Kernel-Based Regression and Objective Nonlinear Measures to Assess Brain Functioning

Roman Rosipal

A thesis submitted in partial fulfillment of the requirements
of the University of Paisley for the degree of doctor of philosophy.

September 2001

Applied Computational Intelligence Research Unit
School of Information and Communications Technology
University of Paisley, Scotland

— *To Eva and Adam* —

Abstract

Two different problems of reflecting brain functioning are addressed. This involves human performance monitoring during the signal detection task and depth of anaesthesia monitoring. The common aspect of both problems is to monitor brain activity through the electroencephalogram recordings on the scalp. Although these two problems create only a fractional part of the tasks associated with physiological data analysis the results and the methodology proposed have wider applicability.

- A theoretical and practical investigation of the different forms of kernel-based nonlinear regression models and efficient kernel-based algorithms for appropriate features extraction is undertaken. The main focus is on solving the problem of providing reduced variance estimates of the regression coefficients when a linear regression in some kernel function defined feature space is assumed. To that end Kernel Principal Component Regression and Kernel Partial Least Squares Regression techniques are proposed. These kernel-based techniques were found to be very efficient when observed data are mapped to a high dimensional feature space where usually algorithms as simple as their linear counterparts in input space are used. The methods are used and compared with existing kernel-based regression techniques in measuring the human signal detection performance from the associated Event Related Potentials.

- The depth of anaesthesia (DOA) problem was addressed by assuming different complexity measures. These measures were inspired by nonlinear dynamical systems and information theories. Data from patients undergoing general anesthesia were used and the results were compared with traditional spectral indices. The promising results of this pilot study suggest the possibility to include these measures into the existing family of DOA descriptors. This opens a new area of more detailed and extensive research into this very important medical problem.

Declaration

The work contained in this thesis is the result of my own investigations and has not been accepted nor concurrently submitted in candidature for any other academic award.

Acknowledgements

First, I would like to thank Dr Alan Hope for all his help and guidance throughout clinical aspects of anaesthesia problem. His eagerness to help have been essential for the part of research associated with experimental settings, measurements and evaluation of the achieved results on depth of anaesthesia problem. Further, I would like to thank Dr Leonard J. Trejo for providing me the ERP data acquired in a signal detection task and for his continuous discussion on many different aspects and the results on the problem associated with this data set.

I would like to thank my supervisor Prof Mark Girolami for providing me opportunity to carry out this PhD studies, for all his support during the studies and valuable comments, criticism and suggestions which were invaluable.

Thanks are also extended to many other researchers and PhD students for the many discussions, suggestions, comments and helpful feedback. I would like to thank Dr Lucius Chudý, Dr Andrzej Cichocki, Prof Colin Fyfe, Dr Bogdan Gabrys, Ms Ata Kabán and Mr Alexei Vinokurov.

On a more personal level, I would like to thanks my wife, son and my parents for their continued love, patience and support over the PhD study.

I was funded be a research grant for the project Objective Measures of Depth of Anaesthesia, University of Paisley and Glasgow Western Infirmary NHS trust, and I am partially supported by Slovak Grant Agency for Science (grants No. 2/5088/00 and No. 00/5305/468). I am grateful to them for their financial support.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

List of Abbreviations and Symbols

Frequently used abbreviations

ApEn	Approximate Entropy
BIS	Bispectral Index
CCEn	Corrected Conditional Entropy
CD	Correlation Dimension
CEn	Conditional Entropy
CER	Coarse-grain Entropy Rates
CV	Cross Validation
DOA	Depth of Anaesthesia
EEG	Electroencephalogram
EM	Expectation Maximization
EMKPCA	Expectation Maximization Approach to Kernel Principal Component Analysis
ERM	Empirical Risk Minimization
ERP	Event Related Potential
FFT	Fast Fourier Transform
GPER	Gaussian Process Entropy Rates
iid	independent identically distributed
KSE	Kolmogorov-Sinai Entropy
LV	Latent Variables
ML	Maximum Likelihood
MLSVR	Multi-layer Support Vector Regression
NCI	Nonlinear Correlation Index
NIPALS	Nonlinear Iterative Partial Least Squares
NRMSE (NMSE)	Normalized (Root) Mean Squared Error
OLS	Ordinary Least Squares
PCA	Principal Component Analysis
PCR	Principal Component Regression
PLS	Partial Least Squares
PRESS	Prediction Error Sum of Squares
RR	Ridge Regression
RKHS	Reproducing Kernel Hilbert Space
RN	Regularization Networks
SE95	Spectral Edge 95%
SpEn	Spectral Entropy
SRC	Spearman's Ranked Correlation Coefficient
SRM	Structural Risk Minimization
SVM	Support Vector Machines
SVR	Support Vector Regression
TCI	Target Controlled Infusions
TPC	Test Proportion Correct
VC	Vapnik-Chervonenkis dimension

Important Symbols

$\mathcal{R}$	the set of reals
$\mathcal{R}^+$	the set of positive reals
$\mathcal{N}$	the set of natural numbers
$\mathcal{R}^N$	the N-dimensional vector space
$\mathcal{X}, \mathcal{Y}$	space of data points
$\mathcal{F}$	feature space
$\mathcal{H}$	Reproducing Kernel Hilbert Space
$L_2(\mathcal{X})$	a Hilbert space of real valued functions defined over $\mathcal{X}$
$\langle f, g \rangle$	an inner product in $L_2(\mathcal{X})$ space; $\langle f, g \rangle = \int_{\mathcal{X}} f(\mathbf{x})g(\mathbf{x})d\mathbf{x}$
$(\mathbf{x}.\mathbf{y}) = \mathbf{x}^T\mathbf{y}$	(canonical) dot product between vectors $\mathbf{x}$ and $\mathbf{y}$
$\lvert x \rvert$	absolute value (magnitude) of x
$\lVert . \rVert$	norm (normally Euclidean norm (length))
$\lVert . \rVert_{\mathcal{H}}$	Reproducing Kernel Hilbert Space norm
$[., .]$	closed interval
$(., .)$	open interval
$\Pr[X = x]$	probability of event x
$E[X]$	statistical expectation with respect to probability distribution function $P(x)$; $E[X] = \int_{\mathcal{X}} xdP(x)$
$N(a, \sigma)$	normal distribution with mean a and variance σ
δ_{ij}	Kronecker symbol
$\ln$	logarithm to base e (natural logarithm)
$\log_a$	logarithm to base a
$\mathbf{x}^T, \mathbf{X}^T$	vector, matrix transpose
$\mathbf{X}^{-1}$	matrix inverse
$\mathbf{1}_n$	vector of ones of length n
$\mathbf{I}$	identity matrix
$\mathbf{I}_n$	diagonal matrix with the elements on diagonal equal to n
$diag(.)$	diagonal matrix
$rank(.)$	rank of matrix
$trace(.)$	trace of matrix
$var(.)$	variance
$cov(.)$	covariance
$\in$	symbol for “belongs to”
$\subset$	symbol for “subset of”
$\subseteq$	symbol for “subset of or equals to”
$\equiv$	symbol for “denote, define”

1. INTRODUCTION

The work which follows springs from two sources which may seem to be rather different, however, under more detailed inspection we discover that the results of the individual parts are contributory to each other. Two different problems of reflecting brain functioning are attacked. These are human performance monitoring during the signal detection task and depth of anaesthesia monitoring. The common aspect of both problems is to monitor brain activity through the electroencephalogram (EEG) recordings on the scalp. However, the major difference between both problems is how these recordings are related to the behavior of the humans under investigation. While in the case of human performance monitoring we may objectively evaluate the ability of individual subjects to detect a desired signal occurring on the monitor by observing the correctness and reaction time of subjects, in the case of depth of anaesthesia monitoring we need to focus on the possible extraction of such features (descriptors) which may help us to make a decision about the anaesthetic level of patients during the surgery.

We provide a more detailed introduction to the investigated problems in the individual parts of this thesis, however, for the moment we would like to briefly discuss general aspects of the use of EEG in clinical or experimental practice and provide several different examples and experiments of brain function monitoring which are related to our problems. The EEG was first described in 1875 by Richard Caton, a physician in Liverpool, who made experiments on exposed cortical surface of animals where he observed electrical oscillations. However, it was Hans Berger, a psychiatrist in Jena, who first reported more systematic descriptions of human EEG. From that time there began a large interest in the medical community to use EEG in clinical practice. Complicated, and on first inspection irregular and random behavior of EEG traces gave rise to the necessity to involve a more rigorous approach to EEG than simple visual inspection of EEG recordings. A well established methodology to investigate the changes in EEG recordings is based on transformation of the signal into the frequency domain where inspection of the waveforms belonging to different frequency bands is usually conducted. The former observations of EEG recordings of a wide range of different humans of varying age, physique, psychological, healthy or unhealthy conditions provides a reference for the detection of possible anomalies in investigated patients. Spikes, short term lower amplitude oscillations or other features occurring in EEG traces may further serve for better detection of brain activity changes. Another, recently applied promising methodology of EEG signals processing is based on theory of nonlinear dynamical systems, chaos theory and theory of stochastic processes.

Another domain of brain functioning monitoring is the recording of brain event related potentials (ERP), that is electroencephalographic recordings time-locked to a specific stimulus or cognitive activity. ERP reflect mental processes and are known to be related to human performance, including signal detection, target identification and recognition, memory, tracking and mental computation.

We already sketched at the beginning of this chapter that we may generally consider two different tasks of brain function monitoring. The first category is given by problems where the objective measurement of human behavior may be observed. Simple eyes opening and closing during the EEG recording will create two categories and the classification task to recognize these events from EEG traces is well defined in the way that we may construct a good classifier by observing this objective information. Examples of these tasks are found in the domain of neurocontrol techniques using the detection of stimuli or cognitive activities determined EEG signals (e.g. ERP) as an indicator of human intentions. This provides another channel of possible interactions between human and machines (brain computer interface, robotic control) or opens new possibilities of communication between humans (rehabilitative medicine). By giving a human a mental task (e.g. counting) for the case when the letter of interest is shortly flashed on the monitor we may observe a significant ERP component occurring approximately 300ms after the stimulus (P300 component). This allows us to discriminate between the letter of interest and remaining letters of an alphabet also randomly flashed on the screen. This may produce a new tool for communication between paralyzed, motor limited patients and physicians. The problem of human performance monitoring during the signal detection task investigated in this thesis belongs to this category. The regression model reflecting the dependence between measured ERP and performance measure consisting of correctness, reaction time and confidence as provided by investigated subject is constructed.

However, in many clinical practice problems we usually do not have this reliable, objective information we want to reference to the measured EEG signals. An example of this may be the automatic detection of different stages of sleep. There exist generally five different stages of sleep and one standard criterion of their classification is defined by the Rechtschaffen and Kales scoring system. However, the actual classification by visual inspection of the EEG traces depends on experience and up to some level on subjective decision of the electroencephalographer. Although this subjective decision may by partially removed by appropriate processing of the raw EEG data with the aim to more reliably detect individual features included in the general scoring systems, we cannot assess a fully objective decision as this depends on electroencephalographer or laboratory conditions. Thus the construction of a fully automatic classifier will depend on the scoring provided and in reverse the spurious results of the system have to be consulted with an electroencephalographer. Even more complicated may be the problem of the depth of anaesthesia monitoring investigated in the second part of this thesis. In this case there does not exist a generally acceptable standard scoring system and the decision of how deeply the patients are anaesthetized during the surgery depends on an anaesthesiologist who needs to combine a wide range of clinical aspects and different auxiliary measures. Finding a reliable, objective measure derived from EEG recordings attracted the attention of the research community over the past few decades and this thesis also contributes to this area.

1.1 Contribution of the thesis

In this subsection we briefly summarize the novelty of the study to make the later presentation clearer and the contribution of the thesis more evident. All contributions to the field are summarized in refereed journal papers and conference publications as well as in technical reports which reflect both the theoretical and experimental results of the author.

The two problems described in the previous section and addressed in this thesis will be studied in parallel.

- In the case of human performance monitoring during the signal detection task the new kernel based regression techniques were studied. The introductory study of the support vector regression (SVR) technique resulted in a conference publication [106], where a new method of on-line prediction of the chaotic Mackey-Glass time series was proposed. Although this part of the study is not directly connected to the work presented in the thesis, the superior performance of SVR in comparison to a Resource Allocating Radial Basis Function network provides an indication of the potential applicability of the methodology of constructing a linear regression model in a feature space where the original data are nonlinearly mapped. In the next step Kernel Principal Component Analysis (PCA) will be studied as a potential tool for nonlinear features extraction. This study also motivates our investigation into Kernel Principal Component Regression models, including an extensive comparison of the existing Kernel Ridge Regression, SVR and Multi-layer SVR techniques. Nonlinear, kernel-based regression models using features extracted by the linear PCA and Kernel PCA were compared in the conference publication [108] and journal paper [110]. Statistically significant improvement using Kernel PCA in comparison to linear PCA preprocessing indicate the usefulness of the approach. The computational and memory allocation constraints when Kernel PCA is applied to large data sets motivated our derivation of the Expectation Maximization approach to Kernel PCA published in [107]. In the thesis we will discuss other theoretical aspects of the algorithm. In [111] we have experimentally demonstrated the ability of the algorithm to extract principal components which lead to the Kernel PCR models with the same accuracy when the Kernel PCA algorithm is used. However, we demonstrated that in the case when a subset of main principal components is required the algorithm is more efficient. Moreover, lower memory requirements of the algorithm allows its use also in the situation of large data sets. Finally, the family of regularized least squares models in a feature space will be extended by Kernel Partial Least Squares (PLS) regression. We will demonstrate that exploiting existing correlations between regressors and response variables Kernel PLS provides models with significantly lower, qualitatively different components in comparison to Kernel PCR, while the prediction accuracy remains the same or is superior. This work was published in [112] and was accepted to a journal. Finally, the performance of all kernel based regression models will be compared on the problem of human performance monitoring during the signal detection task.

- In the second part of the thesis the usefulness of entropy rates and other complexity measures for the purpose of extracting depth of anaesthesia information from the EEG is explored. These measures will be applied to EEG data from patients undergoing general anaesthesia and will be compared with traditional spectral indices. Eight EEG series will be investigated. Two representative parts of these EEG series will be used to quantify the discriminative power of each method: a series containing moderate and light anaesthesia; and one containing the patient's emergence from anaesthesia. We will show that the complexity measures (some of them for the first time applied to EEG data measured under anaesthesia) are as good as, or better than the spectral methods at distinguishing light from moderate anaesthetic depth. Different theoretical and practical aspects of individual measures will be discussed. Part of the presented work is under review for publication in a journal. Finally, a large amount of experimental work associated with the measurement and preprocessing of EEG data measured under real surgical conditions is implicitly hidden behind the published results, however, this part of work and obtained data sets are highly contributory to the further investigations which will be undertaken in the domain.

PART A

In this part of the thesis we start with the formulation of the nonlinear regression problem and will provide several examples of nonparametric regression models. Then we focus our attention on regression methods constructed using the theoretical results of a new learning paradigm – Structural Risk Minimization (SRM) Inductive Principle – developed over the last few decades. We provide a description of the SRM Inductive Principle in sufficient detail to understand how the principle motivates the practical development of new learning algorithms. We also highlight the close connection of the SRM principle to regularization theory when the construction of regression models is considered. Finally, the first chapter of part A is concluded by providing basic definitions of a Reproducing Kernel Hilbert Space (RKHS) and by a description of the Representer Theorem which is one of the main results in the theory of learning in a RKHS. The introductory part is mainly motivated by the work published in [153, 23, 35, 156].

In the second chapter we discuss the nonlinear, kernel-based Principal Component Analysis (PCA) method. Motivated by the probabilistic linear PCA model we provide the Expectation Maximization (EM) approach to Kernel PCA and then we discuss several aspects and properties of the approach. Both standard Kernel PCA and the EM approach to Kernel PCA have been further used for the extraction of principal components employed in regression tasks.

The next chapter summarizes several nonlinear, kernel-based regression models considered in a RKHS. First, the regularized least-squares models – Kernel Partial Least Squares, Kernel Principal Component Regression and Kernel Ridge Regression – are described. The short description of Support Vector Regression (SVR) and Multi-Layer SVR is also provided. The problem of multicollinearity and its influence on the variance of the estimate of regression coefficients in the case of least-squares regression models is addressed. The connections among the individual regression models are provided and some of their properties are discussed. Finally, the model selection approaches as used in the experiments conducted are described.

In the next, experimental part, the construction and experimental settings for acquisition of the data sets employed is provided. The numerical results achieved with individual regression techniques are described and compared.

The main results and observations are summarized in the last chapter.

2. INTRODUCTION TO KERNEL LEARNING

2.1 Nonlinear Regression

One of the important tasks in mathematical statistics is to find the relationships between a set of independent variables, usually called predictor variables (inputs), and the set of dependent variables called responses (outputs). If at least one of the sets of variables is being subject to random fluctuations, possible measurement noise or other forms of randomness the problem is known as regression. While linear regression considers only linear relations between predictors and response variables, in nonlinear regression more general forms of dependencies may be assumed. Although the traditional linear regression models are attractively simple, many real life problems have a nonlinear character.

In the nonlinear regression formulation the goal is to estimate an unknown (desired) continuous and real valued function g assuming the model

$$y = g(\mathbf{x}) + \eta, \tag{2.1}$$

where η represents a random part (noise) of the observed output values $y \in \mathcal{Y} \subseteq \mathcal{R}$ and $\mathbf{x} \in \mathcal{X} \subseteq \mathcal{R}^N$ is an N-dimensional vector of input variables. We assume that the random noise η is zero mean and distributed according to the unknown probability distribution function $P(\eta)$. Based on a set of the observed input-output pairs $\{(\mathbf{x}_i, y_i)_{i=1}^n \in \mathcal{X} \times \mathcal{Y}\}$ the estimate of $g(\mathbf{x})$ is constructed. We will consider this estimate to be a linear combination of the functions $\{\Psi_i(.)\}_{i=1}^p$, $p \in \mathcal{N}$, weighted through the coefficients $\{w_i\}_{i=1}^p$ and denote it

$$f(\mathbf{x}) \equiv f(\mathbf{x}, \theta) = \sum_{i=1}^{p} w_i \Psi_i(\mathbf{B}^T \mathbf{x}) + b. \tag{2.2}$$

We assume that $\theta \in \mathbf{\Theta}$, where $\mathbf{\Theta}$ represents a set of unknown parameters $\mathbf{w} = (w_1, \ldots, w_p)^T$, $b, \mathbf{B}, p$ and parameters characterizing the functions $\{\Psi_i\}_{i=1}^p$. The coefficient b is usually called the bias and it makes the model (2.2) translation invariant. The matrix $\mathbf{B}$ extends the model by assuming different linear transformations or weighting of the original predictors[1]. Generally, the construction of the estimate $f(\mathbf{x})$ consists of two different tasks:

a) the selection of the number p and appropriate forms of the basis functions $\{\Psi_i\}_{i=1}^p$

b) the estimation of the parameters $\mathbf{B}, \mathbf{w}$ and b.

In the next two subsections we describe and provide several examples of two main statistical approaches considered in nonlinear regression.

