petroleum Engineers*, 1992, pp. 833–842.

[20] Pawelzik, K., Kohlmorgen, J., Müller, K.-R., "Annealed Competition of Experts for a Segmentation and Classification of Switching Dynamics", *Neural Computation*, vol. 8, no. 2, pp. 340–356, 1996.

[21] Cetin, M, Comert, G, "Short-term traffic flow prediction with regime switching models", *Journal of the Transportation Research Board*, pp. 23–31, 2006.

[22] Widmer, C., Kloft, M., **Görnitz, N.**, Raetsch, G., "Efficient Training of Graph-Regularized Multitask SVMs", in *European Conference on Machine Learning & Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2012, pp. 633–647.

[23] Nasir, J. A., **Görnitz, N.**, Brefeld, U., "An Off-the-shelf Approach to Authorship Attribution", in *International Conference on Computational Linguistics (COLING)*, 2014, pp. 895–904.

[24] Porbadnigk, A., **Görnitz, N.**, Kloft, M., Müller, K.-R., "Decoding Brain States during Auditory Perception by Supervising Unsupervised Learning.", *Journal of Computing Science and Engineering (JCSE)*, vol. 7, no. 2, pp. 112–121, 2013.

[25] **Görnitz, N.**, Braun, M., Kloft, M., "Hidden Markov Anomaly Detection", in *International Conference on Machine Learning (ICML)*, 2015, pp. 1833–1842.

[26] **Görnitz, N.**, Widmer, C., Zeller, G., Kahles, A., Sonnenburg, S., Rätsch, G., "Hierarchical Multitask Structured Output Learning for Large-scale Sequence Segmentation", in *Advances in Neural Information Processing Systems (NIPS)*, 2011, pp. 2690–2698.

[27] Zeller, G., **Görnitz, N.**, Kahles, A., Behr, J., Mudrakarta, P., Sonnenburg, S., Rätsch, G., "mTim: rapid and accurate transcript reconstruction from RNA-Seq data", *ArXiv*, 2013.

[28] Porbadnigk, A. K., **Görnitz, N.**, Sannelli, C., Binder, A., Braun, M., Kloft, M., Müller, K.-R., "When Brain and Behavior Disagree: Tackling systematic label noise in EEG data with Machine Learning", in *IEEE International Winter Workshop on Brain-Computer Interface (BCI)*, 2014.

[29] Porbadnigk, A. K., **Görnitz, N.**, Sannelli, C., Binder, A., Braun, M., Kloft, M., Müller, K.-R., "Extracting latent brain states — Towards true labels in cognitive neuroscience experiments", *NeuroImage*, vol. 120, pp. 225–253, 2015.

[30] **Görnitz, N.**, Lima, L. A., Varella, L. E., Müller, K.-R., Nakajima, S., "Transductive Regression for Data with Latent Dependency Structure", *IEEE Transactions on Neural Networks and Learning (TNNLS)*, 2017.

[31] **Görnitz, N.**, Lima, L. A., Müller, K.-R., Kloft, M., Nakajima, S., "Support vector data descriptions and k-means clustering: one class?", *IEEE Transactions on Neural Networks and Learning (TNNLS)*, 2017.

[32] Lima, L. A., **Görnitz, N.**, Varella, L. E., Vellasco, M., Müller, K.-R., Nakajima, S., "Porosity Estimation by Semi-supervised Learning with Sparsely Available Labeled Samples", *Computers & Geosciences*, vol. 106, pp. 33–48, 2017.

[33] Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K., Schölkopf, B., "An Introduction to Kernel-based Learning Algorithms", English, *IEEE Transactions on Neural Networks (TNNLS)*, vol. 12, no. 2, pp. 181–201, Jan. 2001.

[34] Schölkopf, B., Mika, S., Burges, C. J., Knirsch, P., Müller, K.-R., Rätsch, G., Smola, A. J., "Input Space Versus Feature Space in Kernel-Based Methods", *IEEE Transactions on Neural Networks (TNN)*, vol. 10, no. 5, pp. 1000–1017, 1999.

[35] Schölkopf, B., Smola, A., Müller, K.-R., "Nonlinear Component Analysis as a Kernel Eigenvalue Problem", *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[36] ——, "Kernel principal component analysis", in *Artificial Neural Networks (ICANN)*, 1997, pp. 583–588.

[37] Cortes, C., Vapnik, V., "Support-Vector Networks", *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[38] Boser, B. E., Guyon, I. M., Vapnik, V. N., "A Training Algorithm for Optimal Margin Classifiers", in *Proceedings of the 5th Annual Workshop on Computational Learning Theory (COLT)*, 1992, pp. 144–152.

[39] Mohri, M., Rostamizadeh, A., Talwalkar, A., *Foundations of Machine Learning*. MIT Press, 2012.

[40] Smola, A. J., Schölkopf, B., Müller, K.-R., "The connection between regularization operators and support vector kernels", *Neural Networks*, vol. 11, no. 4, pp. 637–649, 1998.

[41] Kimeldorf, G., Wahba, G., "Some results on Tchebycheffian spline functions", *Journal of Mathematical Analysis and Applications*, vol. 33, no. 1, pp. 82–95, 1971.

[42] Schölkopf, B., Herbrich, R., Williamson, R., Smola, A. J., "A Generalized Representer Theorem", in *International Conference on Learning Theory (COLT)*, 2001, pp. 416–426.

[43] Boyd, S., Vandenberghe, L., *Convex Optimization*. Cambridge University Press, 2004.

[44] Schmidt, M., "Convergence Rates of Stochastic Optimization Algorithms", University of British Columbia, Tech. Rep., 2010.

[45] Shalev-Shwartz, S., "Online Learning and Online Convex Optimization", *Foundations and Trends in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2011.

[46] Tao, P. D., An, L. T. H., "A D.C. Optimization Algorithm for Solving the Trust-Region Subproblem", *SIAM Journal on Optimization*, vol. 8, no. 2, pp. 476–505, 1998.

[47] Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers", *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2010.