2.1.1 Parametric vs. Nonparametric Regression

The classical statistical learning paradigm was introduced in the 1920-1930s by Fisher [24]. The paradigm is based on the estimation of an unknown, desired functional dependency

[1] In many linear or nonlinear regression models the matrix $\mathbf{B}$ is taken to be the $(N \times N)$ identity matrix; i.e. the original input representation is used. However, in some applications it may be profitable to assume the reduction of a possibly high-dimensional input space N to more compact $K < N$ representation. Examples of this can be the transformation of $\mathbf{x}$ given by the $(N \times K)$ matrix $\mathbf{B}$ consisting of the K eigenvectors found by Principal Component Analysis. Assuming $\mathbf{B}$ to be the $(N \times N)$ diagonal with different values on the diagonal leads to a weighting of the original predictors.

g using *a priori* knowledge about the dependency up to the values of a finite number of parameters. Therefore estimation of the parameters represents the problem of dependency estimation. An example may be the estimator with the polynomial basis functions up to the order p; i.e.

$$f(\mathbf{x}) = \sum_{i=1}^{p} w_i \mathbf{x}^i + b.$$

Based on the data provided we only need to estimate the unknown coefficients $\{w_i\}_{i=1}^{p}$ and b whilst the maximum order p and thus the form of nonlinearity is given *a priori*.

In the last 20-30 years a new statistical learning theory has been investigated which overcomes some of the deficiencies of Fisher's paradigm. The new paradigm is based on the assumption that in order to estimate dependency from the data, it is sufficient to know some general properties of the set of functions to which the unknown dependency belongs [153]. In our context this paradigm represents the nonparametric approach to the regression. An example of nonparametric regression, which makes minimal assumptions about the dependency of the outputs on the input variables, is the spline smoothing method [64, 156]. In this case we are giving no explicit parameterization to (2.2) and assume that the estimate lives in the infinite dimensional space of all continuous functions of $\mathbf{x}$. The final model which will take the form (2.2) is then given by the learning process itself. In [153], Vapnik describes the basic principles of the new theory and explains the core of the new approach – the *Structural Risk Minimization* (SRM) inductive principle through which a learning process is defined. Before we give a more detailed description of this paradigm in section 2.2, we briefly review several nonparametric regression approaches to the construction of the estimate (2.2).

2.1.2 Nonparametric Regression Models

Additive Models (AM)

AM were proposed in [12]. The simplest AM have the form

$$f(\mathbf{x}) = \sum_{j=1}^{N} f_j(x_j) + b,$$

where x_j is the jth predictor in the observation $\mathbf{x}$ and b is a bias term. The functions f_j are *a priori* unknown and are estimated from the data. AM assume that the predictors have an additive effect. Thus, the response variable is modeled as the sum of arbitrary smooth univariate functions of the predictors. The backfitting algorithm described in [43] is used to find the best AM model based on the data provided.

Projection Pursuit Regression (PPR)

PPR was designed to handle cases when the underlying function is additive with respect to linearly transformed predictors rather than the original predictor variables [28]. PPR has a form similar to AM

$$f(\mathbf{x}) = \sum_{j=1}^{p} f_j(\mathbf{w}_j^T \mathbf{x}) + b,$$

where the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_p$ determine a set of p linear combinations of the predictors. These vectors may be found through the cross-validation technique and are not necessarily orthogonal. The second step then consists of selecting the appropriate functions f_j and the same approaches as used in AM may be used here.

Spline Models

Consider the interval $[a, b]$ and k *knots* $\{t_i\}_{i=1}^{k}$ splitting the interval in the following way $-\infty \leq a < t_1 < t_2 < \ldots < t_k < b \leq \infty$. The (univariate, natural) polynomial spline is a real valued function $s(x)$ having the following properties:

1. $s(x) \in p^{m-1}$ for $x \in [a, t_1], x \in [t_k, b]$
2. $s(x) \in p^{2m-1}$ for $x \in [t_i, t_{i+1}], i = 1, \dots, k-1$
3. $s(x) \in C^{2m-2}$ for $x \in (-\infty, \infty)$

where p^r is the set of polynomials of degree r or less and C^r is a set of functions with r continuous derivatives.

Spline models are based on the assumption that in regression we prefer smooth regression estimates rather than interpolation of the observed data. Then, for $\xi > 0$ we look for the estimate $f^{\xi}(x)$ given by minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(t_i))^2 + \xi \int_a^b (f^m(x))^2 dx,$$

where f belongs to a Hilbert space of functions with $m-1$ continuous derivatives and the mth derivative square integrable. It was shown by Schoenberg [118, 117], that the minimizer is a natural polynomial spline. The parameter ξ controls the trade off between fit to the data and the *smoothness* given by the squared integral of the mth derivative of the solution. For further details on spline models and extension to additive spline models in the case of multivariable inputs see [156].

Kernel Regression (KR)

KR or the Nadaraya-Watson estimator is based on the estimation of the joint probability distribution $P(\mathbf{x}, y)$ of predictor and response variables from the final number n of observed samples $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ using the Parzen window estimator

$$P(\mathbf{z}) = \frac{1}{nh}\sum_{i=1}^{n} u\left(\frac{\mathbf{z} - \mathbf{z}_i}{h}\right),$$

where $u(.)$ is an appropriate kernel and h is a positive parameter. Choosing $u(.)$ to be of the form

$$u(\mathbf{z}) = K(\|\mathbf{x}\|)K(y),$$

where K is a one-dimensional, symmetric kernel, leads to the Nadaraya-Watson estimator [83, 159]

$$f(\mathbf{x}) = \frac{\sum_{i=1}^{n} y_i K(\|\mathbf{x} - \mathbf{x}_i\|)}{\sum_{i=1}^{n} K(\|\mathbf{x} - \mathbf{x}_i\|)}.$$

There is a wide literature published on this type of regression and the appropriate forms of the kernel K and its parameters are discussed there (see e.g. [32, 125, 42] and ref. therein).

Regularization Networks (RN)

RN [35] represent a larger family of models minimizing the functional

$$H[f] = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2 + \xi\Omega(f),$$

where f belongs to some space of real-valued functions, ξ is a positive constant (*regularization term*) and $\Omega(f)$ is a *smoothness functional* defining properties of the final estimate $f(\mathbf{x})$. We provide more detailed description of RN later, however, for the moment we note that AM, PPR, splines models and KR approaches may be straightforwardly extended also into the context of RN [35]. One of the most well-known examples of RN are radial-basis function networks [45, 35].

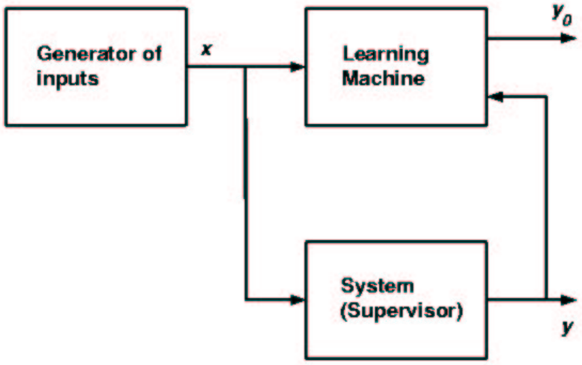

Fig. 2.1: Block diagram of general learning process from examples. During the learning process the learning machine receives the pairs of samples (x, y) from the generator of inputs and the system, respectively. The goal of learning is to approximate the behavior of the system; i.e. for any given input x return a value y_0 close to the system response y.

Artificial Neural Networks (ANN)

A typical representative of an ANN is a feed-forward network of the form

$$f(\mathbf{x}) = w_0 + \sum_{i=1}^{p} w_i u(\mathbf{a}_i^T \mathbf{x} + a_{i0}),$$

where $u(.)$ is an activation function (usually logistic or hyperbolic tangent function) [45]. The back-propagation algorithm is used to estimate unknown parameters $\{w_i, \mathbf{a}_i, a_{i0}\}_{i=1}^{p}$. The parameter p represents the number of hidden nodes (neurons) and is usually tuned during the learning process. As ANN are not our main interest in this thesis we refer the reader to [45] describing the model and learning strategies in more detail.

2.2 Structural Risk Minimization Inductive Principle

Consider the general process of learning from examples as it is shown in Figure 2.1 and described in [153]. The first block represents a generator of random input vectors $\mathbf{x} \in \mathcal{X}$, drawn independently from a fixed but unknown probability distribution $P(\mathbf{x})$. The system (supervisor) represents a functional block which for every input $\mathbf{x}$ returns an output value y according to a fixed, unknown conditional distribution function $P(y|\mathbf{x})$. Learning machines represent a set of approximation functions $\mathcal{A}$.[2] The problem of learning is to find the function $f \in \mathcal{A}$ which best approximates the system's response. This selection is based on the observed independent identically distributed (iid) data pairs $(\mathbf{x}_i, y_i)_{i=1}^{n}$ drawn according to the joint probability distribution function $P(\mathbf{x}, y) = P(\mathbf{x})P(y|\mathbf{x})$.

Now, consider the problem of regression estimation. We assume that the system output y is real valued and that the set $\mathcal{A}$ is now a set of real functions $f(\mathbf{x})$. The unknown function $g(\mathbf{x})$ from (2.1) is the mean of the output conditional probability; i.e. the *regression function*

$$g(\mathbf{x}) = \int y dP(y|\mathbf{x}). \tag{2.3}$$

The goal of regression estimation is to find a function $f_0(\mathbf{x})$ from $\mathcal{A}$ which provides the best approximation to the regression function. Although, later we will consider that this

[2] SRM principle is a general learning theory and does not restrict the set of admissible functions to a specific form. In fact, we may consider any set of functions.

approximation function $f_0(\mathbf{x})$ will have the form (2.2), for the current theoretical description of the SRM inductive principle we assume a rather general form for functions belonging to the set $\mathcal{A}$.

To find the function $f_0(\mathbf{x})$ the following *risk functional*

$$R[f] = \int V(y, f(\mathbf{x}))dP(\mathbf{x}, y) \tag{2.4}$$

is minimized over the class of functions $\mathcal{A}$. $V(y, f(\mathbf{x}))$ is an appropriately chosen cost-function to measure the difference between the system's response y and the response of the learning machine $f(\mathbf{x})$ to a given input $\mathbf{x}$. In the case that V is of the form $(y - f(\mathbf{x}))^2$ the ideal estimator $f_0(\mathbf{x})$ is the regression function (2.3). This simply means that we may obtain the desired regression function $g(\mathbf{x})$ only if it is contained in $\mathcal{A}$ otherwise the estimate $f_0(\mathbf{x})$ is the closest to the regression function in the metric $L_2(\mathcal{X})$:

$$d(g(\mathbf{x}), f_0(\mathbf{x})) = \sqrt{\int (g(\mathbf{x}) - f_0(\mathbf{x}))^2 dP(\mathbf{x})}.$$

Whereas in practice we usually do not know the probability distribution function $P(\mathbf{x}, y)$, the following inductive principle consists of replacing risk functional (2.4) by the so-called *empirical risk functional* [153]

$$R_{emp}[f, n] = \frac{1}{n}\sum_{i=1}^{n} V(y_i, f(\mathbf{x}_i)). \tag{2.5}$$

Thus, using a limited set of observations $(\mathbf{x}_i, y_i)_{i=1}^n \in \mathcal{X} \times \mathcal{Y}$ we approximate the function $f_0(\mathbf{x})$ which minimizes risk (2.4) by the function $f_0^n(\mathbf{x})$ minimizing the empirical risk (2.5). This principle of replacing the risk functional (2.4) by (2.5) is called the *Empirical Risk Minimization* inductive principle (ERM principle) [153]. The ERM principle is quite a general concept and is also included in the classical solution to the problem of regression estimation. This will be discussed in the next subsection.

The obvious question when the ERM principle is used is how close the estimate $f_0^n(\mathbf{x})$ will be to the ideal estimate $f_0(\mathbf{x})$ given by the minimization of (2.4). Consider the case where the set $\mathcal{A}$ is too complex; i.e. it contains functions which can almost perfectly fit the outputs y of the system, and that we have only a restricted, relatively small number of observed examples. It is possible that we may attain a zero value of (2.5) for some $f_0^n(\mathbf{x})$, however this will not guarantee that this estimate will be close to the $f_0(\mathbf{x})$. This intuitive result is given by the fact that $f_0^n(\mathbf{x})$ will also perfectly 'copy' the noise component of the observed outputs. In addition, minimization of (2.5) is usually badly determined in the sense that several different solutions may exist [142].

The problem of the convergence of $f_0^n(\mathbf{x})$ to the ideal estimate $f_0(\mathbf{x})$ was intensively studied by Vapnik and resulted in the method of *Structural Risk Minimization* (SRM) [152, 153]. The SRM principle is based on the idea of the restriction of the complexity of the space $\mathcal{A}$ by reducing a set of all assumed approximation functions. This will guarantee the avoidance of too 'tight' a fit of the observed noisy outputs; i.e. the problem of *overfitting* or bad generalization properties of the final model $f_0^n(\mathbf{x})$. The basic principle of SRM is to construct a nested sequence of the smaller function spaces $\mathcal{A}_1 \subset \mathcal{A}_2 \subset \ldots \subset \mathcal{A}_P$. The ordering of the $\mathcal{A}_i$'s is given by the order $c_1 \leq c_2 \leq \ldots \leq c_P$, where c_i is the finite quantity measuring the *capacity* (complexity) of the function space $\mathcal{A}_i$. The Vapnik-Chervonenkis (VC) dimension, introduced by Vapnik and Chervonenkis [154], became one the most popular measures to quantify the capacity of a set of functions. The SRM then combines the quality of the approximation given by the minimization of the empirical risk functional (2.5) as well as controlling the complexity of the approximating function. Mathematically this allows us

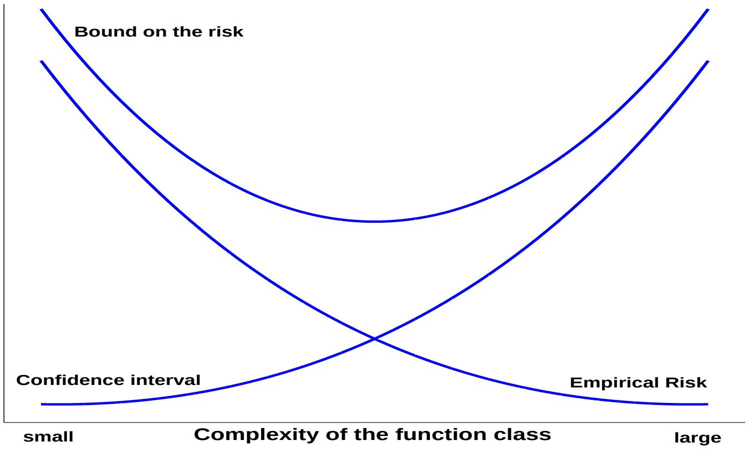

Fig. 2.2: Schematic illustration of bound on the risk as a summation of the empirical risk and of the confidence interval. Increasing the complexity of the function class decreases the empirical risk but upper bound of the risk is increased due to the increase of confidence interval. The smallest bound of the risk is a trade off between the empirical risk and complexity (confidence).

to define general forms for probabilistic bounds on the distance between the risk functional (2.4) and the empirical risk functional (2.5); i.e. with probability at least κ

$$R[f] \leq R_{emp}[f,n] + \beta(\sqrt{\frac{c}{n}}, \kappa), \tag{2.6}$$

where c is the capacity, n the number of samples and β is an increasing function of $\frac{c}{n}$ and κ. Following [153] (page 92) this can be graphically illustrated as in Figure 2.2.

Although, the SRM principle provides a theoretical framework for the most accurate estimate of $f_0(\mathbf{x})$, the practical implementation of SRM with the VC dimension or more appropriate V_γ dimension (or closely related *fat-shattering* dimension) [62, 1, 5, 22] in the case of real valued functions is a very difficult problem (see e.g. [23]). In spite of this fact, the SRM principle together with the existing regularization theory motivated the construction of new types of learning machines – Support Vector Machines (SVM). This connection between SRM principle and regularization theory will be discussed in section 2.3. However, before doing this, we will examine regression estimation in the classical paradigm.

2.2.1 Classical Paradigm of Regression Estimation

The classical paradigm for regression estimation is usually based on an *a priori* given parametric form of an unknown functional dependency. The maximum likelihood (ML) method is then used as the basic inductive tool for the estimation of unknown parameters based on the data provided.

Consider an unknown function $g(\mathbf{x})$ from (2.1) which has the parametric form $f(\mathbf{x}, \theta_0), \theta_0 \in \boldsymbol{\Theta}$. Further assume that the additive noise component η is distributed according to a known probability density function $p(\eta)$. Given the observed data pairs $(\mathbf{x}_i, y_i)_{i=1}^n$ the ML principle can be used to estimate an unknown vector of parameters θ_0. This estimate is given by maximizing of the functional

$$L(\theta) = \sum_{i=1}^{n} \ln p(y_i - f(\mathbf{x}_i, \theta)), \quad \theta \in \boldsymbol{\Theta}.$$

It is well known that in the case that the noise is normally distributed with zero mean and some fixed covariance matrix the ML estimate coincides with the minimization of the functional [52]

$$L^*(\theta) = \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i, \theta))^2, \;\; \theta \in \mathbf{\Theta}.$$

Effectively it means we are minimizing the empirical risk functional (2.5) using the quadratic loss function V. This will provide the function which gives the best least squares approximation to the data. However, when the noise is distributed according to a different law the choice of a different cost function may be profitable. This issue is discussed in the following subsection.

2.2.2 Cost Functions

In the case of the considered regression models (2.2) (i.e. regression models linear in regression coefficients $\{w_i\}_{i=1}^{p}$) the Gauss-Markov theorem states that the least squares estimate of the vector of unknown coefficients $\mathbf{w}$ is the linear unbiased estimate with the minimum variance among all the other linear unbiased estimates (see e.g. [93]). Thus, in the case that the additive noise η is normally distributed it also provides the best approximation to the regression function. This is not true if the additive noise is distributed according to a different unknown distribution and the optimal ERM principle approximation to the regression function leads to loss-function associated with this distribution. The choice of different cost functions in dependence on a considered noise distribution was intensively studied by Huber and gave rise to the so-called *robust regression* [52]. Motivated by this result, for the class of densities 'close' to the uniform distribution, Vapnik [151] introduced a ϵ-insensitive cost function of the form

$$V(y, f(\mathbf{x})) = |f(\mathbf{x}) - y|_\epsilon = \left\{ \begin{array}{rcl} 0 & : & |f(\mathbf{x}) - y| \leq \epsilon \\ |f(\mathbf{x}) - y| - \epsilon & : & otherwise \end{array} \right. \tag{2.7}$$

Other types of noise distributions and the corresponding cost functions generally used in the context of studied kernel-based learning were discussed in detail in [128, 127].

2.3 Regularization Theory

In the previous section we pointed out that the direct minimization of the functional (2.5) may lead to the problem of overfitting, i.e. the problem of bad generalization. It can happen when the capacity of the set of functions $\mathcal{A}$ is very high and we are dealing with a data set which is too small or contains only a limited amount of information about the desired dependency. We have shown that the recently elaborated SRM inductive principle provides a very general, theoretically founded tool to give the solution to the problem. However, the problem of bad generalization is not new; in fact, Hadamard, at the beginning of the XX. century, observed that solving the linear operator (mapping from a metric space $\mathcal{M}_1$ to a metric space $\mathcal{M}_2$) equation:

$$Lh = H, \;\; h \in \mathcal{M}_1, H \in \mathcal{M}_2$$

is *ill-posed*[3] in the sense that, even if there exists a unique solution, a small deviation of $H \rightarrow H_\delta$ (e.g. by some noise level δ) can lead to large deviations of the solution h_δ from the ideal solution h. This may happen also in the case where the level of noise corruption δ decreases to zero. To overcome this problem, in 1962-63 regularization theory was proposed by Tikhonov [141], Ivanov [55] and Philips [96]. With the aim of solving ill-posed problems, it was discovered that minimizing a regularized functional

$$R_{reg}(h) = ||Lh - H_\delta||^2 + \zeta_\delta \Omega(h),$$

[3] Under some (very general) conditions.

where $\Omega(h)$ is some functional with some specific regularizing properties and ζ_δ is an appropriately chosen, noise dependent constant, leads to the sequence of solutions that converges to the desired solution h as δ tends to zero. In our case the concepts of regularization theory and ERM inductive principle lead to the problem of minimizing the following functional:

$$R_{reg}(h) = R_{emp} + \xi\Omega(h) \tag{2.8}$$

where ξ is a regularization constant to control the trade off between model complexity and approximation accuracy in order to achieve good generalization performance. Several algorithms which lead to minimizing a similar regularized risk functional were described in [8, 45].

In this thesis, we will construct approximation functions of the form (2.2) belonging to some functional Hilbert space, more specifically to a Reproducing Kernel Hilbert Space $\mathcal{H}$ described in the following section. Evgeniou et al. [23] have shown the connection between the estimation methodologies motivated by the SRM principle and regularization theory, respectively, in the case that an approximation of the desired functional dependency g (2.1) is considered to belong to a RKHS. Inspired by the SRM principle they proposed a learning strategy based on the construction of a nested structure of sets of functions ordered with respect to an increasing real-valued sequence $a_1 < a_2 < \ldots < a_P$ associated with a norm $\|.\|_\mathcal{H}$ defined in $\mathcal{H}$. The sequence of sets of functions has the form:

$$F_1 \subset F_2 \subset \ldots \subset F_P, \tag{2.9}$$

where $\{F_j = \{f \in \text{RKHS} : \|f\|_\mathcal{H} \leq a_j\}\}_{j=1}^P$. By construction, the capacity of the sets of functions $\{F_j\}_{j=1}^P$ (in terms of V_γ dimension) will increase according to the increase of a_i [23, 22]. Having this nested structure, we need to minimize the empirical risk (2.5) over the individual sets of functions $\{F_j\}_{j=1}^P$; i.e. to solve the following constrained minimization problem for all $j = 1, \ldots P$

$$\begin{aligned} &\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n V(y_i, f(\mathbf{x}_i)) \\ &\text{subject to} : \|f\|_\mathcal{H}^2 \leq a_j^2 \end{aligned} \tag{2.10}$$

The problem can be solved by the technique of Lagrange multipliers leading to the minimization of the form

$$\frac{1}{n}\sum_{i=1}^n V(y_i, f(\mathbf{x}_i)) + \xi_j(\|f\|_\mathcal{H}^2 - a_j^2) \quad \forall j = 1, \ldots P \tag{2.11}$$

with respect to f and maximization with respect to Lagrange multiplier $\xi_j \geq 0$. Assuming the structure (2.9) this will provide us the sequence of the solutions $\{f_j^*\}_{j=1}^P$ and the sequence of the corresponding optimal Lagrange multipliers $\{\xi_j^*\}_{j=1}^P$. Then, the optimal solution f_{opt} selected from the set of all solutions $\{f_j^*\}_{j=1}^P$ is given by the trade off between two terms of the right hand side of (2.6); i.e. the empirical error and the capacity of the corresponding functional subsets. However, in practice, this theoretical concept is difficult to implement mainly because of computational difficulties when the solutions of a large number of the constrained optimization problems (2.10) have to be found[4]. To overcome this difficulty Evgeniou et al. [23] proposed to search for the minimum of

$$\frac{1}{n}\sum_{i=1}^n V(y_i, f(\mathbf{x}_i)) + \xi\|f\|_\mathcal{H}^2 \tag{2.12}$$

[4] Other practical difficulties associated with the direct implementation of SRM principle were discussed in [23].

instead. This is motivated by the fact that having the optimal value ξ_j^* to find the optimal solution f_{opt} we could simply replace (2.11) by the unconstrained minimization problem

$$\frac{1}{n}\sum_{i=1}^{n} V(y_i, f(\mathbf{x}_i)) + \xi_j^* \|f\|_{\mathcal{H}}^2. \tag{2.13}$$

Thus, in practice we need to solve (2.12) for different values of ξ and to pick up the best ξ usually based on some model selection criteria. We have to stress that the main goal of the approach motivated by the SRM principle was to show the connection between this theoretically well founded statistical learning principle and its possible practical approximation. For more detailed discussions on implementation of the SRM principle and some aspects of replacing problem (2.10) by (2.12) we refer the reader to [153, 23].

In the last section of this introductory chapter we will give the basic formal definition and describe some of the properties of a RKHS. We will also provide a general form to the solution of (2.12) when a functional space is a RKHS.

2.4 Learning in Kernel-Induced Feature Spaces

Before we will proceed to the definition and description of the properties of a RKHS we will review the definition of real Hilbert and $L_2(\mathcal{X})$ spaces.

Definition: A Hilbert space is a separable real inner product space that is complete in the metric derived from its inner product. An inner product space $\mathcal{V}$ is separable if it contains a sequence of elements $m_1, m_2, \ldots$ that span a dense subspace of $\mathcal{V}$.

Definition: Let $\mathcal{X}$ be a compact subset of $\mathcal{R}^N$. A Hilbert space $L_2(\mathcal{X})$ is the set of real valued functions f defined over $\mathcal{X}$ for which

$$\|f\|_{L_2} = \left[\int_{\mathcal{X}} f(\mathbf{x})^2 d\mathbf{x}\right]^{1/2}.$$

The formula

$$\langle f, g\rangle = \int_{\mathcal{X}} f(\mathbf{x})g(\mathbf{x})d\mathbf{x}$$

defines an inner product on $L_2(\mathcal{X})$.

2.4.1 Reproducing Kernel Hilbert Space

A RKHS is a Hilbert space $\mathcal{H}$ of the functions defined over some compact set $\mathcal{X} \subset \mathcal{R}^N$ with the property that all the evaluation functionals $T_{\mathbf{x}}[f] = f(\mathbf{x})$, $\forall f \in \mathcal{H}$ are bounded [4].

To better understand this formal definition we will now provide several basic properties of a RKHS. First, consider a symmetric function $K(\mathbf{x}, \mathbf{y})$ of two variables satisfying the Mercer theorem conditions [76]:

Theorem (Mercer): Let $\mathcal{X}$ be a compact subset of $\mathcal{R}^N$. Suppose K is a continuous symmetric function such that the integral operator $T_K : L_2(\mathcal{X}) \to L_2(\mathcal{X})$,

$$(T_K f)(.) = \int_{\mathcal{X}} K(., \mathbf{x})f(\mathbf{x})d\mathbf{x} \tag{2.14}$$

is positive, i.e. for all $f \in L_2(\mathcal{X})$ we have

$$\int_{\mathcal{X}\times\mathcal{X}} K(\mathbf{x}, \mathbf{y})f(\mathbf{x})f(\mathbf{y})d\mathbf{x}d\mathbf{y} \geq 0.$$

Then the kernel function $K(\mathbf{x},\mathbf{y})$ can be expanded in a uniformly convergent series

$$K(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{M} \lambda_i \psi_i(\mathbf{x})\psi_i(\mathbf{y}) \quad M \leq \infty, \tag{2.15}$$

where $\{\psi_i(.)\}_{i=1}^{M}$, are the normalized eigenfunctions (and so $\{\|\psi_i\|_{L_2} = 1\}_{i=1}^{M}$) of the integral operator T_K (2.14) and $\{\lambda_i > 0\}_{i=1}^{M}$ are the corresponding positive eigenvalues. In the case $M = \infty$ the series converges absolutely and uniformly for almost all $(\mathbf{x}, \mathbf{y})$.

The fact that for any such positive definite kernel, there exists a unique RKHS is well established by the *Moore-Aronszajn theorem* [4]. Further, the form $K(\mathbf{x},\mathbf{y})$ has the following *reproducing property*

$$f(\mathbf{x}) = \langle f(.), K(\mathbf{x}, .)\rangle_{\mathcal{H}} \;\; \forall f \in \mathcal{H}, \tag{2.16}$$

where $\langle ., .\rangle_{\mathcal{H}}$ is the scalar product in $\mathcal{H}$. The function K is called a *reproducing kernel* for $\mathcal{H}$ (hence the terminology RKHS). This reproducing property implies that the evaluation functionals defined by $T_{\mathbf{x}}[f] = f(\mathbf{x})$, $\forall f \in \mathcal{H}$ are linear and bounded. The boundedness means that there exists $M_{\mathbf{x}} \in \mathcal{R}^+$ such that $|T_{\mathbf{x}}[f]| \leq M_{\mathbf{x}}\|f\|_{\mathcal{H}}$. In our case it simply means that by using the Cauchy-Schwarz inequality we have

$$|T_{\mathbf{x}}[f]| = |f(\mathbf{x})| = \langle f(.), K(\mathbf{x}, .)\rangle_{\mathcal{H}} \leq \|K(\mathbf{x}, .)\|_{\mathcal{H}}\|f\|_{\mathcal{H}} = M_{\mathbf{x}}\|f\|_{\mathcal{H}} \;\; \forall f \in \mathcal{H},$$

where $\|.\|_{\mathcal{H}}$ is a norm defined in $\mathcal{H}$ and its exact form will be described below.

It follows from Mercer's theorem that the sequence $\{\psi_i(.)\}_{i=1}^{M}$ creates an orthonormal basis in $\mathcal{H}$ and we can express any function $f \in \mathcal{H}$ as $f(\mathbf{x}) = \sum_{i=1}^{M} d_i\psi_i(\mathbf{x})$ for some $d_i \in \mathcal{R}$. However, it is worth noting that, we can also construct a RKHS by choosing a sequence of linearly independent functions (not necessary orthogonal) $\{\phi_i(\mathbf{x})\}_{i=1}^{M}$ and positive numbers $\{\alpha_i\}_{i=1}^{M}$ to define a uniformly convergent series (in the case of $M = \infty$ absolutely and uniformly convergent)

$$K(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{M} \alpha_i\phi_i(\mathbf{x})\phi_i(\mathbf{y}). \tag{2.17}$$

This construction also gives the connection between the RKHS and stochastic processes [156] where the K is assumed to represent the correlation function of a zero-mean Gaussian stochastic process evaluated at points $\mathbf{x}$ and $\mathbf{y}$. Similar to the orthogonal basis case we can express any function $f \in \mathcal{H}$ in the form $f(\mathbf{x}) = \sum_{i=1}^{M} b_i\phi_i(\mathbf{x})$ for some $b_i \in \mathcal{R}$. This allows us to define a scalar product in $\mathcal{H}$:

$$\langle h(\mathbf{x}), f(\mathbf{x})\rangle_{\mathcal{H}} = \langle \sum_{i=1}^{M} a_i\phi_i(\mathbf{x}), \sum_{i=1}^{M} b_i\phi_i(\mathbf{x})\rangle_{\mathcal{H}} \equiv \sum_{i=1}^{M} \frac{a_i b_i}{\alpha_i}$$

and the norm

$$\|f\|_{\mathcal{H}} = \langle f(\mathbf{x}), f(\mathbf{x})\rangle_{\mathcal{H}}^{1/2} = \left(\sum_{i=1}^{M} \frac{b_i^2}{\alpha_i}\right)^{1/2}. \tag{2.18}$$

The motivation to define this kind of norm is that we need to satisfy the reproducing property (2.16) of the kernel function $K(\mathbf{x},\mathbf{y})$; i.e.

$$f(\mathbf{x}) = \langle f(\mathbf{y}), K(\mathbf{x},\mathbf{y})\rangle_{\mathcal{H}} = \sum_{i=1}^{M} \frac{b_i\alpha_i\phi_i(\mathbf{x})}{\alpha_i} = \sum_{i=1}^{M} b_i\phi_i(\mathbf{x})$$

Rewriting (2.17) in the form

$$K(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{M} \sqrt{\alpha_i}\phi_i(\mathbf{x})\sqrt{\alpha_i}\phi_i(\mathbf{y}) = (\Phi(\mathbf{x}).\Phi(\mathbf{y})) \tag{2.19}$$

it becomes clear that any kernel $K(\mathbf{x}, \mathbf{y})$ also corresponds to a canonical (Euclidean) dot product in a possibly high dimensional space $\mathcal{F}$ where the input data are mapped by

$$\begin{aligned} \Phi: \quad & \mathcal{X} \to \mathcal{F} \\ & \mathbf{x} \to (\sqrt{\alpha_1}\phi_1(\mathbf{x}), \sqrt{\alpha_2}\phi_2(\mathbf{x}), \ldots, \sqrt{\alpha_M}\phi_M(\mathbf{x})) \end{aligned} \tag{2.20}$$

The space $\mathcal{F}$ is usually denoted as a *feature space* and $\{\{\phi_i(\mathbf{x})\}_{i=1}^M, \mathbf{x} \in \mathcal{X}\}$ as *feature mappings.* The number of basis functions $\phi_i(.)$ also defines the dimensionality of $\mathcal{F}$.

2.4.2 Representer Theorem

One of the main results in the theory of learning in a RKHS $\mathcal{H}$ was given by Kimeldorf and Wahba [65, 156, 157] and is known as the

Representer Theorem (simple case)*:* Let the loss function $V(y_i, f)$ be a functional of f which depends on f only pointwise, that is, through $\{f(\mathbf{x}_i)\}_{i=1}^n$ – the values of f at the data points. Then any solution to the problem: find $f \in \mathcal{H}$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n} V(y_i, f(\mathbf{x}_i)) + \xi \|f\|_{\mathcal{H}}^2 \tag{2.21}$$

has a representation of the form

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i K(\mathbf{x}_i, \mathbf{x}), \tag{2.22}$$

where $\{c_i\}_{i=1}^n \in \mathcal{R}$.

In regularization theory, ξ is a positive number (regularization term) to control the trade-off between approximating properties and the smoothness of f and the squared norm $\|f\|_{\mathcal{H}}^2$ is sometimes called the 'stabilizer'. Moreover, the above results can be extended even for the case when K is positive semidefinite. In such a case a RKHS $\mathcal{H}$ contains a subspace of functions f with a zero norm $\|f\|_{\mathcal{H}}$ (the null space). Kimeldorf and Wahba have also shown [65, 156, 157] that in such a case the solution of (2.21) has the more general form

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i K(\mathbf{x}_i, \mathbf{x}) + \sum_{j=1}^{l} d_j \zeta_j(\mathbf{x}), \tag{2.23}$$

where the functions $\{\zeta_j(.)\}_{j=1}^l$ span the null space of $\mathcal{H}$ and the coefficients $\{c_i\}_{i=1}^n$, $\{d_j\}_{j=1}^l$ are again given by the data. In the thesis we will consider only the case when $l = 1$ and $\zeta_1(\mathbf{x}) = const \ \ \forall \mathbf{x}$.

3. NONLINEAR, KERNEL-BASED PCA

Linear principal component analysis (PCA) is a well established, and one of the oldest, techniques of multivariate analysis. The central idea of PCA is the dimensionality reduction of a data set when there exist some correlations among the variables. PCA transforms a number of correlated variables into a smaller number of orthogonal, i.e. uncorrelated, variables called *principal components*. Thus the reduction or so-called *feature extraction* allows us to restrict the entire space to a subspace of a lower dimensionality. Before we give the description of the nonlinear, kernel-based PCA in some possibly high-dimensional feature space $\mathcal{F}$ we will describe the standard linear PCA algorithm and its kernel version [60, 171].

3.1 Linear PCA

Let $\tilde{\mathbf{X}}$ denote an N-dimensional random vector representing the data domain of interest and assume we have n samples (realizations) $\{\mathbf{x}_i\}_{i=1}^n \in \mathcal{X} \subseteq \mathcal{R}^N$ of $\tilde{\mathbf{X}}$. Further, assume the random vector $\tilde{\mathbf{X}}$ has zero-mean (i.e. $E[\tilde{\mathbf{X}}] = \mathbf{0}$, where E represents statistical expectation) and that the sample-based estimate of the positive semidefinite $(N \times N)$ covariance matrix $\mathbf{C} = E[\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}]$ of $\tilde{\mathbf{X}}$ has the form

$$\hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T = \frac{1}{n}\mathbf{X}^T\mathbf{X}, \qquad (3.1)$$

where $\mathbf{X}$ represents the $(n \times N)$ matrix consisting of the observed samples. The main goal of PCA is to find the direction of maximum variance, i.e. the directions where the data $\mathbf{x}$ have maximal spread. This is given by the solution of the *eigenvalue problem*

$$\mathbf{Cu} = \lambda\mathbf{u} \qquad (3.2)$$

which has a nontrivial solution only for special values of λ that are called *eigenvalues* of the covariance matrix $\mathbf{C}$. The associated unit vectors $\mathbf{u}$ are called *eigenvectors*. Numerically the problem is solved by the diagonalization of the $(N \times N)$ matrix $\hat{\mathbf{C}}$ leading to the estimate of the sequence of eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N \geq 0$ and the corresponding eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N$.

After the extraction of the eigenvalues and eigenvectors we may project the original data onto $p \leq N$ eigenvectors based on some *a priori* given criterion. The projection is simply given by $\mathbf{P} = \mathbf{XU}$ where $\mathbf{U}$ is a $(N \times p)$ matrix with columns created by the selected p eigenvectors.

3.1.1 Kernel PCA

It is not uncommon in real world problems that the number of observed variables significantly exceeds the number of measurements (samples). In such a case $\mathbf{X}$ will be a 'wide' $(n << N)$ matrix, usually consisting of highly collinear data; i.e. there exist linear or near-linear dependencies among the variables. Further, if the number of variables N is high the diagonalization and the storage of the $(N \times N)$ sample covariance matrix $\hat{\mathbf{C}}$ will lead to high computational and memory requirements. Under the assumption that the first $p < n$ principal components cover almost all variance in the observed data structure the kernel based approaches to the extraction of principal components can be used [171]. We will describe the method when the principal components are extracted based on the diagonalization of $\mathbf{K} = \mathbf{X}\mathbf{X}^T$ the $(n \times n)$ matrix as this will have the straightforward connection to the nonlinear kernel-based PCA algorithm described in the next section.