[48] Bach, F., Jenatton, R., Mairal, J., Obozinski, G., "Optimization with Sparsity-Inducing Penalties", *Foundations and Trends in Machine Learning*, vol. 4, no. 1, pp. 1–106, 2012.

[49] Parikh, N., Boyd, S., "Proximal Algorithms", *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[50] Yen, E., Peng, N., Wang, P.-W., Lin, S.-D., "On convergence rate of concave-convex procedure", in *NIPS Optimization Workshop*, 2012.

[51] An, L. T. H., Tao, P. D., "The DC (Difference of Convex Functions) Programming and DCA Revisited with DC Models of Real World Nonconvex Optimization Problems", *Annals of Operations Research*, vol. 133, no. 1-4, pp. 23–46, 2005.

[52] Horst, R, Thoai, N. V., "DC Programming : Overview", *Journal of Optimization Theory and Applications*, vol. 103, no. 1, pp. 1–43, 1999.

[53] Hawkins, D. M., *Identification of outliers*. Chapman and Hall, 1980, vol. 11.

[54] Grubbs, F. E., "Procedures for Detecting Outlying Observations in Samples", *Technometrics*, vol. 11, no. 1, pp. 1–21, 1969.

[55] Samek, W., Nakajima, S., Kawanabe, M., Müller, K.-R., "On robust parameter estimation in brain-computer interfacing", *Journal of Neural Engineering (JNE)*, vol. 14, no. 6, 2017.

[56] Höhner, J., Nakajima, S., Bauer, A., Müller, K.-R., **Görnitz, N.,** "Minimizing Trust Leaks for Robust Sybil Detection", in *International Conference on Machine Learning (ICML)*, Jul. 2017, pp. 1520–1528.

[57] Akoglu, L., Tong, H., Koutra, D., "Graph-based Anomaly Detection and Description: A Survey", *Data Mining and Knowledge Discovery*, p. 49, Apr. 2014.

[58] Stokes, J. W., Platt, J. C., Kravis, J., "ALADIN : Active Learning of Anomalies to Detect Intrusion", Tech. Rep., 2008.

[59] He, X., Niyogi, P., "Locality Preserving Projections", in *Advances in Neural Information Processing Systems (NIPS)*, 2003.

[60] Tibshirani, R., "Regression Selection and Shrinkage via the Lasso", *Journal of the Royal Statistical Society*, vol. 58, no. 1, pp. 267–288, 1996.

[61] Tax, D., Duin, R., "Data domain description using support vectors", in *Proceedings of the European Symposium on Artificial Neural Networks*, vol. 256, 1999, pp. 251–256.

[62] Breunig, M. M., Kriegel, H.-P., Ng, R. T., Sander, J., "LOF: Identifying Density-Based Local Outliers", *ACM SIGMOD International Conference on Management of Data*, pp. 1–12, 2000.

[63] Kriegel, H.-P., Kröger, P., Schubert, E., Zimek, A., "LoOP: Local Outlier Probabilities", in *ACM Conference on Information and Knowledge Management (CIKM)*, 2009, pp. 1649–1652.

[64] MacQueen, J., "Some methods for classification and analysis of multivariate observations", in *Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1967, pp. 281–297.

[65] Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U., "When Is "Nearest Neighbor" Meaningful?", in *International Conference on Database Theory (ICDT)*, 1999, pp. 217–235.

[66] Angiulli, F., "Concentration Free Outlier Detection", in *European Conference on Machine Learning (ECML)*, 2017, pp. 3–19.

[67] Zimek, A., Schubert, E., Kriegel, H. P., "A survey on unsupervised outlier detection in high-dimensional numerical data", *Statistical Analysis and Data Mining*, vol. 5, no. 5, pp. 363–387, 2012.

[68] Rissanen, J., "Modeling by shortest data description", *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[69] Grünwald, P., *The Minimum Description Length Principle*. MIT Press, 2007.

[70] ——, "A tutorial introduction to the minimum description length principle", *ArXiv*, 2004.

[71] Moya, M. M., Koch, M. W., Hostetler, L. D., "One-class classifier networks for target recognition applications", Sandia National Labs., Albuquerque, NM (United States), Tech. Rep., 1993.

[72] Moya, M. M., Hostetler, L. D., "One-class generalization in second-order backpropagation networks for image classification", Tech. Rep., 1989.

[73] Minter, T. C., "Single-Class Classification", *Symposium on Machine Processing of Remotely Sensed Data*, 1975.

[74] John H.J. Einmahl, David M. Mason, "Generalized Quantile Processes", *The Annals of Statistics*, vol. 20, no. 2, pp. 1062–1078, 1992.

[75] Polonik, W., "Minimum Volume Sets in Statistics: Recent Developments", in *Annual Conference of the Gesellschaft f{ü}r Klassifikation e.V.* R. Klar and O. Opitz, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 1997, pp. 187–194.

[76] ——, "Minimum volume sets and generalized quantile processes", *Stochastic Processes and their Applications*, vol. 69, no. 1, pp. 1–24, Jul. 1997.

[77] Tsybakov, A. B., "On nonparametric estimation of density level sets", *Annals of Statistics*, vol. 25, pp. 948–969, 1997.

[78] Polonik, W., "Measuring Mass Concentrations and Estimating Density Contour Clusters - An Excess Mass Approach", *The Annals of Statistics*, vol. 23, no. 3, pp. 855–881, 1995.

[79] Ghasemi, A., Rabiee, H. R., Manzuri, M. T., Rohban, M. H., "A Bayesian Approach to the Data Description Problem", in *AAAI Conference on Artificial Intelligence*, 2012, pp. 907–913.

[80] Xiao, Y., Wang, H., Xu, W., "Hyperparameter Selection for Gaussian Process One-Class Classification", *IEEE Transactions on Neural Networks (TNNLS)*, vol. 26, no. 9, pp. 2182–2187, 2015.

[81] Hovelynck, M., Chidlovskii, B., "Multi-modality in one-class classification", in *Proceedings of the 19th international conference on World wide web - WWW '10*, 2010, p. 441.