Assuming that in (3.2) we replaced the covariance matrix $\mathbf{C}$ by its sample estimate (3.1) we can write

$$\hat{\mathbf{C}}\mathbf{u} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \mathbf{u} = \frac{1}{n}\sum_{i=1}^{n} a_i \mathbf{x}_i,$$

where $\{a_i = \mathbf{x}_i^T \mathbf{u} = (\mathbf{x}_i.\mathbf{u})\}_{i=1}^n$ represents the canonical dot product between the $\mathbf{x}_i$ and $\mathbf{u}$. Thus, using (3.2) we can see that for $\lambda \neq 0$ we have $\mathbf{u} = \frac{1}{\lambda n}\sum_{i=1}^n a_i \mathbf{x}_i$; i.e. all the non-zero solutions of (3.2) corresponding to non-zero values of λ have to lie in the span of the data samples $\{\mathbf{x}_i\}_{i=1}^n$. This allows us to rewrite (3.2) for the sample estimate $\hat{\mathbf{C}}$ in the form

$$\mathbf{x}_j^T \hat{\mathbf{C}}\mathbf{u} = \lambda \mathbf{x}_j^T \mathbf{u} \quad \text{for all } \; j = 1, 2, \ldots n$$

or in the matrix form

$$\mathbf{X}\hat{\mathbf{C}}\mathbf{u} = \lambda \mathbf{X}\mathbf{u}.$$

Using the fact that $\mathbf{X}\hat{\mathbf{C}}\mathbf{u} = \frac{1}{n}\mathbf{X}\mathbf{X}^T\mathbf{X}\mathbf{u} = \frac{1}{n}\mathbf{K}\mathbf{X}\mathbf{u}$ we obtain the eigenvalue problem

$$\mathbf{K}\tilde{\mathbf{u}} = n\lambda\tilde{\mathbf{u}} = \tilde{\lambda}\tilde{\mathbf{u}} \tag{3.3}$$

solution of which will lead to the extraction of the eigenvectors $\tilde{\mathbf{u}} = \mathbf{X}\mathbf{u}$ and eigenvalues $\tilde{\lambda} = n\lambda$ of the $\mathbf{K}$ matrix. At the beginning of the previous section we assumed that the individual observed variables are zero-mean. Having constructed the $\mathbf{K}$ matrix from the given raw data we can construct its 'centralized' version corresponding to the kernel matrix of the centralized data by [172, 121]

$$\mathbf{K} \leftarrow (\mathbf{I} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)\mathbf{K}(\mathbf{I} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T), \tag{3.4}$$

where $\mathbf{I}$ is n dimensional identity matrix and $\mathbf{1}_n$ represent the vector of ones of length n. Because the matrix $(\mathbf{I} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)$ is of rank $(n-1)$ the centralized $\mathbf{K}$ matrix will have rank less than or equal to $(n-1)$; i.e. $rank(\mathbf{K}) \leq (n-1)$. Effectively it means that by solving (3.3) using the centralized $\mathbf{K}$ matrix, we may obtain up to $(n-1)$ different non-zero eigenvectors in the case $n \leq N$ and up to the N eigenvectors in the case $n > N$. Further, using the eigenvalue problem equation (3.2) we can write

$$\mathbf{X}^T\tilde{\mathbf{u}} = \mathbf{X}^T\mathbf{X}\mathbf{u} = n\lambda\mathbf{u}.$$

However, due to the fact that $(n\lambda)^{-2}\tilde{\mathbf{u}}^T\mathbf{X}\mathbf{X}^T\tilde{\mathbf{u}} = (n\lambda)^{-1}$ we need to correctly normalize the $\mathbf{X}^T\tilde{\mathbf{u}}$ vectors to keep the eigenvectors $\mathbf{u}$ orthonormal. This normalization leads to the form

$$\mathbf{u} = (n\lambda)^{-1/2}\mathbf{X}^T\tilde{\mathbf{u}} = \tilde{\lambda}^{-1/2}\mathbf{X}^T\tilde{\mathbf{u}}, \tag{3.5}$$

where $\tilde{\mathbf{u}}, \tilde{\lambda}$ are given by the solution of (3.3). After the extraction of $p \leq (n-1)$ non-zero eigenvectors $\{\tilde{\mathbf{u}}_i\}_{i=1}^p$ and corresponding eigenvalues $\{\tilde{\lambda}_i\}_{i=1}^p$ we can rewrite (3.5) in matrix form

$$\mathbf{U} = \mathbf{X}^T\tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}^{-1/2},$$

where columns of $\tilde{\mathbf{U}}$ are created by the eigenvectors $\{\tilde{\mathbf{u}}_i\}_{i=1}^p$ and $\tilde{\mathbf{\Lambda}}$ is a diagonal matrix $diag(\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_p)$. The projection of the original data $\{\mathbf{x}_i\}_{i=1}^n$ onto the desired principal directions is now given by

$$\mathbf{P} = \mathbf{X}\mathbf{X}^T\tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}^{-1/2} = \mathbf{K}\tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}^{-1/2} = \tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}^{1/2}, \tag{3.6}$$

where the last equality follows from equation (3.3).

3.2 Nonlinear, Kernel-based PCA

Consider now that after observing the data $\{\mathbf{x}_i\}_{i=1}^n \in \mathcal{X} \subseteq \mathcal{R}^N$ we apply a nonlinear mapping (2.20); i.e mapping of the following form

$$\Phi : \mathcal{X} \rightarrow \mathcal{F}, \quad \mathbf{x} \rightarrow \Phi(\mathbf{x})$$

where $\mathcal{F}$ is an $M \leq \infty$ dimensional feature space corresponding to some kernel function $K(\mathbf{x}, \mathbf{y})$. The (linear) PCA problem in $\mathcal{F}$ can be formulated as the diagonalization of an n-sample estimate of the $(M \times M)$ covariance matrix

$$\hat{\mathbf{C}}_{\mathcal{F}} = \frac{1}{n} \sum_{i=1}^{n} \Phi(\mathbf{x}_i)\Phi(\mathbf{x}_i)^T = \frac{1}{n} \mathbf{\Phi}^T \mathbf{\Phi}, \tag{3.7}$$

where $\Phi(\mathbf{x}_i)$ are centered nonlinear mappings of the input variables $\{\mathbf{x}_i\}_{i=1}^n \in \mathcal{X}$ and the $(n \times M)$ matrix $\mathbf{\Phi}$ is their matrix representation. Using the same derivation as described in the previous section the equivalent eigenvalue problem can be derived

$$\mathbf{K}\tilde{\mathbf{u}} = n\lambda\tilde{\mathbf{u}} = \tilde{\lambda}\tilde{\mathbf{u}}, \tag{3.8}$$

where $\mathbf{K}$[1] now represents the symmetric $(n \times n)$ *Gram matrix* with the elements

$$\mathrm{K}_{ij} = (\Phi(\mathbf{x}_i).\Phi(\mathbf{x}_j)) = K(\mathbf{x}_i, \mathbf{x}_j). \tag{3.9}$$

In fact, using a very similar derivation, this Kernel PCA algorithm was described by Schölkopf et al. [121]. Another kernel-based description of PCA can be easily recovered from the method of *snapshots* derived by Sirovich [126] for a discrete point approximation of the continuous Karhunen-Loève expansion. Note, that using the polynomial kernel $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}.\mathbf{y})$ (Appendix A.3) leads to the same (linear) kernel-based PCA as described in the subsection (3.1.1).

Again, using the derivation from the previous section leading to (3.5), we can express the desired eigenvectors $\mathbf{u}$ of the covariance matrix $\hat{\mathbf{C}}_{\mathcal{F}}$ in a feature space $\mathcal{F}$ as

$$\mathbf{u} = \tilde{\lambda}^{-1/2}\mathbf{\Phi}^T\tilde{\mathbf{u}}$$

or in the matrix form

$$\mathbf{U} = \mathbf{\Phi}^T\tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}^{-1/2}, \tag{3.10}$$

where similar to the previous subsection the columns of $\tilde{\mathbf{U}}$ are created by the extracted eigenvectors $\{\tilde{\mathbf{u}}^i\}_{i=1}^p$ of $\mathbf{K}$ and $\tilde{\mathbf{\Lambda}}$ is a diagonal matrix $diag(\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_p)$ of the corresponding eigenvalues. The k-th nonlinear principal component of $\mathbf{x}$ is now given as the projection of $\Phi(\mathbf{x})$ onto the eigenvector $\mathbf{u}^k$

$$\beta_k(\mathbf{x}) \equiv \Phi(\mathbf{x})^T\mathbf{u}^k = \tilde{\lambda}_k^{-1/2}\Phi(\mathbf{x})^T\mathbf{\Phi}^T\tilde{\mathbf{u}}^k = \tilde{\lambda}_k^{-1/2}\sum_{i=1}^{n}\tilde{u}_i^k K(\mathbf{x}_i, \mathbf{x}). \tag{3.11}$$

Re-writing this projection in matrix form we have, for the projection of original data points $\{\mathbf{x}_i\}_{i=1}^n$,

$$\mathbf{P} = \mathbf{\Phi}\mathbf{U} = \mathbf{\Phi}\mathbf{\Phi}^T\tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}^{-1/2} = \mathbf{K}\tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}^{-1/2} = \tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}^{1/2}. \tag{3.12}$$

In practice we are usually also interested in the projection of test data points $\{\mathbf{x}_i\}_{i=n+1}^{n+n_t}$ which were not used to estimate the eigenvectors and eigenvalues. This can be simply given by

$$\mathbf{P}_t = \mathbf{\Phi}_t\mathbf{\Phi}^T\tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}^{-1/2} = \mathbf{K}_t\tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}^{-1/2}, \tag{3.13}$$

[1] Because we assume centered nonlinear mappings $\{\Phi(\mathbf{x}_i)\}_{i=1}^n$ we need to centralize the matrix $\mathbf{K}$. Centralization in a feature space $\mathcal{F}$ is again given by (3.4) [121].

where $\mathbf{\Phi}_t$ is the $(n_t \times M)$ matrix of the mapped testing data points $\{\Phi(\mathbf{x}_i)\}_{i=n+1}^{n+n_t}$ and $\mathbf{K}_t$ is the $(n_t \times n)$ 'test' matrix whose elements are

$$(\mathrm{K}_t)_{ij} = (\Phi(\mathbf{x}_i).\Phi(\mathbf{x}_j)) = K(\mathbf{x}_i, \mathbf{x}_j),$$

where $\{\mathbf{x}_i\}_{i=n+1}^{n+n_t}$ and $\{\mathbf{x}_j\}_{j=1}^{n}$ are testing and training points, respectively. The centralization of $\mathbf{K}_t$ is given by [172, 121]

$$\mathbf{K}_t \leftarrow (\mathbf{K}_t - \frac{1}{n}\mathbf{1}_{n_t}\mathbf{1}_n^T\mathbf{K})(\mathbf{I} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T), \tag{3.14}$$

where $\mathbf{I}$ is again n dimensional identity matrix and $\mathbf{1}_{n_t}$ represent the vector of ones of the length n_t.

3.2.1 The Estimation of Kernel Eigenfunctions

In this subsection we focus on the estimation of the eigenfunctions forming the expansion of a kernel function $K(\mathbf{x}, \mathbf{y})$. This will provide us with the connection between these estimates and the projections (3.12) onto the eigenvectors estimated by kernel PCA.

Based on Mercer's theorem (subsection 2.4.1) each kernel function $K(\mathbf{x}, \mathbf{y})$ can be expanded into a uniformly convergent series (2.15); i.e.

$$K(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{M} \lambda_i \psi_i(\mathbf{x})\psi_i(\mathbf{y}) \quad M \leq \infty$$

where $\{\psi_i(.)\}_{i=1}^{M}$ are the normalized eigenfunctions of the integral operator

$$\int_{\mathcal{X}} K(\mathbf{y}, \mathbf{x})\psi_i(\mathbf{x})d\mathbf{x} = \lambda_i\psi_i(\mathbf{y}).$$

The normalization condition implies the constraint

$$\int_{\mathcal{X}} \psi_i(\mathbf{x})\psi_j(\mathbf{x})d\mathbf{x} = \delta_{ij},$$

where δ_{ij} is the Kronecker delta function. Assuming we use a simple quadrature rule with weights equal to $1/n$, where n represents the number of equally spaced data points $\{\mathbf{x}_k\}_{k=1}^{n}$, we can write

$$\frac{1}{n}\sum_{k=1}^{n} K(\mathbf{y}, \mathbf{x}_k)\psi_i(\mathbf{x}_k) \approx \lambda_i\psi_i(\mathbf{y}) \tag{3.15}$$

and

$$\frac{1}{n}\sum_{k=1}^{n} \psi_i(\mathbf{x}_k)\psi_j(\mathbf{x}_k) \approx \int_{\mathcal{X}} \psi_i(\mathbf{x})\psi_j(\mathbf{x})d\mathbf{x} = \delta_{ij}.$$

Now, solving (3.15) at the data points $\{\mathbf{x}_k\}_{k=1}^{n}$, i.e. using the Nyström method for the solution of integral equations [18], we obtain the eigenvalue problem (3.8). This simply means that we approximate the kernel function expansion $K(\mathbf{x}, \mathbf{y})$ by

$$K(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{M} \lambda_i\psi_i(\mathbf{x})\psi_i(\mathbf{y}) \approx \sum_{i=1}^{n} n\lambda_i \frac{\psi_i(\mathbf{x})}{\sqrt{n}}\frac{\psi_i(\mathbf{y})}{\sqrt{n}}.$$

The approximation to the pth eigenfunction is then given by $\psi_p(\mathbf{x}_k) \approx \frac{1}{\sqrt{n}}\tilde{u}_k^p$ where $\tilde{u}_k^p$ is the kth element of the pth eigenvector given by (3.8). It is also easy to see that there is the relation $\{\lambda_i \approx \frac{\tilde{\lambda}_i}{n}\}_{i=1}^{n}$ where $\{\tilde{\lambda}_i\}_{i=1}^{n}$ are the eigenvalues of the Gram matrix $\mathbf{K}$.

Thus, we can see that the projections (3.12) are nothing other than the scaled eigenfunction $\{\psi_i(\mathbf{x})\}_{i=1}^{n}$ estimates. This different scaling for the estimate of the ith eigenfunction

$\psi_i(\mathbf{x})$ is given by $\sqrt{\lambda_i}$. This fact was recently pointed out in [164, 165]. Moreover, the authors in [173, 164, 165] studied the more realistic case when the eigenfunctions are estimated based on the observed data distributed according to some density function $p(\mathbf{x})$. They proposed to use the same estimate as described above, however, using the finite samples drawn according to $p(\mathbf{x})$. This effectively leads to the estimate of the modified eigenfunctions $p^{1/2}(\mathbf{x})\psi_i(\mathbf{x})$ due to the fact that we assume the generalized eigenproblem

$$\int_{\mathcal{X}} K(\mathbf{y},\mathbf{x})p(\mathbf{x})\psi_i(\mathbf{x})d\mathbf{x} = \lambda_i\psi_i(\mathbf{y})$$

with the orthonormality constraint

$$\int_{\mathcal{X}} \psi_i(\mathbf{x})p(\mathbf{x})\psi_j(\mathbf{x})d\mathbf{x} = \delta_{ij}.$$

3.3 An EM Approach to Kernel PCA

Although, the diagonalization of the $\mathbf{K}$ matrix in the eigenvalue problem (3.8) (or (3.3)) provides a unique solution to the estimation of eigenvectors and eigenvalues, in the case of a high number of data points the problem is computationally burdensome[2]. Fortunately, in practice, we usually do not need to extract the whole spectrum of eigenvectors and eigenvalues, rather the extraction only of the first leading eigenvalues and eigenvectors is desired. In such a case, several algorithms for the more efficient extraction of a subset of eigenvectors and eigenvalues exist. An example is the *power method* [162, 36] which extracts eigenvectors one after an other, in decreasing order of their corresponding eigenvalues. This method scales as $\mathcal{O}(p^2)$ where p is the number of extracted eigenvectors. Another method, based on a probabilistic formulation of PCA using the *expectation-maximization* (EM) algorithm [19] for the extraction of desired principal components was proposed in [113, 143]. The aim of this section is to provide a description of the modification of the EM approach to PCA in the case of nonlinear (kernel-based) PCA. This algorithm was proposed in [107].

3.3.1 Probabilistic PCA

A probabilistic PCA model in the input space $\mathcal{R}^N$ is defined as the latent variable model

$$\mathbf{x} = \mathbf{Q}\mathbf{y} + \boldsymbol{\eta}, \tag{3.16}$$

where $\mathbf{Q}$ is an $(N \times p)$ parameter matrix and the observation vector and latent variable vectors are given as $\mathbf{x} \in \mathcal{R}^N$ and $\mathbf{y} \in \mathcal{R}^p$, respectively. The latent variables are assumed to be normally distributed with zero mean and identity covariance; i.e. $\mathbf{y} \sim N(\mathbf{0}, \mathbf{I})$. The zero mean noise $\boldsymbol{\eta} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ is also normally distributed with a covariance matrix defined as $\boldsymbol{\Sigma}$. Further, we assume that the latent and noise variables are independent and also that their samples are iid. In such a case model (3.16) reduces to a single Gaussian model for $\mathbf{x}$; i.e. $\mathbf{x} \sim N(\mathbf{0}, \mathbf{Q}\mathbf{Q}^T + \boldsymbol{\Sigma})$. If the $\boldsymbol{\Sigma}$ matrix is restricted to be a diagonal matrix with positive elements the model is known as *factor analysis*. Further, if we assume that $\boldsymbol{\Sigma} = \lim_{\sigma^2 \to 0} \sigma^2\mathbf{I}$ the PCA model is obtained. In this case the noise $\boldsymbol{\eta}$ is assumed to be isotropic (equal in all directions) and infinitesimally small.

Given an observation vector $\mathbf{x}$ we are interested in the posterior probability distribution $P(\mathbf{y}|\mathbf{x})$. Assuming that the noise is distributed as $\boldsymbol{\eta} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ and using Bayes's rule we can write for the posterior density function [143, 114]

$$p(\mathbf{y}|\mathbf{x}) = N((\mathbf{Q}^T\mathbf{Q} + \sigma^2\mathbf{I})^{-1}\mathbf{Q}^T\mathbf{x}, \sigma^2(\mathbf{Q}^T\mathbf{Q} + \sigma^2\mathbf{I})^{-1}).$$

[2] The direct diagonalization of a symmetric $(n \times n)$ matrix scales as $\mathcal{O}(n^3)$.

It is now clear that as the noise level in the model becomes infinitesimal the posterior density becomes a delta function

$$p(\mathbf{y}|\mathbf{x}) = N((\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^T\mathbf{x}, \mathbf{0}) = \delta(\mathbf{y} - (\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^T\mathbf{x})$$

and the EM algorithm is effectively a straightforward least squares projection [113, 114]

$$\begin{array}{ll} \mathbf{E-Step} & \mathbf{Y} = (\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^T\mathbf{X} \\ \mathbf{M-Step} & \mathbf{Q}^{new} = \mathbf{X}\mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T)^{-1}. \end{array}$$

Where, now, we denoted the $(N \times n)$ matrix of data observations as $\mathbf{X}$ and the $(p \times n)$ matrix of latent variables as $\mathbf{Y}$.[3]

It has been shown in [143] that, in the case of infinitesimal small noise in our model, the maximum-likelihood estimate of $\mathbf{Q}$ at convergence will be equal to

$$\mathbf{Q}_{ML} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{R}, \tag{3.17}$$

where the columns of the $\mathbf{U}$ matrix are the eigenvectors of the sample covariance matrix with corresponding eigenvalues $\lambda_1, ..., \lambda_p$ being the diagonal elements of the diagonal matrix $\mathbf{\Lambda}$, and $\mathbf{R}$ is an arbitrary orthogonal rotation matrix. In [143], the authors also pointed out, that taking the columns of $\mathbf{R}^T$ to be equal to the eigenvectors of the $\mathbf{Q}_{ML}^T\mathbf{Q}_{ML}$ matrix, we can recover the true principal axes.

3.3.2 EM approach to Kernel PCA

Motivated by the previously described probabilistic PCA results, in [107] we proposed an EM approach to Kernel PCA which, similar to section 3.2, is based on the nonlinear mapping of the input data to feature space $\mathcal{F}$ by a map $\Phi : \mathcal{X} \subseteq \mathcal{R}^N \to \mathcal{F}$.