[82] Schölkopf, B, Platt, J. C., Shawe-Taylor, J, Smola, A. J., Williamson, R. C., "Estimating the Support of a High-dimensional Distribution", Microsoft Research, Tech. Rep., 1999.

[83] Rätsch, G., Mika, S., Schölkopf, B., Müller, K.-R., "Constructing Boosting Algorithms from SVMs: An Application to One-Class Classification", *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 24, pp. 1184–1199, 2002.

[84] Lee, G., Scott, C. D., "The one class support vector machine solution path", in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, 2007.

[85] Munoz, A., Moguerza, J. M., "One-class support vector machines and density estimation: The precise relation", in *Iberoamerican Congress on Pattern Recognition (CIARP)*, vol. 9, Springer, 2004, pp. 216–223.

[86] Davenport, M. A., Baraniuk, R. G., Scott, C. D., "Learning minimum volume sets with support vector machines", in *IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing (MLSP)*, 2007, pp. 301–306.

[87] Vert, R., Vert, J.-P., "Consistency and Convergence Rates of One-Class SVMs and Related Algorithms", *Journal for Machine Learning Research (JMLR)*, vol. 7, pp. 817–854, 2006.

[88] Glazer, A., Lindenbaum, M., Markovitch, S., "q -OCSVM : A q -Quantile Estimator for High-Dimensional Distributions", in *Advances in Neural and Information Processing Systems (NIPS)*, 2013, pp. 503–511.

[89] Lee, G., Scott, C. D., "Nested Support Vector Machines", *IEEE Transactions on Signal Processing (TSP)*, vol. 58, no. 3, 2010.

[90] Muandet, K., Schölkopf, B., "One-Class Support Measure Machines for Group Anomaly Detection", *ArXiv*, 2013.

[91] Erfani, S. M., Rajasegarar, S., Karunasekera, S., Leckie, C., "High-Dimensional and Large-Scale Anomaly Detection using a Linear One-Class SVM with Deep Learning", *Pattern Recognition*, vol. 58, pp. 121–134, 2016.

[92] David Tax, "One-class classification", PhD thesis, Delft University of Technology, 2001.

[93] Sjöstrand, K., Larsen, R., "The entire regularization path for the support vector domain description", *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 9, pp. 241–248, 2006.

[94] Fawcett, T., "An introduction to ROC analysis", *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[95] Hajizadeh, S., Li, Z., Dollevoet, R. P.B. J., Tax, D. M. J., "Evaluating Classification Performance with only Positive and Unlabeled Samples", in *Structural, Syntactic, and Statistical Pattern Recognition*, 2014, pp. 233–242.

[96] Goix, N., "How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms?", in *ICML 2016 Anomaly Detection Workshop*, 2016.

[97] Tax, D., Müller, K.-R., "A Consistency-Based Model Selection for One-class Classification", in *International Conference on Pattern Recognition (ICPR)*, 2004, pp. 363–366.

[98] Thomas, A., Clémençon, S., Feuillard, V., Gramfort, A., "Learning Hyperparameters for Unsupervised Anomaly Detection", in *ICML 2016 Anomaly Detection Workshop*, 2016.

[99] Legendre, A. M., *Nouvelles méthodes pour la détermination des orbites des comètes*. F. Didot, 1805.

[100] Gauss, C. F., *Theoria motus corporum coelestium sectionibus conicis solem ambientium*. Hamburgi: sumtibus Frid. Perthes et IH Besser, 1809, pp. 1–227.

[101] Stigler, S. M., "Gauss and the Invention of Least Squares", *The Annals of Statistics*, vol. 9, no. 3, pp. 465–474, 1981.

[102] Hoerl, A. E., Kennard, R. W., "Ridge Regression: Biased Estimation for Nonorthogonal Problems", *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[103] Lee, S., Zhu, J., Xing, E. P., "Adaptive Multi-Task Lasso: with Application to eQTL Detection", in *Advances in Neural and Information Processing Systems (NIPS)*, 2010.

[104] Huang, J., Zhang, T., Metaxas, D., "Learning with Structured Sparsity", *Journal for Machine Learning Research (JMLR)*, vol. 12, pp. 3371–3412, 2011.

[105] Kloft, M., Brefeld, U., Sonnenburg, S., Zien, A., "lp-Norm Multiple Kernel Learning", *Journal of Machine Learning Research (JMLR)*, vol. 12, 953–997, 2011.

[106] Kloft, M., "Lp-Norm Multiple Kernel Learning", PhD thesis, Berlin Institute of Technology (TU Berlin), 2011.

[107] Rakotomamonjy, A., Bach, F. R., Canu, S., Grandvalet, Y., "SimpleMKL", *Journal for Machine Learning Research (JMLR)*, vol. 9, pp. 2491–2521, 2008.

[108] Gönen, M., Alpaydın, E., "Multiple Kernel Learning Algorithms", *Journal of Machine Learning Research (JMLR)*, vol. 12, pp. 2211–2268, 2011.

[109] Rätsch, G., Mika, S., Schölkopf, B., Müller, K.-R., "Constructing boosting algorithms from SVMs: An application to one-class classification", *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 24, no. 9, pp. 1184–1199, 2002.

[110] Rätsch, G., Schölkopf, B., Mika, S., Müller, K.-R., "SVM and boosting: One class", GMD-Forschungszentrum Informationstechnik, Tech. Rep., 2000.

[111] Zou, H., Hastie, T., "Regularization and variable selection via the elastic-net", *Journal of the Royal Statistical Society*, vol. 67, no. 2, pp. 301–320, 2005.

[112] Mika, S., Rätsch, G., Müller, K.-R., "A Mathematical Programming Approach to the Kernel Fisher Algorithm", in *Advances in Neural Information Processing Systems (NIPS)*, vol. 13, 2001, pp. 591–597.

[113] Yuan, M., Lin, Y., "Model selection and estimation in regression with group variables", *Journal of the Royal Statistical Society*, vol. 68, no. 1, pp. 49–67, 2006.

[114] Gehler, P. V., Nowozin, S., "Infinite Kernel Learning", in *NIPS Workshop on Kernel Learning: Automatic Selection of Optimal Kernels*, 2008.

[115] Sören Sonnenburg, Gunnar Rätsch, Christin Schäfer, Bernhard Schölkopf, "Large Scale Multiple Kernel Learning", *Journal of Machine Learning Research (JMLR)*, vol. 7, pp. 1531–1565, 2006.

[116] Bach, F., "Consistency of the group Lasso and multiple kernel learning", *Journal of Machine Learning Research (JMLR)*, vol. 9, pp. 1179–1225, 2007.

[117] Kloft, M., Brefeld, U., Sonnenburg, S., Zien, A., "Non-Sparse Regularization for Multiple Kernel Learning", in *NIPS Workshop on Kernel Learning: Automatic Selection of Optimal Kernels*, 2008.

[118] Kloft, M., Brefeld, U., Sonnenburg, S., Laskov, P., Müller, K.-R., Zien, A., "Efficient and Accurate Lp-Norm Multiple Kernel Learning", in *Advances in Neural Information Processing Systems (NIPS)*, 2009, pp. 997–1005.

[119] Kloft, M., Brefeld, U., Düssel, P., Gehl, C., Laskov, P., "Automatic Feature Selection for Anomaly Detection", in *ACM Workshop on Artificial Intelligence and Security (AISec)*, 2008, 71–76.

[120] Lanckriet, G. R., Cristianini, N., Bartlett, P., Ghaoui, L. E., Jordan, M. I., "Learning the kernel matrix with semi-definite programming", *Journal of Machine Learning Research (JMLR)*, vol. 5, pp. 27–72, 2004.

[121] Bach, F. R., Lanckriet, G. R. G., Jordan, M. I., "Multiple kernel learning, conic duality, and the smo algorithm", in *International Conference on Machine Learning (ICML)*, 2004, pp. 6–14.

[122] Porbadnigk, A. K., Antons, J.-N., Blankertz, B., Treder, M. S., Schleicher, R., Möller, S., Curio, G., "Using ERPs for assessing the (sub)conscious perception of noise", in *IEEE Engineering in Medicine and Biology Society (EMBC)*, 2010, pp. 2690–2693.

[123] Porbadnigk, A. K., Scholler, S., Blankertz, B., Ritz, A., Born, M., Scholl, R., Müller, K.-R., Curio, G., Treder, M. S., "Revealing the neural response to imperceptible peripheral flicker with machine learning", in *IEEE Engineering in Medicine and Biology Society (EMBC)*, 2011, pp. 3692–3695.

[124] Scholler, S., Bosse, S., Treder, M. S., Blankertz, B., Curio, G., Müller, K.-R., Wiegand, T., "Towards a direct measure of video quality perception using EEG", *IEEE Transactions on Image Processing*, vol. 21, no. 5, pp. 2619–2629, 2012.

[125] Sutton, S., Braren, M., Zubin, J., John, E., "Evoked-potential correlates of stimulus uncertainty", *Science*, vol. 150, no. 3700, pp. 1187–1188, 1965.

[126] Müller, K.-R., Anderson, C. W., Birch, G. E., "Linear and non-linear methods for brain-computer interfaces", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, no. 2, pp. 165–169, 2003.

[127] Blankertz, B., Lemm, S., Treder, M. S., Haufe, S., Müller, K.-R., "Single-trial analysis and classification of ERP components – a tutorial", *NeuroImage*, vol. 56, pp. 814–825, 2011.

[128] Bünau, P., Meinecke, F. C., Király, F., Müller, K.-R., "Finding stationary subspaces in multivariate time series", *Physical Review Letters*, vol. 103, 2009.

[129] Shenoy, P., Krauledat, M., Blankertz, B., Rao, R. P. N., Müller, K.-R., "Towards adaptive classification for BCI", *Journal of Neural Engineering*, vol. 3, no. 1, R13, 2006.

[130] Vidaurre, C., Sannelli, C., Müller, K.-R., Blankertz, B., "Machine-learning based co-adaptive calibration", *Neural Computation*, vol. 23, no. 3, pp. 791–816, 2011.

[131] Kübler, A., Müller, K.-R., "An introduction to brain computer interfacing", in *Toward Brain-Computer Interfacing*, G. Dornhege, J. del R. Millán, T. Hinterberger, D. McFarland, and K.-R. Müller, Eds., Cambridge, MA: MIT press, 2007, pp. 1–25.

[132] Mendenhall, T. C., "The characteristic curves of composition", *Science*, vol. 9, no. 214, pp. 237–246, 1887.

[133] Maurer, H., Kappe, F., Zaka, B., "Plagiarism – a survey", *Journal of Universal Computer Science*, vol. 12, no. 8, pp. 1050–1084, 2006.

[134] Tan, E., Guo, L., Chen, S., Zhang, X., Zhao, Y. E., "Unik: Unsupervised social network spam detection", in *International Conference on Information & Knowledge Management (CIKM)*, 2013, pp. 479–488.

[135] Mosteller, F., Wallace, D. L., *Inference and disputed authorship: The federalist*. New York: Addison-Wesley, 1964.

[136] Fung, G., "The disputed federalist papers: Svm feature selection via concave minimization", in *Conference on Diversity in Computing*, 2003, pp. 42–46.

[137] Burrows, J. F., *Computation into criticism: A study of jane austen's novels and an experiment in method*. Oxford: Clarendon Press, 1987.

[138] Houvardas, J., Stamatatos, E., "N-gram feature selection for authorship identification", in *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, 2006, pp. 77–86.

[139] Stamatatos, E., Fakotakis, N., Kokkinakis, G., "Computer-based authorship attribution without lexical measures", *Computers and the Humanities*, vol. 35, no. 2, pp. 193–214, 2001.

[140] Raghavan, S., Kovashka, A., Mooney, R., "Authorship attribution using probabilistic context-free grammars", in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010, pp. 38–42.