Realizing that the $\mathbf{Q}$ matrix may by obtained by scaling and rotation of the $\mathbf{U}$ matrix (3.17) consisting of eigenvectors computed by diagonalization of the sample covariance matrix we can express the r^{th} column of $\mathbf{Q}$ as $\mathbf{Q}^r = \sum_{j=1}^n \gamma_j^r \Phi(\mathbf{x}_j)$ and write it in matrix notation as $\mathbf{\Phi}^T\mathbf{\Gamma}$, where the matrix $\mathbf{\Phi}$ is the $(n \times M)$ matrix which has individual rows consisting of the vectors $\Phi(\mathbf{x}_1), \ldots, \Phi(\mathbf{x}_n)$ of the centered mappings[4] of the observed data and $\mathbf{\Gamma}$ is an $(n \times p)$ matrix of the coefficients $\{\gamma_i^r : i = 1, \ldots, n; r = 1, \ldots, p\}$. Using the 'kernel' trick, i.e. $\Phi(\mathbf{x}_1)^T\Phi(\mathbf{x}_2) = K(\mathbf{x}_1, \mathbf{x}_2)$ we can see that the E-step will now be

$$\mathbf{Y} = (\mathbf{\Gamma}^T\mathbf{K}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{K}. \tag{3.18}$$

Now let us consider the M-Step. Denote the term $\mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T)^{-1}$ by $\mathbf{A}$. Then we may write

$$\mathbf{Q}^{new} = \mathbf{\Phi}^T\mathbf{A},$$

where $\mathbf{Q}^{new} = \mathbf{\Phi}^T\mathbf{\Gamma}^{new}$. Thus we have the M-step

$$\mathbf{\Gamma}^{new} = \mathbf{A} = \mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T)^{-1}. \tag{3.19}$$

This choice of $\mathbf{\Gamma}^{new}$ is unique for the case when the $\mathbf{\Phi}^T$ matrix has $rank(\mathbf{\Phi}^T) = n$, otherwise it is one of the possible solutions for $\mathbf{\Phi}^T\mathbf{\Gamma}^{new} = \mathbf{\Phi}^T\mathbf{A}$. Finally, after convergence of the proposed kernel-based EM algorithm, the projection of the new point $\mathbf{x}$ onto the corresponding p nonlinear principal components is given by

$$\boldsymbol{\beta}(\mathbf{x}) \equiv (\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^T\Phi(\mathbf{x}) = (\mathbf{\Gamma}^T\mathbf{K}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{k}, \tag{3.20}$$

[3] To be consistent with the probabilistic PCA method as described in [113, 114, 143], here, we assume this data structure, in spite of the fact that in the thesis we usually assume data samples creating the rows and variables the columns.

[4] Again, the centering in the feature space $\mathcal{F}$ can be carried out in a straightforward manner by 'centering' the kernel matrix $\mathbf{K}$ outlined in the previous sections (eqs. (3.4) and (3.14)).

where $\mathbf{k}$ is the vector $[K(\mathbf{x}_1, \mathbf{x}), ..., K(\mathbf{x}_n, \mathbf{x})]^T$. This projection is up to the scaling and rotation identical to the projection of the data point $\mathbf{x}$ using the eigenvectors of the covariance matrix $\hat{\mathbf{C}}_{\mathcal{F}}$ given by (3.11). In the next chapter these projections are used as input data to rotationally and scaling invariant ordinary least squares regression method and in such a case we even do not need to find true principal axes as given by Kernel PCA algorithm.

We already pointed out that the maximum likelihood estimate $\mathbf{Q}_{ML}$ at convergence will be of the form (3.17). Using this theoretical result, relation (3.10) and the fact that we defined $\mathbf{Q} = \mathbf{\Phi}^T \mathbf{\Gamma}$ we can write

$$\mathbf{\Phi}^T \mathbf{\Gamma}_{ML} = \mathbf{U} \mathbf{\Lambda}^{1/2} \mathbf{R} = \mathbf{\Phi}^T \tilde{\mathbf{U}} \tilde{\mathbf{\Lambda}}^{-1/2} \mathbf{\Lambda}^{1/2} \mathbf{R} = \mathbf{\Phi}^T \tilde{\mathbf{U}} \mathbf{I}_n^{-1/2} \mathbf{R},$$

where $\mathbf{I}_n$ is the diagonal $(p \times p)$ matrix with the elements on the diagonal equal to n and $\mathbf{\Gamma}_{ML}$ denotes the matrix $\mathbf{\Gamma}$ corresponding to maximum likelihood estimate $\mathbf{Q}_{ML}$. Thus, at convergence, the orthogonality of the $\mathbf{\Gamma}_{ML} = \tilde{\mathbf{U}} \mathbf{I}_n^{-1/2} \mathbf{R}$ matrix will be achieved which may be seen from the fact that

$$\mathbf{\Gamma}_{ML}^T \mathbf{\Gamma}_{ML} = \mathbf{R}^T \mathbf{I}_n^{-1/2} \tilde{\mathbf{U}}^T \tilde{\mathbf{U}} \mathbf{I}_n^{-1/2} \mathbf{R} = \mathbf{I}_n^{-1}.$$

Further, it is easy to see that $\mathbf{Q}_{ML}^T \mathbf{Q}_{ML} = \mathbf{\Gamma}_{ML}^T \mathbf{K} \mathbf{\Gamma}_{ML} = \mathbf{R}^T \mathbf{\Lambda} \mathbf{R}$ and we may write the projection of the training data points

$$\begin{aligned} \mathbf{P} &= \left\{ (\mathbf{Q}_{ML}^T \mathbf{Q}_{ML})^{-1} \mathbf{Q}_{ML}^T \mathbf{K} \right\}^T = \left\{ (\mathbf{R}^T \mathbf{\Lambda} \mathbf{R})^{-1} \mathbf{R}^T \mathbf{I}_n^{-1/2} \tilde{\mathbf{U}}^T \mathbf{K} \right\}^T = \\ &= \left\{ \mathbf{R}^T \mathbf{\Lambda}^{-1} \mathbf{I}_n^{-1/2} \tilde{\mathbf{U}}^T \mathbf{K} \right\}^T = \left\{ \mathbf{R}^T \mathbf{\Lambda}^{-1} \mathbf{I}_n^{-1/2} \tilde{\mathbf{\Lambda}} \tilde{\mathbf{U}}^T \right\}^T = \tilde{\mathbf{U}} \mathbf{I}_n^{1/2} \mathbf{R} = \mathbf{\Gamma}_{ML} \mathbf{I}_n. \end{aligned}$$

Similarly for the projection of testing data points we may write

$$\mathbf{P}_t = \mathbf{K}_t \tilde{\mathbf{U}} \mathbf{I}_n^{-1/2} \mathbf{\Lambda}^{-1} \mathbf{R} = \mathbf{K}_t \mathbf{\Gamma}_{ML} \mathbf{\Lambda}^{-1} \mathbf{R}$$

and it is easy to see that these projections are up to the scaling $\mathbf{\Lambda}^{1/2}$ and rotation $\mathbf{R}^T$ identical to the projections (3.12) and (3.13), respectively. In the next subsection we will describe how the estimation of the eigenvalues $\{\lambda_i\}_{i=1}^p$ and consequently the normalization of the projection to avoid the different scaling in the individual eigendirections can be achieved. However, before that, we would like to make several notes about the proposed method for performing kernel PCA.

Firstly, due to the use of Mercer kernels the method is independent of the dimensionality of the input space. Secondly, the computational complexity, per iteration, of the proposed EM method for Kernel PCA is $\mathcal{O}(pn^2)$ where n is the number of data points and p is the number of extracted components. Where a small number of eigenvectors require to be extracted and there are a large number of data points available this method is comparable in complexity to the iterative power method which has complexity $\mathcal{O}(n^2)$. As we noted before, direct diagonalization of a symmetric $\mathbf{K}$ matrix to solve the eigenvalue problem (3.8) has complexity of the order $\mathcal{O}(n^3)$. In [111, 107] we compared these three methods in terms of the number of floating point operations and we observed that the proposed EM approach to Kernel PCA may also be profitable when the extraction of a higher number of eigenvectors is needed. Moreover, in [111], on several examples the convergence of the eigenvectors extracted by the proposed approach to the eigenvectors obtained by the solution of the 'classical' kernel PCA (section 3.2) was demonstrated.

From the equations (3.18) and (3.19) we can also see that individual EM steps can be performed without storing the whole $(n \times n)$ matrix $\mathbf{K}$. In such a case memory requirements scale as $\mathcal{O}(p^2) + \mathcal{O}((p+1)n)$. However, this will slow down the computations as the elements of $\mathbf{K}$ have to be computed repeatedly. We discuss possible implementations of the algorithm in Appendix A.1.

3.3.3 Eigenvalues Estimation

From the definition of our probabilistic PCA model it is clear that the latent variables $\mathbf{y}$ have identity covariance matrix. Thus at convergence the projection of the observed data to the p-dimensional subspace will lead to the sphering of the projected data. Using the fact that at convergence $\mathbf{\Gamma}_{ML} = \tilde{\mathbf{U}}\mathbf{I}_n^{-1/2}\mathbf{R}$ is an orthogonal matrix, we will show how the matrix $\mathbf{\Lambda}$ can be recovered.

The diagonalization of the symmetric matrix $\mathbf{K}^2$ instead of $\mathbf{K}$ leads to the same eigenvectors and squared eigenvalues [78, 119]. Further the matrix $\frac{1}{n}\mathbf{K}^2$ can be seen as the sample estimate of the covariance matrix of *the empirical kernel map* Φ_{emp} defined for a given set of points $\{\mathbf{x}_i\}_{i=1}^n$ as

$$\Phi_{emp} : \mathcal{R}^N \rightarrow \mathcal{R}^n$$

$$\mathbf{x} \rightarrow K(.,\mathbf{x})|_{\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}} = (K(\mathbf{x}_1,\mathbf{x}),\ldots,K(\mathbf{x}_n,\mathbf{x})).$$

This fact was recently also used in [79] where a similar EM algorithm to Kernel PCA was proposed. It is easy to see that applying the defined Φ_{emp} mapping on all data points will lead to the construction of the Gram matrix $\mathbf{K}$. However, this is now supposed to be a data matrix with the n observations in rows and n variables in columns. Further note, that the centralization procedure (3.4) provides the matrix with zero-mean rows and columns [172]. Thus, we can formulate the eigenvalue problem

$$\frac{1}{n}\mathbf{K}^T\mathbf{K}\tilde{\mathbf{u}} = \frac{1}{n}\mathbf{K}^2\tilde{\mathbf{u}} = \lambda^2\tilde{\mathbf{u}},$$

where the centralized $\mathbf{K}$ matrix is used and $\tilde{\mathbf{u}}, \tilde{\lambda} = n\lambda$ are also the solutions of (3.8).

In the next step we can take the orthonormal basis created by orthogonalization of the columns of $\mathbf{\Gamma}$ and project the observed data to the p-dimensional subspace defined by this orthonormal matrix $\mathbf{\Gamma}_{orth}$. By applying standard PCA on the covariance matrix of the projected data $\mathbf{Y} = \mathbf{K}\mathbf{\Gamma}_{orth}$ we can recover the desired squared eigenvalues of the covariance matrix $\hat{\mathbf{C}}_{\mathcal{F}}$ (3.7).

4. KERNEL-BASED REGRESSION

4.1 *Introduction*

In this chapter the construction of the estimates of desired functional dependencies g (2.1) in a RKHS $\mathcal{H}$ will be described. Although different regression models will be constructed the same variational problem of finding the estimate $f(\mathbf{x}) \in \mathcal{H}$ (2.2) minimizing the functional (2.21) will be assumed. We have already mentioned that, based on Representer Theorem (subsection 2.4.2), the solution to (2.21) is of the form

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i K(\mathbf{x}_i, \mathbf{x}) = \sum_{i=1}^{n} c_i \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}), \tag{4.1}$$

where $\Phi : \mathcal{X} \subseteq \mathcal{R}^N \to \mathcal{F}$ again represents a mapping to $M \leq \infty$ dimensional feature space $\mathcal{F}$ given by the selected kernel function. We also consider that $\Phi_i(\mathbf{x}) = \sqrt{\alpha_i}\phi_i(\mathbf{x}),\ \mathbf{x} \in \mathcal{X}$ where $\{\phi_i(.)\}_{i=1}^M$ is a set of linearly independent functions (not necessary orthogonal) creating the basis of $\mathcal{H}$. Thus, any $f \in \mathcal{H}$ can be expanded into the form

$$f(\mathbf{x}) = (\sum_{i=1}^{n} c_i \Phi(\mathbf{x}_i))^T \Phi(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x}) = \sum_{i=1}^{M} w_i \sqrt{\alpha_i}\phi_i(\mathbf{x}) = \sum_{i=1}^{M} b_i \phi_i(\mathbf{x}) \tag{4.2}$$

and we can see that (4.1) can also be interpreted as an estimate of a linear regression model[1] in $\mathcal{F}$ where the M dimensional vector $\mathbf{w} = \sum_{i=1}^n c_i \Phi(\mathbf{x}_i)$ now represents a vector of regression coefficients. Using the norm definition (2.18) we can write

$$\|f\|_{\mathcal{H}}^2 = \sum_{i=1}^{M} \frac{b_i^2}{\alpha_i} = \sum_{i=1}^{M} \frac{(w_i\sqrt{\alpha_i})^2}{\alpha_i} = \sum_{i=1}^{M} w_i^2 = \|\mathbf{w}\|^2.$$

The squared norm $\|f\|_{\mathcal{H}}^2$ actually stands for a large class of smoothness functionals $\Omega(f)$ defined over elements of the form (4.2). It simply means, that the smoothing properties of a final estimate are determined by the appropriate choice of a kernel function $K(\mathbf{x}, \mathbf{y})$ which induces a RKHS and a corresponding norm $\|f\|_{\mathcal{H}}$. We demonstrate this on an example of translation invariant kernels $K(\mathbf{x} - \mathbf{y})$ in Appendix A.4.

The concept of regularization also provided the connection [33] between Regularization Networks and Support Vector Regression (section 4.5) where the minimization of the risk functional

$$\frac{1}{n}\sum_{i=1}^{n} V(y_i, f(\mathbf{x}_i)) + \xi\|\mathbf{w}\|^2 \tag{4.3}$$

is assumed. However, the idea of a '*flat*' linear regression estimate; i.e. the penalization of large in absolute value weights $\{w_i\}_{i=1}^M$ through the regularization term ξ, was rather motivated by the aim of finding a separating hyperplane of maximum distance between classes used in the pattern recognition domain [153]. It has been stressed later that choosing the *flattest* function in a feature space can, based on the smoothing properties of the selected kernel function, lead to a smooth function in the input space [129].

[1] The expansion (4.1) is usually called the *dual* representation of f whilst (4.2) is called the *primal* representation of f.

To summarize this section, we can see that the solution of (4.1) given by the Representer Theorem can be also formulated in its primal representation (4.2). This simply means that having the finite set of data samples $\{(\mathbf{x}_i, y_i)_{i=1}^n \in \mathcal{X} \times \mathcal{Y}\}$ we approximate the desired functional dependency g (2.1) by the estimate given by the solution of a linear regression model

$$\mathbf{y} = \mathbf{\Phi}\mathbf{w} + \boldsymbol{\eta}, \tag{4.4}$$

where we assume $\boldsymbol{\eta}$ is a $(n \times 1)$ vector of error terms whose elements have equal variance σ^2 and are independent of each other, and $\mathbf{\Phi} = [\Phi(\mathbf{x}_1), \Phi(\mathbf{x}_2), \ldots, \Phi(\mathbf{x}_n)]^T$ represents the $(n \times M)$ matrix of predictors (regressors) in $\mathcal{F}$. Assuming that the appropriate kernel mapping was selected, we need to guarantee the smoothness of the final estimate by choosing a $flat$ regression function in $\mathcal{F}$. In the case of least-squares estimates this will bring us to the problem of multicollinearity.

4.1.1 Multicollinearity and Regularized Least-Squares Regression Models

One of the main problems in multiple regression is a linear or near-linear dependence of the regressors – *multicollinearity.* The multicollinearity of regressors is a serious problem that can dramatically influence the usefulness of a regression model. Multicollinearity results in large variances and covariances for the least-squares estimators of the regression coefficients. Multicollinearity can also produce estimates of the regression coefficients that are too large in absolute values (Appendix A.2). Thus the values and signs of estimated regression coefficients may change considerably given different data samples. This effect can lead to a regression model which fits the training data reasonably well, but in general bad generalization of the model can occur. This fact is in a very close relation to the requirement of choosing the *flattest* function in a feature space stressed in the previous section. In fact, in a kernel-type of regression we usually nonlinearly transform the original data to the high dimensional space whose dimension M is in many cases significantly higher than the number of observations; i.e. $M \gg n$. In such a case, there are many linear as well as possibly many approximate dependencies among the regressors. This can be easily seen from the fact that in the model (4.4), the $rank(\mathbf{\Phi}) \leq n$. Further, in the case of polynomial kernels of the type $K(\mathbf{x}, \mathbf{y}) = ((\mathbf{x}.\mathbf{y}) + c)^d$ (Appendix A.3) the existing input space multicollinearities will be exactly 'mapped' into a feature space representation.

There exist several methods to deal with multicollinearity. Generally they are based on the requirement to shrink the solution to the regression from the areas of lower data spread. From a statistical point of view this leads to a biased but lower variance estimate of the regression coefficients. These methods create a class of regression techniques usually called *shrinkage* or *regularized* regression. We may distinguish two main principles in their construction. The first principle is based on the transformation of the original regressors into latent variables (LV). LV are usually created to be orthogonal with the aim of reflecting the 'real' intrinsic structure of the original regressors. Principal Component Regression (PCR) and Partial Least Squares (PLS) regression are the main techniques belonging to this category. In contrast to LV techniques, Ridge Regression (RR) operates on the original regressors and the desired lower variance estimate is achieved by penalizing the weights with the aim of shrinking the solution to the origin. Some other techniques belonging either to the first or second category are discussed in [60, 131, 140, 37].

In the thesis we discuss RR, PCR and PLS Regression approaches. Using the theoretical basis of these techniques in input space, we discuss their counterparts in a kernel defined feature space. In the next three sections we provide a brief description of these methods followed by their kernel based implementation.

4.2 Kernel Principal Component Regression

PCR is based on the projection of the original regressors onto the principal components extracted by PCA. As there is a straightforward connection between PCR in input space and its kernel-based implementation [108, 110] we directly start with a feature space linear regression model (4.4) and further assume that regressors $\{\Phi_j(\mathbf{x})\}_{j=1}^M$ are zero-mean. Thus $\mathbf{\Phi}^T\mathbf{\Phi}$ is proportional to the sample covariance matrix and Kernel PCA can be performed to extract its M eigenvalues $\{\tilde{\lambda}_i\}_{i=1}^M$ and corresponding eigenvectors $\{\mathbf{u}^i\}_{i=1}^M$[2] (3.10). Having the eigensystem $\{\tilde{\lambda}_i, \mathbf{u}^i\}_{i=1}^M$ the spectral decomposition [60] of $\mathbf{\Phi}^T\mathbf{\Phi}$ has the form

$$\mathbf{\Phi}^T\mathbf{\Phi} = \sum_{i=1}^M \tilde{\lambda}_i \mathbf{u}^i(\mathbf{u}^i)^T. \tag{4.5}$$

The k-th principal component of $\Phi(\mathbf{x})$ is given by (3.11). By projection of all original regressors onto the principal components we can rewrite (4.4) as

$$\mathbf{y} = \mathbf{B}\mathbf{v} + \boldsymbol{\eta}, \tag{4.6}$$

where $\mathbf{B} = \mathbf{\Phi}\mathbf{U}$ is now an $(n \times M)$ matrix of transformed regressors and $\mathbf{U}$ is an $(M \times M)$ matrix whose k-th column is the eigenvector $\mathbf{u}^k$. The columns of the matrix $\mathbf{B}$ are now orthogonal and the least squares estimate of the coefficients $\mathbf{v}$ becomes

$$\hat{\mathbf{v}} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{y} = \tilde{\mathbf{\Lambda}}^{-1}\mathbf{B}^T\mathbf{y}, \tag{4.7}$$

where $\tilde{\mathbf{\Lambda}} = \mathrm{diag}(\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_M)$. It is worth noting, that PCR, as well as other biased regression techniques, is not invariant to the relative scaling of the original regressors [26]. However, similar to ordinary least squares (OLS) regression, the solution of (4.6) does not depend on a possibly different scaling in individual eigendirections used in the Kernel PCA transformation. Further, the results obtained using all principal components—the PCA projection of the original regressor variables—in (4.6) is equivalent to that obtained by least squares using the original regressors. In fact we can express the estimate $\hat{\mathbf{w}}$ of the original model (4.4) as

$$\hat{\mathbf{w}} = \mathbf{U}\hat{\mathbf{v}} = \mathbf{U}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{y} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{y} = \sum_{i=1}^M \tilde{\lambda}_i^{-1}\mathbf{u}^i(\mathbf{u}^i)^T\mathbf{\Phi}^T\mathbf{y}$$

and its corresponding variance-covariance matrix [60] as

$$cov(\hat{\mathbf{w}}) = \sigma^2\mathbf{U}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{U}^T = \sigma^2\mathbf{U}\tilde{\mathbf{\Lambda}}^{-1}\mathbf{U}^T = \sigma^2\sum_{i=1}^M \tilde{\lambda}_i^{-1}\mathbf{u}^i(\mathbf{u}^i)^T. \tag{4.8}$$

where we used the fact that $\mathbf{y} \sim N(\mathbf{\Phi}\mathbf{w}, \sigma^2\mathbf{I})$. To avoid the problem of multicollinearity, PCR uses only some of the principal components. It is clear from (4.8) that the influence of small eigenvalues can significantly increase the overall variance of the estimate. PCR simply deletes the principal components corresponding to small values of the eigenvalues $\tilde{\lambda}_i$. The penalty we have to pay for the decrease in variance of the regression coefficient estimate is bias in the final estimate. However, if multicollinearity is a serious problem, the introduced bias can have a less significant effect in comparison to a high variance estimate. If the elements of $\mathbf{v}$ corresponding to deleted regressors are zero, an unbiased estimate is achieved [60].