[141] Koppel, M., Akiva, N., Dagan, I., "Feature instability as a criterion for selecting potential style markers: Special topic section on computational analysis of style", *Journal of the American Society for Information Science and Technology*, vol. 57, no. 11, pp. 1519–1525, 2006.

[142] Forman, G., "An extensive empirical study of feature selection metrics for text classification", *Journal of Machine Learning Research (JMLR)*, vol. 3, pp. 1289–1305, 2003.

[143] Seroussi, Y., Zukerman, I., Bohnert, F., "Authorship attribution with latent dirichlet allocation", in *International Conference on Computational Natural Language Learning*, 2011, pp. 181–189.

[144] Koppel, M., Schler, J., Argamon, S., "Authorship attribution in the wild", *Language Resources & Evaluation*, vol. 45, no. 1, pp. 83–94, 2011.

[145] Li, J., Zheng, R., Chen, H., "From fingerprint to writeprint", *Communications of the ACM*, vol. 49, no. 4, pp. 76–82, 2006.

[146] Zipf, G. K., *Selective studies and the principle of relative frequency in language*. Harvard University Press, 1932.

[147] Yule, U., *The statistical study of literary vocabulary*. Cambridge University Press, 2014.

[148] Holmes, D. I., "The Evolution of Stylometry in Humanities Scholarship", *Literary and Linguistic Computing*, vol. 13, no. 3, pp. 111–117, 1998.

[149] Rudman, J., "The state of authorship attribution studies: Some problems and solutions", *Computers and the Humanities*, vol. 31, no. 4, pp. 351–365, 1997.

[150] Diederich, J., Kindermann, J., Leopold, E., Paass, G., "Authorship attribution with support vector machines", *Applied Intelligence*, vol. 19, no. 1, pp. 109–123, 2003.

[151] Khmelev, D. V., "Disputed authorship resolution through using relative empirical entropy for markov chains of letters in human language texts", *Journal of Quantitative Linguistics*, vol. 7, no. 3, pp. 201–207, 2000.

[152] Stamatatos, E., Fakotakis, N., Kokkinakis, G. K., "Automatic text categorization in terms of genre and author", *Computational Linguistics*, vol. 26, no. 4, pp. 471–495, 2000.

[153] Holmes, D. I., Crofts, D. W., "The diary of a public man: A case study in traditional and non-traditional authorship attribution", *Literary and Linguistic Computing (LLC)*, vol. 25, no. 2, pp. 179–197, 2010.

[154] Stamatatos, E., "A survey of modern authorship attribution methods", *Journal of the American Society for Information Science and Technology*, vol. 60, no. 3, pp. 538–556, 2009.

[155] Koppel, M., Schler, J., "Exploiting stylistic idiosyncrasies for authorship attribution", in *International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 69, 2003, pp. 72–80.

[156] Dietterich, T, "Machine learning for sequential data: A review", *Structural, Syntactic, and Statistical Pattern Recognition*, vol. 2396, pp. 227–246, 2002.

[157] Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y., "Large Margin Methods for Structured and Interdependent Output Variables", *Journal of Machine Learning Research (JMLR)*, vol. 6, pp. 1453–1484, 2005.

[158] Lafferty, J., McCallum, A., Pereira, F., "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", in *International Conference on Machine Learning (ICML)*, 2001, pp. 282–289.

[159] Collins, M., "Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms", in *Conference on Empirical Methods in Natural Language Processing*, 2002, pp. 1–8.

[160] Rätsch, G., Sonnenburg, S., "Large Scale Hidden Semi-Markov SVMs", in *Advances in Neural and Information Processing Systems (NIPS)*, B Schölkopf, J Platt, and T Hoffman, Eds., Cambridge, MA: MIT Press, 2006, pp. 1161–1168.

[161] Rabiner, L. R., "A Tutorial on Hidden Markov Models and Selected Appications in Speech Recognition", *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[162] Bakir, G., Hofmann, T., Schölkopf, B., Smola, A. J., Taskar, B., S.V.N. Vishwanathan, *Predicting Structured Data*. MIT Press, 2007.

[163] Altun, Y., Tsochantaridis, I., Hofmann, T., "Hidden Markov Support Vector Machines", in *International Conference on Machine Learning (ICML)*, 2003, pp. 3–10.

[164] Hazan, T., Urtasun, R., "A Primal-Dual Message-Passing Algorithm for Approximated Large Scale Structured Prediction", in *Neural Information Processing Systems (NIPS)*, 2010, pp. 838–846.

[165] Smola, A. J., Vishwanathan, S. V. N., Le, Q. V., "Bundle methods for machine learning", in *Advances in Neural and Information Processing Systems (NIPS)*, vol. 20, 2008, 1377–1384.

[166] Teo, C. H., Vishwanathan, S. V. N., Smola, A., Le, Q. V., "Bundle Methods for Regularized Risk Minimization", *Journal of Machine Learning Research (JMLR)*, vol. 11, pp. 311–365, 2010.

[167] Do, T. M. T., "Regularized bundle methods for large-scale learning problems with an application to large margin training of hidden Markov models", PhD thesis, Sorbonne University, 2010.

[168] Franc, V., Sonnenburg, S., "Optimized cutting plane algorithm for support vector machines", in *International Conference on Machine Learning (ICML)*, New York, New York, USA: ACM Press, 2008, pp. 320–327.

[169] Nowozin, S., Lampert, C. H., "Structured Learning and Prediction in Computer Vision", *Foundations and Trends in Computer Graphics and Vision*, vol. 6, no. 3-4, pp. 185–365, 2010.

[170] Rifkin, R., Lippert, R., "Value Regularization and Fenchel Duality", *Journal of Machine Learning Research (JMLR)*, vol. 8, 441–479, 2007.

[171] McAllester, D., Keshet, J., "Generalization Bounds and Consistency for Latent Structural Probit and Ramp Loss", in *Advances in Neural and Information Processing Systems (NIPS)*, 2011, pp. 2205–2212.

[172] Lampert, C. H., Blaschko, M. B., "Structured prediction by joint kernel support estimation", *Machine Learning*, vol. 77, no. 2-3, pp. 249–269, Apr. 2009.

[173] Steinwart, I., Christmann, A., *Support vector machines*, 1st. Springer Science, 2008.

[174] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., Wold, B., "Mapping and quantifying mammalian transcriptomes by RNA-seq", *Nature Methods*, vol. 5, no. 7, p. 621, 2008.

[175] Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., Gilad, Y., "RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays", *Genome Research*, vol. 18, no. 9, pp. 1509–17, 2008.

[176] Wang, Z., Gerstein, M., Snyder, M., "RNA-seq: A revolutionary tool for transcriptomics", *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[177] Gan, X., Stegle, O., Behr, J., Steffen, J. G., Drewe, P., Hildebrand, K. L., Lyngsoe, R., Schultheiss, S. J., Osborne, E. J., Sreedharan, V. T., Kahles, A., Bohnert, R., Jean, G., Derwent, P., Kersey, P., Belfield, E. J., Harberd, N. P., Kemen, E., Toomajian, C., Kover, P. X., Clark, R. M., Rätsch, G., Mott, R., "Multiple reference genomes and transcriptomes for arabidopsis thaliana", *Nature*, vol. 477, pp. 419–423, 7365 2011.

[178] Anders, S., Huber, W., "Differential expression analysis for sequence count data", *Genome Biology*, vol. 11, no. 10, 2010.

[179] Bohnert, R., Rätsch, G., "rQuant.web: A tool for RNA-Seq-based transcript quantitation", *Nucleic Acids Research*, vol. 38, pp. 348–351, 2010.

[180] Behr, J., Kahles, A., Zhong, Y., Sreedharan, V. T., Drewe, P., Rätsch, G., "MITIE: Simultaneous RNA-Seq-based transcript identification and quantification in multiple samples", *Bioinformatics*, vol. 29, no. 20, pp. 2529–2538, 2013.

[181] Anders, S., Reyes, A., Huber, W., "Detecting differential usage of exons from rna-seq data", *Genome Research*, pp. 2008–2017, 2012.

[182] Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Palma, F., Birren, B. W., Nusbaum, C., Friedman, K. L.-T. N., Regev, A., "Full-length transcriptome assembly from RNA-Seq data without a reference genome", *Nature Biotechnology*, vol. 29, pp. 644–652, 7 2011.

[183] Schulz, M. H., Zerbino, D. R., Vingron, M., Birney, E., "Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels", *Bioinformatics*, vol. 28, no. 8, pp. 1086–1092, 2012.

[184] Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Baren, M. J., Salzberg, S. L., Wold, B. J., Pachter, L., "Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation", *Nature Biotechnology*, vol. 28, pp. 511–515, 5 2010.

[185] Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M. J., Gnirke, A., Nusbaum, C., Rinn, J. L., Lander, E. S., Regev, A., "Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs", *Nature Biotechnology*, vol. 28, pp. 503–510, 5 2010.

[186] Trapnell, C., Pachter, L., Salzberg, S. L., "TopHat: Discovering Splice Junctions with RNA-Seq", *Bioinformatics*, vol. 25, no. 9, pp. 1105–1111, 2009.

[187] Jean, G., Kahles, A., Sreedharan, V. T., Bona, F. D., Rätsch, G., "RNA-seq read alignments with PALMapper", *Current Protocols in Bioinformatics*, vol. 32, pp. 6–11, 2010.

[188] Schweikert, G., Zien, A., Zeller, G., Behr, J., Dieterich, C., Ong, C. S., Philips, P., Bona, F. D., Hartmann, L., Bohlen, A., Krüger, N., Sonnenburg, S., Rätsch, G., "mGene: Accurate SVM-based gene finding with an application to nematode genomes", *Genome Research*, vol. 19, no. 11, pp. 2133–2143, 2009.

[189] Sonnenburg, S., Schweikert, G., Philips, P., Behr, J., Rätsch, G., "Accurate splice site prediction using support vector machines", *BMC Bioinformatics*, vol. 8, no. 10, 2007.

[190] Jebara, T., Kondor, R., Howard, A., "Probability Product Kernels", *Journal of Machine Learning Research (JMLR)*, vol. 5, pp. 819–844, 2004.

[191] Jaakkola, T, Diekhans, M, Haussler, D, "Using the Fisher kernel method to detect remote protein homologies", in *International Conference on Intelligent Systems for Molecular Biology (ISMB)*, 1999, pp. 149–158.

[192] Tsuda, K., Kawanabe, M., Rätsch, G., Sonnenburg, S., Müller, K.-R., "A New Discriminative Kernel from Probabilistic Models", *Neural Computation*, vol. 2414, pp. 2397–2414, 2002.

[193] Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., Watson, J., *Molecular Biology of the Cell*, 4th. Garland, 2002.

[194] Laptev, N., Amizadeh, S., Flint, I., "Generic and Scalable Framework for Automated Time-series Anomaly Detection", in *International Conference on Knowledge Discovery and Data Mining (KDD)*, 2015, pp. 1939–1947.

[195] Dereszynski, E. W., Dietterich, T. G., "Spatiotemporal Models for Data-Anomaly Detection in Dynamic Environmental Monitoring Campaigns", *ACM Transactions on Sensor Networks*, vol. 8, no. 1, 3:1–3:36, 2011.

[196] Schölkopf, B., Smola, A. J., *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

[197] Yuille, A. L., Rangarajan, A., "The concave-convex procedure", *Neural computation*, vol. 15, no. 4, pp. 915–36, Apr. 2003.

[198] Sriperumbudur, B. K., Lanckriet, G. R. G., "On the Convergence of the Concave-Convex Procedure", in *NIPS*, 2009, pp. 1–9.

[199] Kloft, M., Brefeld, U., Sonnenburg, S., Zien, A., Laskov, P., Müller, K.-R., "Learning Non-sparse Kernel Mixtures", in *PASCAL2 Workshop on Sparsity in Machine Learning and Statistics*, 2009.