Using the first p nonlinear principal components (3.11) to create a linear model based on orthogonal regressors in feature space $\mathcal{F}$ we can formulate the Kernel PCR model [108, 110] in primal form

$$f(\mathbf{x}) = \sum_{k=1}^p v_k\beta_k(\mathbf{x}) + b \tag{4.9}$$

[2] For the moment, we are theoretically assuming that $n > M$. Otherwise we have to deal with a singular case ($n \leq M$) allowing us to extract only up to $n-1$ eigenvectors corresponding to non-zero eigenvalues.

or dual representation

$$f(\mathbf{x}) = \sum_{i=1}^{n} a_i K(\mathbf{x}_i, \mathbf{x}) + b, \tag{4.10}$$

where $\{a_i = \sum_{k=1}^{p} v_k \tilde{\lambda}_k^{-1/2} \tilde{u}_i^k\}_{i=1}^n$ and b is a bias term.

We have shown that by removing the principal components whose variances are very small we can eliminate large variances of the estimate due to multicollinearities. However, if the orthogonal regressors corresponding to those principal components have a large correlation with the dependent variable y such deletion is undesirable. On the data sets employed in this thesis we experimentally demonstrated this fact in [111] and in section 4.7 we will provide an example of these observations. There are several different strategies for selecting the appropriate orthogonal regressors for the final model (see [60, 59] and ref. therein). In section 4.7 we discuss approaches used in our experiments.

4.3 Partial Least Squares Regression

The PLS method [168, 170] was proposed and maintained a popular status as a regression technique in its domain of origin – Chemometrics. Because the technique is not as widely known as PCR or RR we firstly provide the description of linear PLS and in the next subsection we derive its nonlinear, kernel-based variant which we proposed in [112].

PLS regression is a technique for modeling a linear relationship between a set of output variables (responses)[3] $\{\mathbf{y}_i\}_{i=1}^n \in \mathcal{R}^L$ and a set of input variables (regressors) $\{\mathbf{x}_i\}_{i=1}^n \in \mathcal{X} \subseteq \mathcal{R}^N$. In the first step, PLS creates orthogonal, i.e. uncorrelated, latent variables which are linear combinations of the original regressors whilst also utilizing existing correlations among input and output variables. A least squares regression is then performed on the subset of extracted latent variables. This leads to biased but lower variance estimates of the regression coefficients compared to the OLS regression.

In the following $\mathbf{X}$ will represent the $(n \times N)$ matrix of n inputs and $\mathbf{Y}$ will stand for the $(n \times L)$ matrix of corresponding L dimensional responses. Further we assume centered input and output variables; i.e. the columns of $\mathbf{X}$ and $\mathbf{Y}$ are zero mean.

There exists several different modifications (see e.g [75, 73, 47, 17]) of the basic algorithm for PLS regression originally developed in [168]. In its basic form the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm [167] is used to sequentially extract the latent vectors $\mathbf{t}, \mathbf{u}$ and weight vectors $\mathbf{w}, \mathbf{c}$ from the $\mathbf{X}$ and $\mathbf{Y}$ matrices in decreasing order of their corresponding singular values. What follows is a modification of the PLS algorithm as described by [69][4]

1. randomly initialize $\mathbf{u}$
2. $\mathbf{w} = \mathbf{X}^T\mathbf{u}$
3. $\mathbf{t} = \mathbf{X}\mathbf{w}$, $\mathbf{t} \leftarrow \mathbf{t}/\|\mathbf{t}\|$
4. $\mathbf{c} = \mathbf{Y}^T\mathbf{t}$
5. $\mathbf{u} = \mathbf{Y}\mathbf{c}$, $\mathbf{u} \leftarrow \mathbf{u}/\|\mathbf{u}\|$
6. repeat steps 2. – 5. until convergence
7. deflate $\mathbf{X}$,$\mathbf{Y}$ matrices: $\mathbf{X} \leftarrow \mathbf{X} - \mathbf{t}\mathbf{t}^T\mathbf{X}$, $\mathbf{Y} \leftarrow \mathbf{Y} - \mathbf{t}\mathbf{t}^T\mathbf{Y}$

[3] Because, it may be also profitable to use PLS regression in situations when the modeling of several mutually dependent responses is desired, we assume a more general multivariate version of PLS; i.e. $L > 1$.

[4] In comparison to the classical NIPALS algorithm we normalize the latent (scores) vectors $\mathbf{t}, \mathbf{u}$ rather than vectors of weights $\mathbf{c}, \mathbf{w}$. This can be efficient only when the number of observation is greater than the number of variables, however, as we will see later it is important for the Kernel PLS algorithm described below.

The PLS regression is an iterative process, i.e. after extraction of one component the algorithm starts again using the deflated matrices $\mathbf{X}$ and $\mathbf{Y}$ computed in step 7. Thus we can achieve a sequence of models up to the point when the rank of $\mathbf{X}$ is reached. However, in practice the technique of cross-validation is usually used to avoid underfitting or overfitting caused by the use of too small or too large dimensional models. After the extraction of the p components we can create $(n \times p)$ matrices $\mathbf{T}$ and $\mathbf{U}$, $(N \times p)$ matrix $\mathbf{W}$ and $(L \times p)$ matrix $\mathbf{C}$ consisting of the columns created by the vectors $\{\mathbf{t}_i\}_{i=1}^p$, $\{\mathbf{u}_i\}_{i=1}^p$, $\{\mathbf{w}_i\}_{i=1}^p$ and $\{\mathbf{c}_i\}_{i=1}^p$, respectively, extracted during the individual iterations. The estimated matrix of regression coefficients $\mathbf{B}$ will take the form [103]

$$\mathbf{B} = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{C}^T, \tag{4.11}$$

where $\mathbf{P}$ is the $(N \times p)$ matrix consisting of loadings vectors $\{\mathbf{p}_i = \mathbf{X}^T\mathbf{t}_i\}_{i=1}^p$. Due to the fact that $\mathbf{p}_i^T\mathbf{w}_j = 0$ for $i > j$ and in general $\mathbf{p}_i^T\mathbf{w}_j \neq 0$ for $i < j$ the matrix $\mathbf{P}^T\mathbf{W}$ is upper triangular and thus invertible. Moreover, using the fact that $\mathbf{t}_i^T\mathbf{t}_j = 0$ for $i \neq j$ and $\mathbf{t}_i^T\mathbf{u}_j = 0$ for $j > i$ in [103] the following equalities were proved[5]

$$\mathbf{W} = \mathbf{X}^T\mathbf{U} \tag{4.12}$$

$$\mathbf{P} = \mathbf{X}^T\mathbf{T}(\mathbf{T}^T\mathbf{T})^{-1} \tag{4.13}$$

$$\mathbf{C} = \mathbf{Y}^T\mathbf{T}(\mathbf{T}^T\mathbf{T})^{-1}. \tag{4.14}$$

Substituting (4.12–4.14) into (4.11) and using the orthogonality of the $\mathbf{T}$ matrix columns we can write the matrix $\mathbf{B}$ in the following form

$$\mathbf{B} = \mathbf{X}^T\mathbf{U}(\mathbf{T}^T\mathbf{X}\mathbf{X}^T\mathbf{U})^{-1}\mathbf{T}^T\mathbf{Y}. \tag{4.15}$$

4.3.1 Kernel Partial Least Squares Regression

Again, assume the non-linear transformation of the input variables $\{\mathbf{x}_i\}_{i=1}^n$ into a feature space $\mathcal{F}$; i.e. mapping $\Phi : \mathbf{x}_i \in \mathcal{R}^N \rightarrow \Phi(\mathbf{x}_i) \in \mathcal{F}$. Our goal is to construct a linear PLS regression model in $\mathcal{F}$. Effectively it means that we can obtain a non-linear regression model where the form of nonlinearity is given by $\Phi(.)$. As we already noted, depending on the nonlinear transformation $\Phi(.)$ the feature space can be high-dimensional, even infinite dimensional when the Gaussian kernel function is used. However, in practice, we are working only with n observations and we have to restrict ourselves to finding the solution of the linear regression problem in the span of the points $\{\Phi(\mathbf{x}_i)\}_{i=1}^n$. This situation is analogous to the case when the input data matrix $\mathbf{X}$ has more columns than rows, i.e. we are dealing with more variables than measured objects. This motivated several authors to introduce the (input space) linear Kernel PLS algorithm [103] to speed up the computation of the components for a linear PLS model. The idea is to compute the components from the $(n \times n)$ $\mathbf{X}\mathbf{X}^T$ matrix rather than $(N \times N)$ $\mathbf{X}^T\mathbf{X}$ matrix when $n \ll N$. Note that the same approach was also used for the computation of the principal components in the previous chapter.

Now, motivated by the theory of RKHS described in section 2.4 we derive the algorithm for the (non-linear) Kernel PLS model. From the previous section we can see that by the connection of steps 2 and 3 and by using the $(n \times M)$ matrix $\mathbf{\Phi}$ of mapped input data we can modify the PLS algorithm into the form[6]

1. randomly initialize $\mathbf{u}$
2. $\mathbf{t} = \mathbf{\Phi}\mathbf{\Phi}^T\mathbf{u}$, $\mathbf{t} \leftarrow \mathbf{t}/\|\mathbf{t}\|$
3. $\mathbf{c} = \mathbf{Y}^T\mathbf{t}$

[5] In our case $\mathbf{T}^T\mathbf{T}$ is p dimensional identity matrix. This is simply a consequence of the normalization of individual latent vectors $\{\mathbf{t}_i\}_{i=1}^p$.

[6] This kernel form of a linear PLS algorithm was described in [69].

4. $\mathbf{u} = \mathbf{Yc}$, $\mathbf{u} \leftarrow \mathbf{u}/\|\mathbf{u}\|$

5. repeat steps 2. – 5. until convergence

6. deflate $\mathbf{\Phi\Phi}^T$,$\mathbf{Y}$ matrices: $\mathbf{\Phi\Phi}^T \leftarrow (\mathbf{\Phi} - \mathbf{tt}^T\mathbf{\Phi})(\mathbf{\Phi} - \mathbf{tt}^T\mathbf{\Phi})^T$,
$\mathbf{Y} \leftarrow \mathbf{Y} - \mathbf{tt}^T\mathbf{Y}$

Applying the same 'kernel' trick (3.9) as in section 3.2 we can write $\mathbf{\Phi\Phi}^T = \mathbf{K}$. Thus, instead of an explicit nonlinear mapping, the kernel function can be used. The deflation of the Gram matrix $\mathbf{K}$ in step 6 after extraction of the $\mathbf{t}$ component is now given by

$$\mathbf{K} \leftarrow (\mathbf{I} - \mathbf{tt}^T)\mathbf{K}(\mathbf{I} - \mathbf{tt}^T) = \mathbf{K} - \mathbf{tt}^T\mathbf{K} - \mathbf{Ktt}^T + \mathbf{tt}^T\mathbf{Ktt}^T, \tag{4.16}$$

where $\mathbf{I}$ is an n dimensional identity matrix. We would like to point out that a similar Kernel PLS algorithm can be also derived from the approach described in [103] which leads to the extraction of the $\mathbf{t}, \mathbf{u}$ components from the $\mathbf{KYY}^T$ and $\mathbf{YY}^T$ matrices. This approach can be more fruitful when the multivariate Kernel PLS model is considered.

Similarly we can see that the matrix of the regression coefficients $\mathbf{B}$ (4.15) will have the form

$$\mathbf{B} = \mathbf{\Phi}^T\mathbf{U}(\mathbf{T}^T\mathbf{KU})^{-1}\mathbf{T}^T\mathbf{Y} \tag{4.17}$$

and to make prediction on training data we can write

$$\hat{\mathbf{Y}} = \mathbf{\Phi B} = \mathbf{KU}(\mathbf{T}^T\mathbf{KU})^{-1}\mathbf{T}^T\mathbf{Y} = \mathbf{TT}^T\mathbf{Y}, \tag{4.18}$$

where the last equality follows from the fact that the matrix of the components $\mathbf{T}$ may be expressed as $\mathbf{T} = \mathbf{\Phi R}$ where $\mathbf{R} = \mathbf{\Phi}^T\mathbf{U}(\mathbf{T}^T\mathbf{KU})^{-1}$ [17, 47]. It is important to stress that during the iterative process of the estimation of the components $\{\mathbf{t}_i\}_{i=1}^p$ we made the deflation of the $\mathbf{K}$ matrix after the extraction of each new component $\mathbf{t}$. Effectively it means that $\mathbf{T} \neq \mathbf{KU}$. Thus, for predictions made on testing points $\{\mathbf{x}_i\}_{i=n+1}^{n+n_t}$ the matrix of regression coefficients (4.17) have to be used; i.e.

$$\hat{\mathbf{Y}}_t = \mathbf{\Phi}_t\mathbf{B} = \mathbf{K}_t\mathbf{U}(\mathbf{T}^T\mathbf{KU})^{-1}\mathbf{T}^T\mathbf{Y},$$

where $\mathbf{\Phi}_t$ is the matrix of the mapped test points and consequently $\mathbf{K}_t$ is the $(n_t \times n)$ 'test' matrix whose elements are $\mathrm{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$ where $\{\mathbf{x}_i\}_{i=n+1}^{n+n_t}$ and $\{\mathbf{x}_j\}_{j=1}^n$ are testing and training points, respectively.

At the beginning of the previous chapter we assumed a centralized PLS regression problem. To centralize the mapped data in a feature space $\mathcal{F}$, we can simply used the equations (3.4) and (3.14), respectively.

In conclusion, we would like provide two interpretations of the Kernel PLS model. For simplicity we will consider the univariate Kernel PLS regression case (i.e. $L = 1$) and we denote the $(n \times 1)$ vector $\mathbf{d} = \mathbf{U}(\mathbf{T}^T\mathbf{KU})^{-1}\mathbf{T}^T\mathbf{Y}$. Now we can represent the solution of the Kernel PLS regression in its dual form as

$$f(\mathbf{x}) = \sum_{i=1}^n d_i K(\mathbf{x}, \mathbf{x}_i)$$

which agrees with the solution of the regularized formulation of regression (2.22) given by the representer theorem in subsection 2.4.2. Using equation (4.18) we may also interpret the Kernel PLS model as a linear regression model of the form (for more detailed interpretation of linear PLS models we refer the reader to [31])

$$f(\mathbf{x}) = c_1 t_1(\mathbf{x}) + c_2 t_2(\mathbf{x}) + \ldots + c_p t_p(\mathbf{x}) = \mathbf{c}^T\mathbf{t}(\mathbf{x}) = \sum_{i=1}^p c_i t_i(\mathbf{x}), \tag{4.19}$$

where the $\{t_i(\mathbf{x})\}_{i=1}^p$ are the projections of the data point $\mathbf{x}$ onto the extracted p components and $\mathbf{c}$ is the vector of weights given by (4.14).

4.4 Kernel Ridge Regression

Ridge Regression is a well known standard statistical technique proposed in [50, 49]. Similar to Kernel PCR a straightforward connection to linear RR allows us to directly start with the Kernel RR description. Kernel RR deals with multicollinearity by assuming the linear regression model (4.4) whose solution is now achieved by minimizing

$$R_{rr}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i, \mathbf{w}))^2 + \xi\|\mathbf{w}\|^2, \tag{4.20}$$

where $f(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T\Phi(\mathbf{x})$ and ξ is a regularization term. The least-squares estimate of $\mathbf{w}$ is

$$\hat{\mathbf{w}} = (\mathbf{\Phi}^T\mathbf{\Phi} + \xi\mathbf{I})^{-1}\mathbf{\Phi}^T\mathbf{y}$$

which is biased but has lower variance compared to an OLS estimate. To make the connection to the Kernel PCR case we express the estimate $\hat{\mathbf{w}}$ in the eigensystem $\{\tilde{\lambda}_i, \mathbf{u}^i\}_{i=1}^{M}$

$$\hat{\mathbf{w}} = \sum_{i=1}^{M}(\tilde{\lambda}_i + \xi)^{-1}\mathbf{u}^i(\mathbf{u}^i)^T\mathbf{\Phi}^T\mathbf{y}$$

and corresponding variance-covariance matrix as [60]

$$cov(\hat{\mathbf{w}}) = \sigma^2\sum_{i=1}^{M}\tilde{\lambda}_i(\tilde{\lambda}_i + \xi)^{-2}\mathbf{u}^i(\mathbf{u}^i)^T.$$

We can see, that in contrast to Kernel PCR (4.8), the variance reduction in Kernel RR is achieved by giving less weight to small eigenvalue principal components via the factor ξ.

In practice we usually do not know the explicit mapping $\Phi(.)$ or its computation in the high-dimensional feature space $\mathcal{F}$ may be numerically intractable. In [116], using the dual representation of the linear RR model, the authors derived a formula for estimation of the weights $\mathbf{w}$ for the linear RR model in a feature space $\mathcal{F}$; i.e. (non-linear) Kernel RR. Again, using the fact that $K(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x})^T\Phi(\mathbf{y})$ we can express the final Kernel RR estimate of (4.20) in the dot product form [116, 16]

$$f(\mathbf{x}) = \mathbf{c}^T\mathbf{k} = \mathbf{y}^T(\mathbf{K} + n\xi I)^{-1}\mathbf{k}, \tag{4.21}$$

where $\mathbf{K}$ is again an $(n \times n)$ Gram matrix and $\mathbf{k}$ is the vector of dot products of a new mapped input example $\Phi(\mathbf{x})$ and the vectors of the training set; $\mathrm{k}_i = (\Phi(\mathbf{x}_i).\Phi(\mathbf{x}))$. It is worth noting that the same solution to the RR problem in the feature space $\mathcal{F}$ can also be derived based on the dual representation of the Regularization Networks minimizing the cost function (2.21) using the quadratic loss function $V(y_i, f(\mathbf{x}_i)) = (y_i - f(\mathbf{x}_i))^2$ [34, 35, 45] or through the techniques derived from Gaussian processes [163, 16].