[200] Bartlett, P., Mendelson, S., "Rademacher and gaussian complexities: Risk bounds and structural results", *Journal of Machine Learning Research (JMLR)*, vol. 3, pp. 463–482, 2002.

[201] Mohri, M., Rostamizadeh, A., Talwalkar, A., *Foundations of machine learning*. The MIT Press, 2012.

[202] Jain, A. K., "Data clustering: 50 years beyond K-means", *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, Jun. 2010.

[203] Dhillon, I. S., Guan, Y., Kulis, B., "Kernel k-means, Spectral Clustering and Normalized Cuts", in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, New York, USA: ACM Press, Aug. 2004, p. 551.

[204] Girolami, M, "Mercer kernel-based clustering in feature space", English, *IEEE Transactions on Neural Networks and Learning (TNNLS)*, vol. 13, no. 3, pp. 780–4, Jan. 2002.

[205] Forero, P. A., Kekatos, V., Giannakis, G. B., "Robust Clustering Using Outlier-Sparsity Regularization", *IEEE Transactions on Signal Processing*, vol. 60, no. 8, pp. 4163–4177, Apr. 2012.

[206] Kondo, Y., "Robustification of the sparse K-means clustering algorithm", PhD thesis, The University of British Columbia, 2009.

[207] Kondo, Y., Salibian-Barrera, M., Zamar, R., "A robust and sparse K-means clustering algorithm", *ArXiv*, 2012.

[208] Hamerly, G., Elkan, C., "Learning the k in k-means", in *Advances in Neural and Information Processing Systems (NIPS)*, 2004, pp. 281–288.

[209] Jinwen Ma, Taijun Wang, "A cost-function approach to rival penalized competitive learning (RPCL)", *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 4, pp. 722–737, Aug. 2006.

[210] Bacciu, D., Starita, A., "Competitive Repetition Suppression (CoRe) Clustering: A Biologically Inspired Learning Model With Application to Robust Clustering", *IEEE Transactions on Neural Networks*, vol. 19, no. 11, pp. 1922–1941, Nov. 2008.

[211] Yiu-ming Cheung, "On rival penalization controlled competitive learning for clustering with automatic cluster number selection", *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 11, pp. 1583–1588, Nov. 2005.

[212] Jia, H., Cheung, Y.-M., Liu, J., "Cooperative and penalized competitive learning with application to kernel-based clustering", *Pattern Recognition*, vol. 47, pp. 3060–3069, 2014.

[213] Chang, W.-C., Lee, C.-P., Lin, C.-J., "A Revisit to Support Vector Data Description (SVDD)", National Taiwan University, Tech. Rep., 2010.

[214] Chang, C.-C., Tsai, H.-C., "A Minimum Enclosing Balls Labeling Method for Support Vector Clustering", National Taiwan University of Science and Technology (Taipei, Taiwan), Tech. Rep., 2007, pp. 1–27.

[215] Wang, X., Chung, F.-l., Wang, S., "Theoretical analysis for solution of support vector data description", *Neural networks*, vol. 24, pp. 360–369, 2011.

[216] Ng, A. Y., Jordan, M. I., Weiss, Y., "On Spectral Clustering: Analysis and an algorithm", in *Advances in Neural and Information Processing Systems (NIPS)*, 2002, pp. 849–856.

[217] Baltrusaitis, T, Banda, N, Robinson, P, "Dimensional affect recognition using Continuous Conditional Random Fields", *IEEE Automatic Face and Gesture Recognition (FG)*, pp. 1–8, 2013.

[218] Blaschko, M., Lampert, C., "Learning to localize objects with structured output regression", *European Conference on Computer Vision (ECCV)*, 2–15, 2008.

[219] Bo, L., Sminchisescu, C., "Structured output-associative regression", *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*, pp. 2403–2410, 2009.

[220] Peng, J., Bo, L., Xu, J., "Conditional Neural Fields", in *Neural Information Processing Systems (NIPS)*, vol. 9, 2009, pp. 1419–1427.

[221] Ratliff, N. D., Bagnell, J. A., Zinkevich, M. A., "(Online) Subgradient Methods for Structured Prediction", Carnegie Mellon University, Tech. Rep., 2007.

[222] Kim, M., "Semi-supervised learning of hidden conditional random fields for time-series classification", *Neurocomputing*, vol. 119, pp. 339–349, 2013.

[223] Cortes, C., Mohri, M., "On Transductive Regression", in *Advances in Neural and Information Processing Systems (NIPS)*, 2007, pp. 305–312.

[224] Chapelle, O., Vapnik, V., Weston, J., "Transductive Inference for Estimating Values of Functions", in *Advances in Neural and Information Processing Systems (NIPS)*, 2000, pp. 421–427.

[225] Leisch, F., "FlexMix : A General Framework for Finite Mixture Models and Latent Class Regression in R", *Journal of Statistical Software*, vol. 11, no. 8, pp. 1–18, 2004.

[226] Grün, B., Leisch, F., "FlexMix Version 2 : Finite Mixtures with Concomitant Variables and Varying and Constant Parameters", *Journal of Statistical Software*, vol. 28, no. 4, pp. 1–35, 2008.

[227] Sindhwani, V., Niyogi, P., Belkin, M., "Beyond the point cloud: from transductive to semi-supervised learning", in *Proceedings of the International Conference on Machine Learning (ICML)*, vol. 1, ACM, 2005, 824–831.

[228] Belkin, M., Niyogi, P., Sindhwani, V., "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples", *Journal of Machine Learning Research (JMLR)*, vol. 7, no. 2006, pp. 2399–2434, 2006.

[229] Belkin, M., Matveeva, I., Niyogi, P., "Regression and regularization on large graphs", Tech. Rep., 2003.

[230] Yu, C.-N. J., Joachims, T., "Learning Structural SVMs with Latent Variables", in *International Conference on Machine Learning (ICML)*, 2009, pp. 1169–1176.