We can see that including a possible bias term into the model leads to its penalization through the ξ term. However, in the case of regression or classification tasks there is no reason to penalize the shift of f by a constant. It was pointed out in [23], that in the case of a radial kernel we can overcome this by using a new kernel of the form

$$\tilde{K}(\mathbf{x}, \mathbf{y}) = K(\mathbf{x}, \mathbf{y}) - \xi_0$$

The ξ_0 is chosen to construct a new RKHS consisting only of zero-mean functions; i.e. $\tilde{K}$ without the zeroth order Fourier component (see Appendix A.4). Effectively, the new kernel $\tilde{K}$ induces the null space of the constant functions which are not included in a new RKHS norm and based on the cost function (2.21) are not penalized[7]. Now, the solution (2.23) will

[7] In fact, we do not need to constrain ourselves to the construction of a RKHS with only constant functions not included in the norm. Similar to SVR, we can consider a new extra constant term not included in the norm $\|f\|_{\mathcal{H}}$ and thus 'balance' the penalization of the potential constant feature of a initial kernel K by this new, not penalized, term.

take the form [34, 35, 156, 23]

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i \tilde{K}(\mathbf{x}, \mathbf{x}_i) + \tilde{b} = \sum_{i=1}^{n} c_i (K(\mathbf{x}, \mathbf{x}_i) - \xi_0) + \tilde{b} = \sum_{i=1}^{n} c_i K(\mathbf{x}, \mathbf{x}_i) + b \qquad (4.22)$$

and the unknown coefficients $\{c_i\}_{i=1}^n, b = \tilde{b} - \sum_{i=1}^n c_i \xi_0$ can be found by solving the following system of linear equations [34, 23]

$$(\tilde{\mathbf{K}} + n\xi\mathbf{I})\mathbf{c} + \mathbf{1}\tilde{b} = (\mathbf{K} + n\xi\mathbf{I})\mathbf{c} + \mathbf{1}b = \mathbf{y}$$
$$\sum_{i=1}^{n} c_i = 0 \qquad (4.23)$$

Thus we still can use a positive definite kernel K as the only change is to estimate new b term. Let us note that in the case of using a quadratic loss function in the SVR model described in section 4.5, the general quadratic optimization problem for finding the estimate of the weights is transformed to the solution of the linear equations (4.23) [132]. In fact, in such a case the same linear regression models in a feature space are assumed.

Another technique in removing a 'bias' term problem is to 'centralize' the regression problem in feature space; i.e. assume the sample mean of the mapped data $\Phi(\mathbf{x}_i)$ and targets y to be zero. This will lead to the regression estimate $f(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T\Phi(\mathbf{x})$ without the bias term. The centralization of the individual mapped data points $\Phi(\mathbf{x})$ can be achieved by the same 'centralization' of the Gram matrix $\mathbf{K}$ and vector $\mathbf{k}$ given by equation (3.4) and (3.14), respectively.

4.5 Support Vector Regression

As we noted at the beginning of the chapter, SVR is a technique where the solution to (4.4) is found by the minimization of the following regularized risk functional

$$\frac{1}{n}\sum_{i=1}^{n} V(y_i, f(\mathbf{x}_i; \mathbf{w}, b)) + \xi\|\mathbf{w}\|^2, \qquad (4.24)$$

where we assume the estimate of the desired regression function has the form $f(\mathbf{x}; \mathbf{w}, b) = \mathbf{w}^T\Phi(\mathbf{x}) + b$ with an extra, not penalized bias term considered. In contrast to Kernel RR cost functions somewhat different from quadratic are usually employed [128]. In practice, a frequently used cost function is Vapnik's ϵ-insensitive cost function (2.7) and SVR notation is usually associated with this type of regression (in the following we implicitly assume this type of SVR).

Vapnik in [153] has also shown that the regression estimate that minimizes the risk functional (4.24) with the ϵ-insensitive cost function is of the form

$$f(\mathbf{x}) = \sum_{i=1}^{n} (\gamma_i^* - \gamma_i) K(\mathbf{x}_i, \mathbf{x}) + b, \qquad (4.25)$$

where $\{\gamma_i, \gamma_i^*\}_{i=1}^n$ are Lagrange multipliers given by the maximization of the quadratic form

$$\max_{\gamma_i, \gamma_i^*} \; [-\epsilon \sum_{i=1}^{n} (\gamma_i + \gamma_i^*) + \sum_{i=1}^{n} (\gamma_i - \gamma_i^*) y_i - \frac{1}{2} \sum_{i,j=1}^{n} (\gamma_i - \gamma_i^*) K(\mathbf{x}_i, \mathbf{x}_j)(\gamma_j - \gamma_j^*)]$$

$$\text{subject to} \quad \textstyle\sum_{i=1}^{n} (\gamma_i^* - \gamma_i) = 0$$
$$0 \leq \gamma_i, \gamma_i^* \leq \tfrac{1}{2n\xi} \qquad (4.26)$$

To solve this optimization problem a primal-dual interior point method [150] is usually employed. As a consequence of using the ϵ-insensitive cost function only some of the coefficients pairs $\{\gamma_i^*, \gamma_i\}_{i=1}^n$ will become non-zero; i.e. $(\gamma_i^* - \gamma_i) \neq 0$. Effectively it leads to a sparse solution in (4.25). The data points associated with the coefficients where $(\gamma_i^* - \gamma_i) \neq 0$ are called *support vectors*.

4.5.1 Multi-Layer SVR

Combining the Kernel PCA preprocessing step with SVR yields a multi-layer SVR (MLSVR) in the following form [120]

$$f(\mathbf{x}) = \sum_{i=1}^{n}(\gamma_i - \gamma_i^*)K_1(\boldsymbol{\beta}(\mathbf{x}_i), \boldsymbol{\beta}(\mathbf{x})) + b,$$

where components of vectors $\boldsymbol{\beta}$ are defined by (3.11). However, in practice the choice of appropriate kernel function K_1 can be difficult. In this study, a linear kernel $K_1(\mathbf{x}, \mathbf{y}) = (\mathbf{x}.\mathbf{y})$ is employed. We are thus performing a linear SVR on the p-dimensional sub-space of $\mathcal{F}$ spanned by the selected nonlinear principal components.

It was shown in [121] that using the multi-layer support vector machines (SVM) approach on classification problems provides results comparable with direct nonlinear SVM approaches. However, when the linear kernel K_1 is used it speeds up the overall computations. This is generally not true in the case of MLSVR. However, similar to Kernel PCR and Kernel PLS the usefulness of the approach is rather based on the possibility of using the real LV structure of the regressors in feature space and/or their de-noising if the regressors are contaminated by a certain level of noise. We discuss this issue in the next section.

4.6 De-noising in a feature space

White additive noise will change the covariance matrix of the investigated signal by adding a diagonal matrix to it, with corresponding variances of individual noise components on the diagonal. In the case of isotropic noise this will lead to the same increase in all eigenvalues computed from the clean signal. If the signal to noise ratio is sufficiently high we can assume that the noise will mainly affect the directions of the principal components corresponding to smaller eigenvalues. This allows us to discard the finite variance due to the noise by projection of the data onto the principal components corresponding to higher eigenvalues. However, a nonlinear transformation of the measured signal consisting of a signal and additive noise can smear the noise into certain directions. Moreover, the nonlinear transformation into a feature space will also 'destroy' essential additivity and uncorrelatedness of the noise. Thus, discarding the finite variance due to the noise can lead to a higher loss of the signal information; i.e. we have to deal with the balance between noise reduction and information loss. We have investigated this situation in the case of the noisy Mackey-Glass time series (see chapter 5) and the nonlinearity $\Phi(.)$ induced by using the Gaussian kernel. From Figure 4.1 we can see that the noise increases the variance in the directions with smaller eigenvalues but decreases the variance in the main signal components. We can infer from this that a more uniform smearing of the investigated signal into all directions was induced. Cutting the directions with the smaller eigenvalues will provide a level of noise reduction, however loss of information in the main signal direction will also appear.

In spite of the above possible disadvantages, in [77, 119] promising results in digit de-noising were demonstrated using Kernel PCA. In fact, the advantage of using (nonlinear) Kernel PCA over its linear variant is the possibility of extracting up to $(n-1)$ principal components able to extract interesting nonlinear structures in data. If the information about the data structure in N dimensional input space is spread into all dimensions we cannot reduce the data structure by linear PCA without a significant loss of information. On the other hand, the extraction of up to $(n-1) > N$ nonlinear principal components may provide a higher chance to reduce the noisy components in the data while maintaining the information about structure in the data.

We have shown that all regression techniques presented in this chapter shrink the OLS solution from the directions of low data spread; i.e. from eigendirections corresponding to small eigenvalues. We may hypothesize that in situations where these eigendirections represent mainly the noisy part of the signal, the LV projection methods – Kernel PCR,

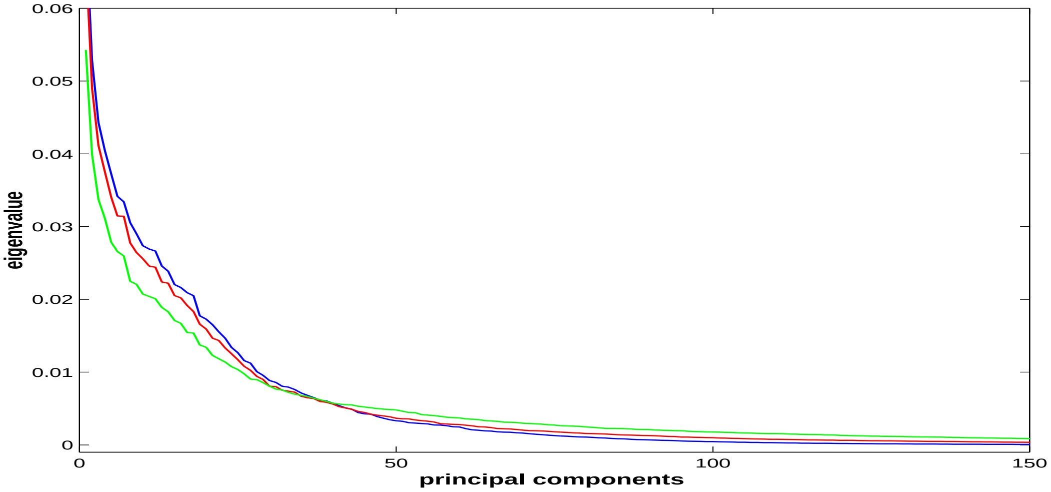

Fig. 4.1: Eigenvalues computed from embedded Mackey-Glass time series transformed to kernel space. Different levels of the noise were added (*n/s* represents the ratio between standard deviation of the noise and signal, respectively); *n/s=0%* (blue line), *n/s=11%* (red line), *n/s=22%* (green line).

Kernel PLS and MLSVR – can be profitable due to the data not being projected onto these eigendirections.

4.7 Model Selection

To determine unknown parameters in all regression models, cross validation (CV) techniques were used [130]. While in Kernel RR, a regularization term and parameters of the kernel function have to be estimated, in Kernel PLS and Kernel PCR it is mainly the problem of appropriate selection of (principal) components. The case of SVR is similar to Kernel RR but extended to the problem of setting the ϵ parameter in the Vapnik's ϵ-insensitive cost function (2.7). In the same way as Kernel PCR, MLSVR further requires the appropriate selection of principal components.

For a comparison of models using particular values of estimated parameters, the prediction error sum of squares (PRESS) statistic was used.

$$\mathrm{PRESS} = \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2,$$

where $f(\mathbf{x}_i)$ represents the prediction of the measured response y_i. PRESS was summed over all CV subsets.

4.7.1 Kernel PLS

In Kernel PLS the number of components gradually increases until the model reaches its optimal dimension. We can use CV to determine the adequacy of the individual components to enter the final model [169] or to use CV for the comparison of whole models of certain dimensionality $1, 2, \ldots, p$. In our study we used the second approach and the validity of individual models was compared in terms of PRESS.

4.7.2 Kernel PCR and MLSVR

The situation is rather more difficult in the case of Kernel PCR due to the fact that principal components are extracted solely based on the description of the input space without using any existing correlations with the outputs. The influence of individual principal components

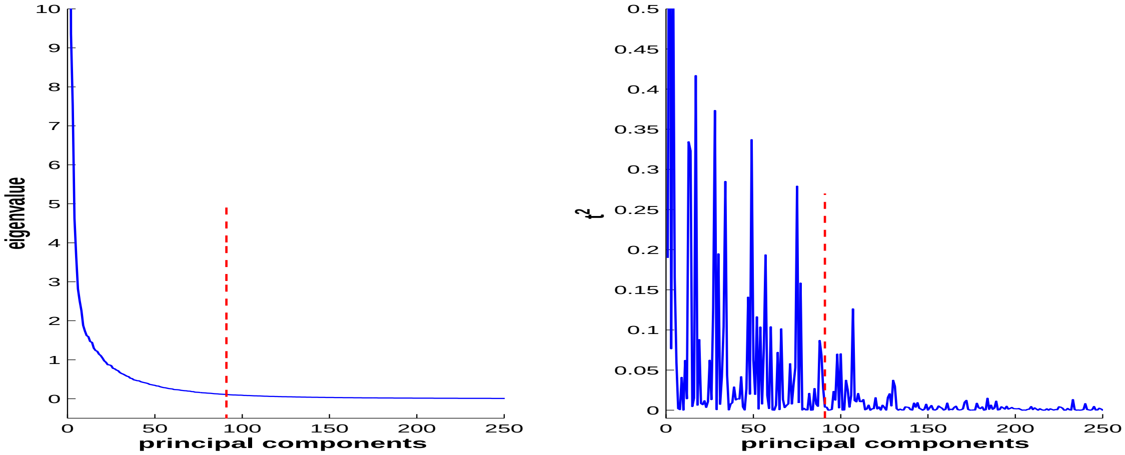

Fig. 4.2: Example of the estimated ordered eigenvalues (left) and t^2 statistic (right) of the regressors created by the projection onto corresponding principal components. Vertical (red) dashed lines indicate a number of principal components describing 95% of the overall variance of the data. One of the training data partitions of subject D from the regression problem described in subsection 5.1.2 was used.

regressors can be consequently measured by the t-test for regression coefficients [80]. By assuming a centralized regression model (4.6) for which the design matrix $\mathbf{B}$ satisfies $\mathbf{B}^T\mathbf{B} = \mathbf{I}$, we can write for the t^2 statistic of k-th regressor $t_k^2 \equiv (\boldsymbol{\beta}_k^T\mathbf{Y})^2$ where $\boldsymbol{\beta}_k$ represents the $(n \times 1)$ vector of the projections of input data onto the k-th principal component. The condition $\mathbf{B}^T\mathbf{B} = \mathbf{I}$ simply means sphering of the projected data which can be achieved on training data by taking $\mathbf{P} = \tilde{\mathbf{U}}$ in equation (3.12).

There are several different situations which can occur in PCR. First, the principal directions with large eigenvalues and significant values of t^2 should always be used in the final model. The principal directions with high eigenvalues and insignificant values of t^2 should also be included in the final model due to the fact that a significant amount of variability of the input data can be lost. The principal directions with low eigenvalues and insignificant values of t^2 should always be deleted. The most difficult problems arise when some of the directions with small eigenvalues have a significant contribution to prediction. This situation on the data sets used was already demonstrated in [111, 110]. Moreover, in Figure 4.2 we also give one of the examples we observed on the used data sets. Comparing the left and right graphs we can see that some of the small eigenvalues principal components may have relatively high prediction properties. In contrast we can see that t^2 values of some higher eigenvalues principal components indicate their low contribution to the overall prediction abilities of the regression model. For further discussion on the principal components selection topic we refer the reader to [60, 131].

First, we would like to stress that as a consequence of the orthogonality of regressors the individual single variable models have an independent contribution to the overall regression model. This significantly simplifies the selection of individual regressors and in our study we decided on the following model selection strategy. We were iteratively increasing the number of large eigenvalue principal components entering the model without considering their values of t^2 statistics. The criterion employed was the amount of explained variance. The rest of the principal components were ordered based on the t^2 statistics. Similar to Kernel PLS, CV was used to compare the whole models of particular dimension. However, in contrast to Kernel PLS the PRESS statistics were used to select the final model over all possible arrangements of the final models; i.e. for a different, fixed number of principal components with large eigenvalues entering the final model.

In the case of MLSVR the selection of principal components was entirely based on a criterion reflecting the described variance of selected principal components.

5. EXPERIMENTS

5.1 Data Sample Construction

The performance of the kernel-based regression techniques described in the previous chapter were compared using two data sets. Selection of the first artificially generated Mackey-Glass time series was motivated by the fact that the differential equation describing this chaotic system was designed to model the control of white blood-cell production [72] and belongs to the category of simulated physiological data sets. The second data set reflects an actual problem of estimating the signal detection performance of humans from measured Event Related Potentials (ERPs).

5.1.1 Chaotic Mackey-Glass Time-Series

The chaotic Mackey-Glass time-series is defined by the differential equation

$$\frac{ds(t)}{dt} = -bs(t) + a\frac{s(t-\tau)}{1+s(t-\tau)^{10}}$$

with $a = 0.2$, $b = 0.1$. The data were generated with $\tau = 17$ and using a second-order Runge-Kutta method with a step size 0.1. Training data is from t=200 to t=3200 while test data is in the range t= 5000 to 5500. To this generated time-series we added white noise with normal distribution and with different levels corresponding to ratios of the standard deviation of the noise and the clean Mackey-Glass time-series.

5.1.2 Human Signal Detection Performance Monitoring

We have used Event Related Potentials and performance data from an earlier study [145, 146, 66]. Eight male Navy technicians experienced in the operation of display systems performed a signal detection task. Each technician was trained to a stable level of performance and tested in multiple blocks of 50–72 trials each on two separate days. Blocks were separated by 1-minute rest intervals. A set of 1000 trials were performed by each subject. Inter-trial intervals were of random duration with a mean of 3s and a range of 2.5–3.5s. The entire experiment was computer-controlled and performed with a 19-inch color CRT display (Figure 5.1). Triangular symbols subtending 42 minutes of arc and of three different luminance contrasts (0.17, 0.43 or 0.53) were presented parafoveally at a constant eccentricity of 2 degrees visual angle. One symbol was designated as the target, the other as the non-target. On some blocks, targets contained a central dot whereas the non-targets did not. However, the association of symbols to targets was alternated between blocks to prevent the development of automatic processing. A single symbol was presented per trial, at a randomly selected position on a 2-degree annulus. Fixation was monitored with an infrared eye tracking device. Subjects were required to classify the symbols as targets or non-targets using button presses and then to indicate their subjective confidence on a 3-point scale using a 3-button mouse. Performance was measured as a linear composite of speed, accuracy, and confidence. A single measure, PF1, was derived using factor analysis of the performance data for all subjects, and validated within subjects. The computational formula for PF1 was

$$\text{PF1} = 0.33{*}\text{Accuracy} + 0.53{*}\text{Confidence - } 0.51{*}\text{Reaction Time}$$

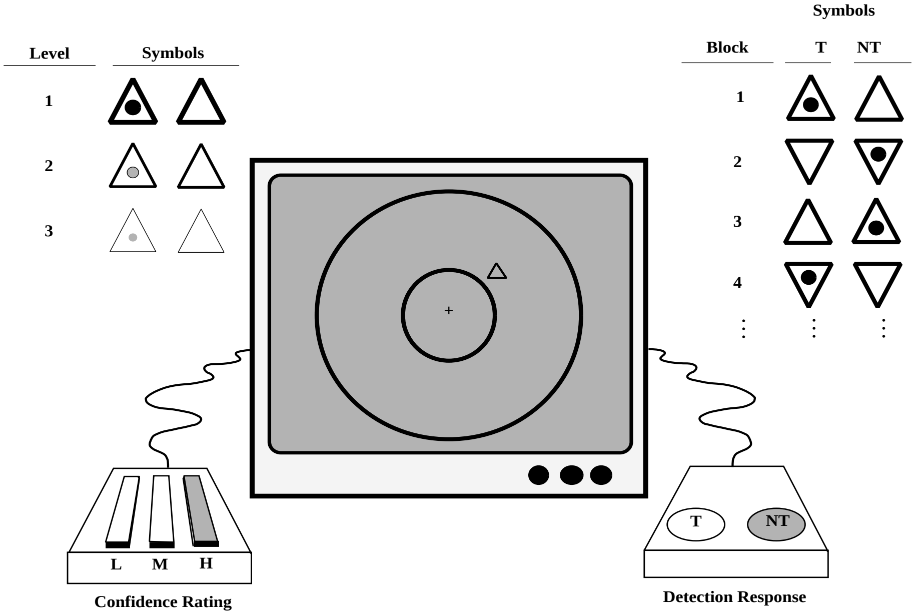

Fig. 5.1: Display, input device configuration and symbols for task-relevant stimuli for the signal detection task.

using standard scores for accuracy, confidence, and reaction time based on the mean and variance of their distributions across all subjects. PF1 varied continuously, being high for fast, accurate, and confident responses and low for slow, inaccurate, and unconfident responses. In our experiments we linearly normalized PF1 to have a range of 0 to 1.