[231] Wainwright, M., Jordan, M. I., "Variational inference in graphical models: The view from the marginal polytope", in *Allerton Conference on Control, Communication and Computing*, 2003.

[232] Wainwright, M. J., Jordan, M. I., "Graphical Models, Exponential Families, and Variational Inference", *Foundations and Trends in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, 2008.

[233] Weiss, Y., Freeman, W. T., "Correctness of belief propagation in Gaussian models of arbitrary topology", *Neural Computation*, vol. 298, no. 0704, 2000.

[234] Besag, J., "Spatial Interaction and the Statistical Analysis of Lattice Systems", *Journal of Royal Statistical Society*, vol. 36, no. 2, pp. 192–236, 1974.

[235] Cristianini, N., Shawe-Taylor, J., Elisseeff, A., Kandola, J., "On Kernel-Target Alignment", in *Advances in Neural and Information Processing Systems (NIPS)*, 2002, pp. 367–373.

[236] Braun, M. L., Buhmann, J. M., Müller, K.-R., "On Relevant Dimensions in Kernel Feature Spaces", *Journal of Machine Learning Research (JMLR)*, vol. 9, pp. 1875–1908, 2008.

[237] Cortes, C., Mohri, M., Rostamizadeh, A., "Algorithms for learning kernels based on centered alignment", *Journal of Machine Learning Research (JMLR)*, vol. 13, no. 1, pp. 795–828, 2012.

[238] Cortes, C., Mohri, M., "Confidence intervals for the area under the roc curve", in *Advances in Neural Information Processing Systems (NIPS)*, 2004.

[239] Porbadnigk, A. K., Treder, M. S., Blankertz, B., Antons, J.-N., Schleicher, R., Möller, S., Curio, G., Müller, K.-R., "Single-trial analysis of the neural correlates of speech quality perception", *Journal of Neural Engineering*, vol. 10, 5 2013.

[240] Falkenstein, M., Hoormann, J., Christ, S., Hohnsbein, J., "ERP components on reaction errors and their functional significance: A tutorial", *Biology and Psychology*, vol. 51, no. 2-3, pp. 87–107, 2000.

[241] Hohnsbein, J, Falkenstein, M, Hoormann, J, "Error processing in visual and auditory choice reaction tasks", *Journal of Psychophysiology*, vol. 3, 1998.

[242] Falkenstein, M., Hohnsbein, J., Hoormann, J., Blanke, L., "Effects of errors in choice reaction tasks on the ERP under focused and divided attention", in *Psychophysiological Brain Research*, C. H. M. Brunia, A. W. K. Gaillard, and A. Kok, Eds., Tilburg University Press, Tilburg, 1990, pp. 192–195.

[243] Gehring, W. J., Coles, M. G. H., Meyer, D. E., Donchin, E., "The error-related negativity: An event-related brain potential accompanying errors", *Journal of Psychophysiology*, vol. 27, pp. 385–390, 1990.

[244] Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., Donchin, E., "A neural system for error detection and compensation", *Psychological Science*, vol. 4, pp. 385–390, 1993.

[245] Brickenkamp, R., Zillmer, E., *D2 test of attention*. Göttingen, Germany: Hogrefe & Huber, 1998.

[246] Castro, S., Caers, J., Mukerji, T., "The Stanford VI Reservoir", *Annual Report of the Stanford Center for Reservoir Forecasting (SCRF)*, pp. 153–154, 2005.

[247] Deutsch, C. V., *Geostatistical Reservoir Modeling*. Oxford University Press, 2002.

[248] Dvorkin, J., Gutierrez, M. A., Grana, D., *Seismic Reflections of Rock Properties*. Cambridge University Press, 2014.

[249] Ratsaby, J., Venkatesht, S. S., "Learning from a Mixture of Labeled and Unlabeled Examples with Parametric Side Information", *Annual Conference on Computational Learning Theory*, pp. 412–417, 1995.

[250] Matheron, G., *Traité de geoestatistique apliquée*. Paris: Editions Technip, 1962, vol. 1.

[251] Deutsch, C. V., Journel, A. G., *GSLIB - Geostatistical Software Library and User's Guide*, 2nd ed. Oxford University Press, 1998.

[252] Siddiqui, M. A., Fern, A., Dietterich, T. G., Wong, W.-K., "Sequential Feature Explanations for Anomaly Detection", *ArXiv*, Feb. 2015.

[253] Vidovic, M. M.-C., **Görnitz, N.**, Müller, K.-R., Kloft, M., "Feature Importance Measure for Non-linear Learning Algorithms", in *NIPS Workshop on Interpretable Machine Learning in Complex Systems*, Nov. 2016.

[254] Vidovic, M. M.-C., **Görnitz, N.**, Müller, K.-R., Rätsch, G., Kloft, M., "Opening the Black Box: Revealing Interpretable Sequence Motifs in Kernel-Based Learning Algorithms", in *European Conference on Machine Learning & Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2015, pp. 137–153.

[255] Vidovic, M. M.-C., Kloft, M., Müller, K.-R., **Görnitz, N.,** "ML2Motif—Reliable extraction of discriminative sequence motifs from learning machines", *PloS one*, vol. 12, no. 3, 2017.

[256] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., Samek, W., "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation", *PloS one*, vol. 10, no. 7, 2015.

[257] Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.-R., "Explaining nonlinear classification decisions with deep taylor decomposition", *Pattern Recognition*, vol. 65, pp. 211–222, 2017.

[258] Ribeiro, M. T., Singh, S., Guestrin, C., "Why Should I Trust You? Explaining the Predictions of Any Classifier", in *Knowledge Discovery and Data Mining (KDD)*, 2016.

[259] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., "Generative Adversarial Nets", *ArXiv*, 2016.

[260] Talagrand, M., "Concentration of measure and isoperimetric inequalities in product spaces", *Publications Mathématiques de l'Institut des Hautes Études Scientifiques*, vol. 81, pp. 73–205, 1 1995.