ERPs were recorded from midline frontal, central, and parietal electrodes (Fz, Cz, and Pz), referred to average mastoids, filtered digitally to a bandpass of 0.1 to 25Hz, and decimated to a final sampling rate of 50Hz. The prestimulus baseline (200 ms) was adjusted to zero to remove any DC offset. Vertical and horizontal electrooculograms (EOG) were also recorded. Epochs containing artifacts were rejected and EOG-contaminated epochs were corrected. Furthermore, any trial in which no detection response or confidence rating was made by a subject was excluded along with the corresponding ERP.

Within each block of trials, a running-mean ERP was computed for each trial (Figure 5.2). Each running-mean ERP was the average of the ERPs over a window that included the current trial plus the 9 preceding trials for a maximum of 10 trials per average. Within this 10-trial window, a minimum of 7 artifact-free ERPs were required to compute the running-mean ERP. If fewer than 7 were available, the running mean for that trial was excluded. Thus each running mean was based on at least 7 but no more than 10 artifact-free ERPs. This 10-trial window corresponds to about 30s of task time. The PF1 scores for each trial were also averaged using the same running-mean window applied to the ERPs, excluding PF1 scores for trials in which ERPs were rejected. Prior to analysis, the running-mean ERPs were clipped to extend from time zero (stimulus onset time) to 1500 ms post-stimulus, for a total of 75 time points.

5.2 *Results*

The present work was carried out with Gaussian kernels; $K(\mathbf{x},\mathbf{y}) = e^{-(\|\mathbf{x}-\mathbf{y}\|^2/d)}$, where d determines the width of the Gaussian function. The Gaussian kernel possesses good smoothness properties (suppression of the higher frequency components) and in the case where we do not have *a priori* knowledge about the regression problem we would prefer a smooth estimate [34, 129].

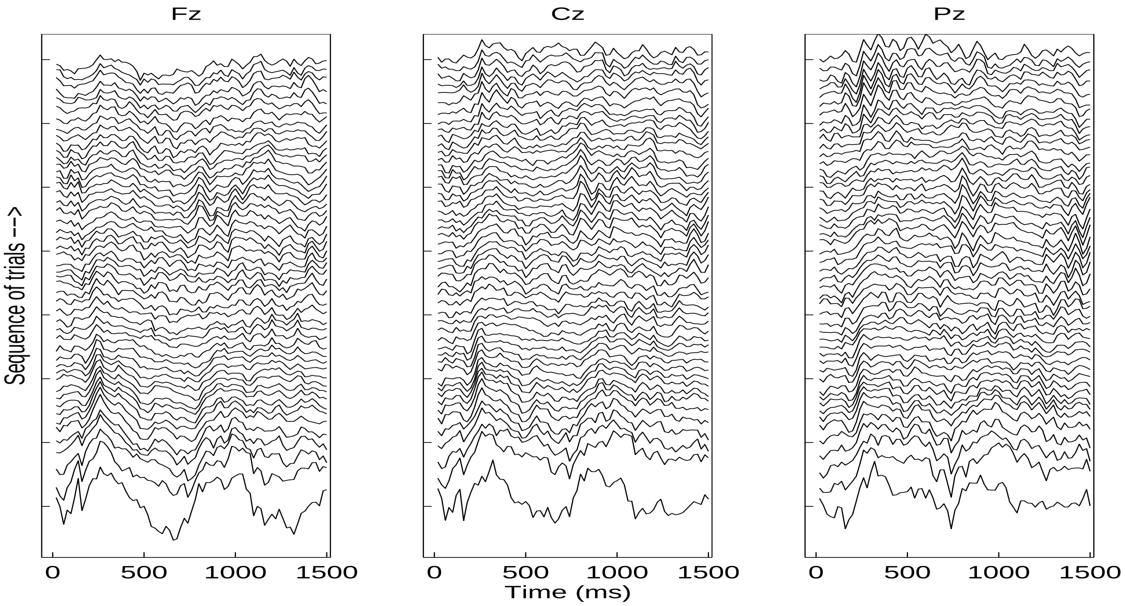

Fig. 5.2: Running-mean ERPs at sites Fz, Cz and Pz for subject B in the first 50 running-mean ERPs.

5.2.1 Chaotic Mackey-Glass Time-Series

The kernel-based regression models with quadratic cost function, i.e. Kernel PCR, Kernel PLS and Kernel RR, were trained to predict the value sampled 85 steps ahead from inputs at time $t, t-6, t-12, t-18$. The training data partitions were constructed by moving a 'sliding window' over the 3000 training samples in steps of 250 samples. This window had a size of 500 samples and 1000 samples, respectively. The validation set was then created by using the following 250 and 500 data points. This created ten partitions of size 500/250 (training/validation) samples and seven partitions of size 1000/500 (training/validation) samples.

We estimated the variance of the overall clean training set and based on this estimate $\hat{\sigma}^2 \doteq 0.05$ the CV technique was used to find the optimal width d from the range $\langle 0.01, 0.2 \rangle$ using the step size 0.01. A fixed test set of size 500 data points (see subsection 5.1.1) was used in all experiments. The performance of the regression models to predict 'clean' Mackey-Glass time series was evaluated in terms of normalized root mean squared error (NRMSE) defined as

$$\text{NRMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \ ,$$

where $\hat{y}$ represents an estimation of the signal of interest on noisy time-series and y the 'clean' (noise free) Mackey-Glass time-series, respectively.

The results achieved using the individual regression models are summarized in Table 5.1. In terms of NRMSE we may observe approximately the same performance of all the methods employed. However, comparing Kernel PLS and Kernel PCR we can observe a significant reduction in the number of the components used in the case of Kernel PLS regression. In some cases Kernel PLS utilizes less than 10% of the components utilized by Kernel PCR.

Increasing the value of d leads to a faster decay of the eigenvalues (see e.g. [166]) and to the potential loss of the 'finer' data structure due to a smaller number of the nonlinear principal components describing the same percentage of all the data variance. Increasing levels of the noise has the tendency to increase the optimal value for the d parameter which coincides with the intuitive assumption about smearing out the local structure. In contrast small values of d will lead to 'memorizing' of the training data structure. Thus, in Figures 5.3 and 5.4 we also compared the results on the noisy time series and their dependence on

Method	*n/s=0.0%*		*n/s=11%*		*n/s=22%*	
	500	1000	500	1000	500	1000
Kernel PLS	0.048 (0.031)	0.007 (0.002)	0.322 (0.030)	0.279 (0.004)	0.455 (0.021)	0.414 (0.010)
# of C	155 (38)	192 (57)	7 (2)	8 (1)	6 (2)	6 (2)
Kernel PCR	0.046 (0.030)	0.009 (0.004)	0.327 (0.030)	0.284 (0.004)	0.462 (0.031)	0.423 (0.011)
# of PC	383 (78)	593 (188)	79 (35)	119 (49)	48 (24)	87 (59)
Kernel RR	0.044 (0.027)	0.007 (0.002)	0.321 (0.041)	0.276 (0.005)	0.451 (0.029)	0.412 (0.008)

Tab. 5.1: The comparison of the approximation errors (NRMSE) of prediction, the number of used components (C) and principal components (PC) for 2 different sizes of Mackey-Glass training set. The values represent an average of 10 simulations in the case of 500 training points and 7 simulations in the case of 1000 training points, respectively. Corresponding standard deviation is presented in parentheses. *n/s* represents the ratio between the standard deviation of the added Gaussian noise and the underlying time-series.

the width d of the Gaussian kernel for the case when training sets were of size 500 data points. We may observe a smaller range of the d values on which Kernel PLS and Kernel PCR achieves the optimal results on the testing set compared to Kernel RR. However, the results also suggest a smaller variance in the case of latent variable projection methods; i.e. Kernel PLS and Kernel PCR.

5.2.2 Human Signal Detection Performance Monitoring

To be consistent with the previous results reported in [145, 66] the validity of the models was measured in terms of normalized mean squared error (NMSE) and in terms of the proportion of data for which PF1 was correctly predicted with 10% tolerance (test proportion correct (TPC)); i.e ±0.1 in our case.

In our pilot study [110] the performance of SVR and Kernel RR methods trained on data preprocessed by linear PCA in the input space was compared with the results achieved by using MLSVR and Kernel PCR on features extracted by Kernel PCA. We trained the models on 50% of the ERPs and tested on the remaining data[1]. The described results, for each setting of the parameters, are an average of 10 runs each on a different partition of training and testing data. In the case of Kernel RR the regularization term ξ was estimated by cross-validation using 20% of the training data set as a validation set. We used $\epsilon = 0.01$ and $\xi = 0.01$ parameter values for SVR and MLSVR models. The criterion for the selection of principal components was the amount of variance described by the selected principal components. To summarize these results in Figures 5.5 and 5.6 the results achieved on subject C(417 ERPs), D(702 ERPs), F(614 ERPs) are depicted. From the figures we can see consistently better results on features extracted by Kernel PCA on subjects D and F. These superior results achieved using the Kernel PCA representation were also observed on the remaining 5 subjects.

[1] In contrast to [146, 145] where one training (odd-numbered blocks of trials)-testing (even-numbered blocks of trials) data pair was used, in our studies we created different training-testing data partitions by random sampling from all blocks of trials. As the within-block correlations are much higher than between blocks this leads to a significant improvement in the results [110]. However, this selection of training-testing partitions is irrelevant for the comparison of the kernel-based regression techniques investigated.

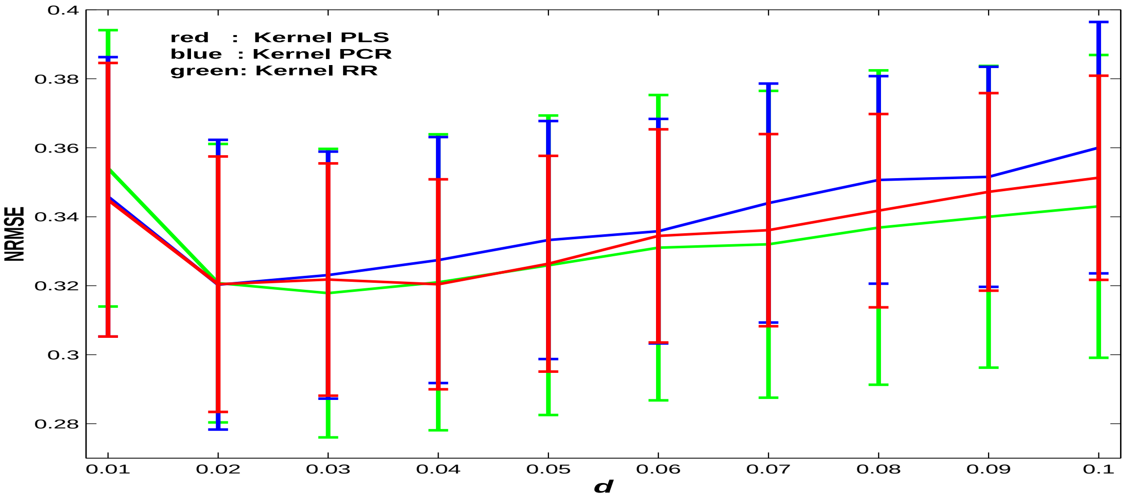

Fig. 5.3: Comparison of the results achieved on the noisy Mackey-Glass (*n/s=11%*) time series with the Kernel PLS(red), Kernel PCR (blue) and Kernel RR (green) methods. Ten different training sets of size 500 data points were used. The performance for different widths (d) of the Gaussian kernel is compared in normalized root mean squared error (NRMSE) terms. The error bars represent the standard deviation on results computed from ten different runs. n/s represents the ratio between the standard deviation of the added Gaussian noise and the underlying time-series.

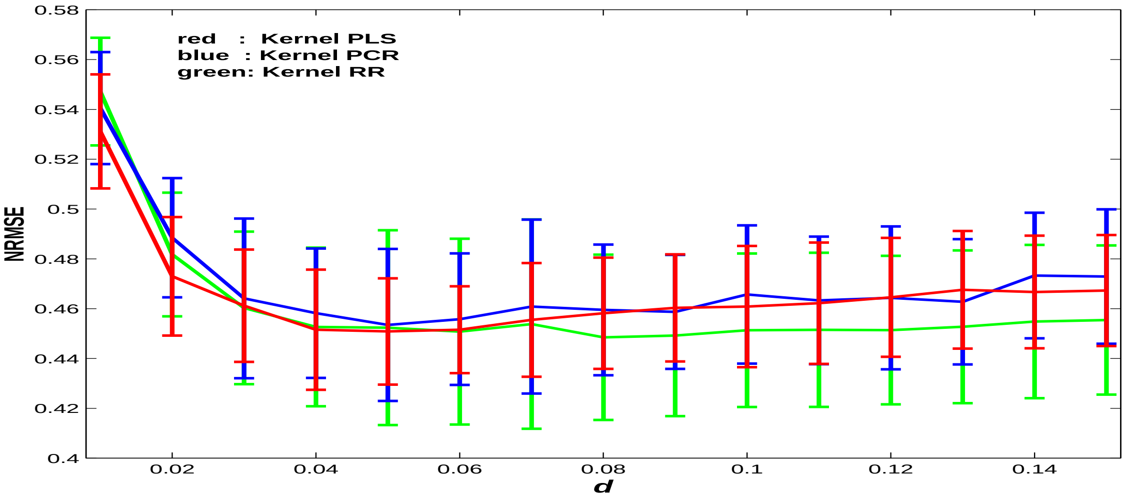

Fig. 5.4: Comparison of the results achieved on the noisy Mackey-Glass (*n/s=22%*) time series with the Kernel PLS(red), Kernel PCR (blue) and Kernel RR (green) methods. Ten different training sets of size 500 data points were used. The performance for different widths (d) of the Gaussian kernel is compared in normalized root mean squared error (NRMSE) terms. The error bars represent the standard deviation on results computed from ten different runs. n/s represents the ratio between the standard deviation of the added Gaussian noise and the underlying time-series.

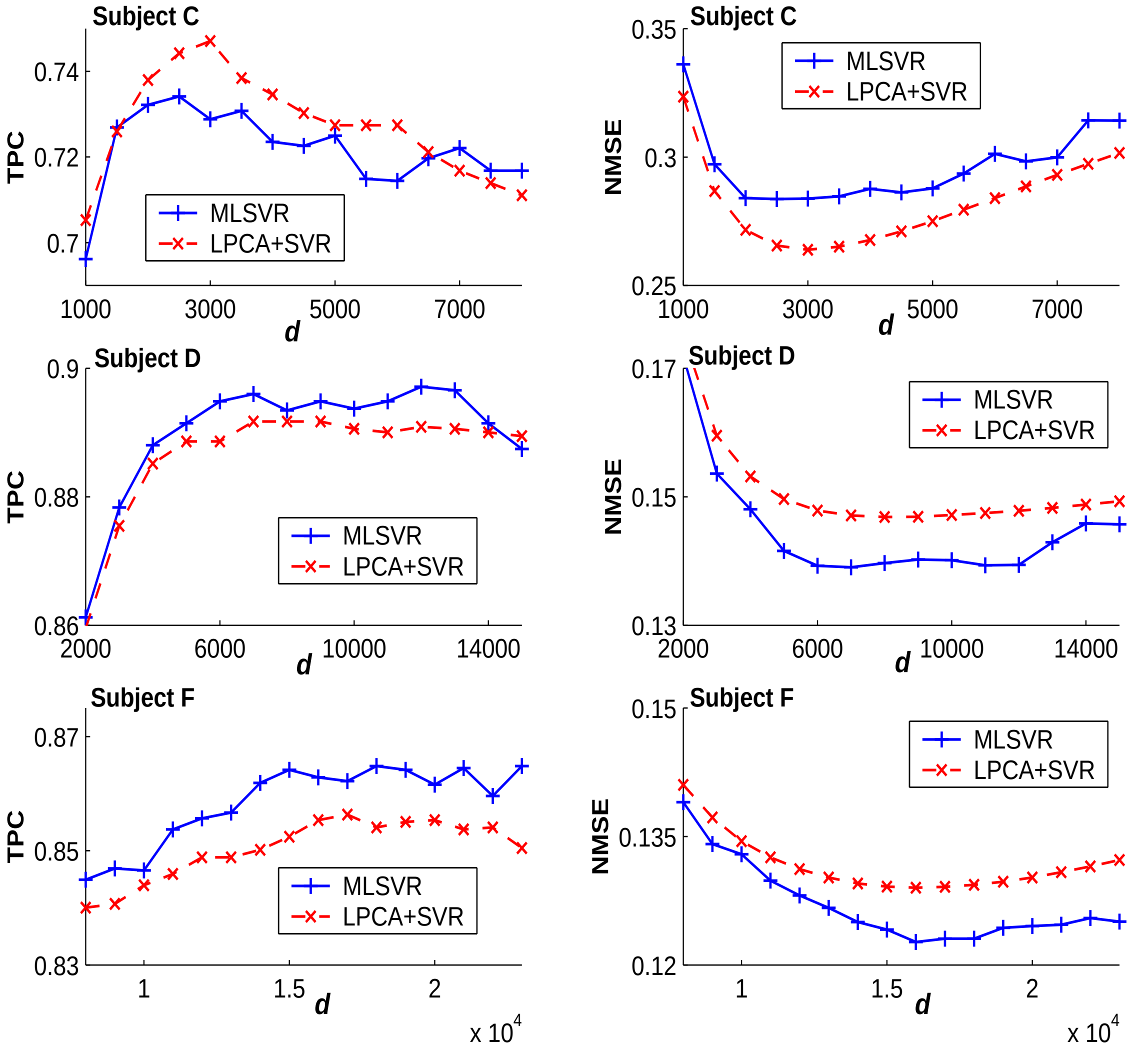

Fig. 5.5: Comparison of the results achieved on subjects C, D and F with MLSVR and SVR on data preprocessed by linear PCA (LPCA + SVR), respectively. In both cases the principal components describing 99% of variance were used. The performance for the different widths (d) of the Gaussian kernel is compared in terms of test proportion correct (TPC) and normalized mean squared error (NMSE).

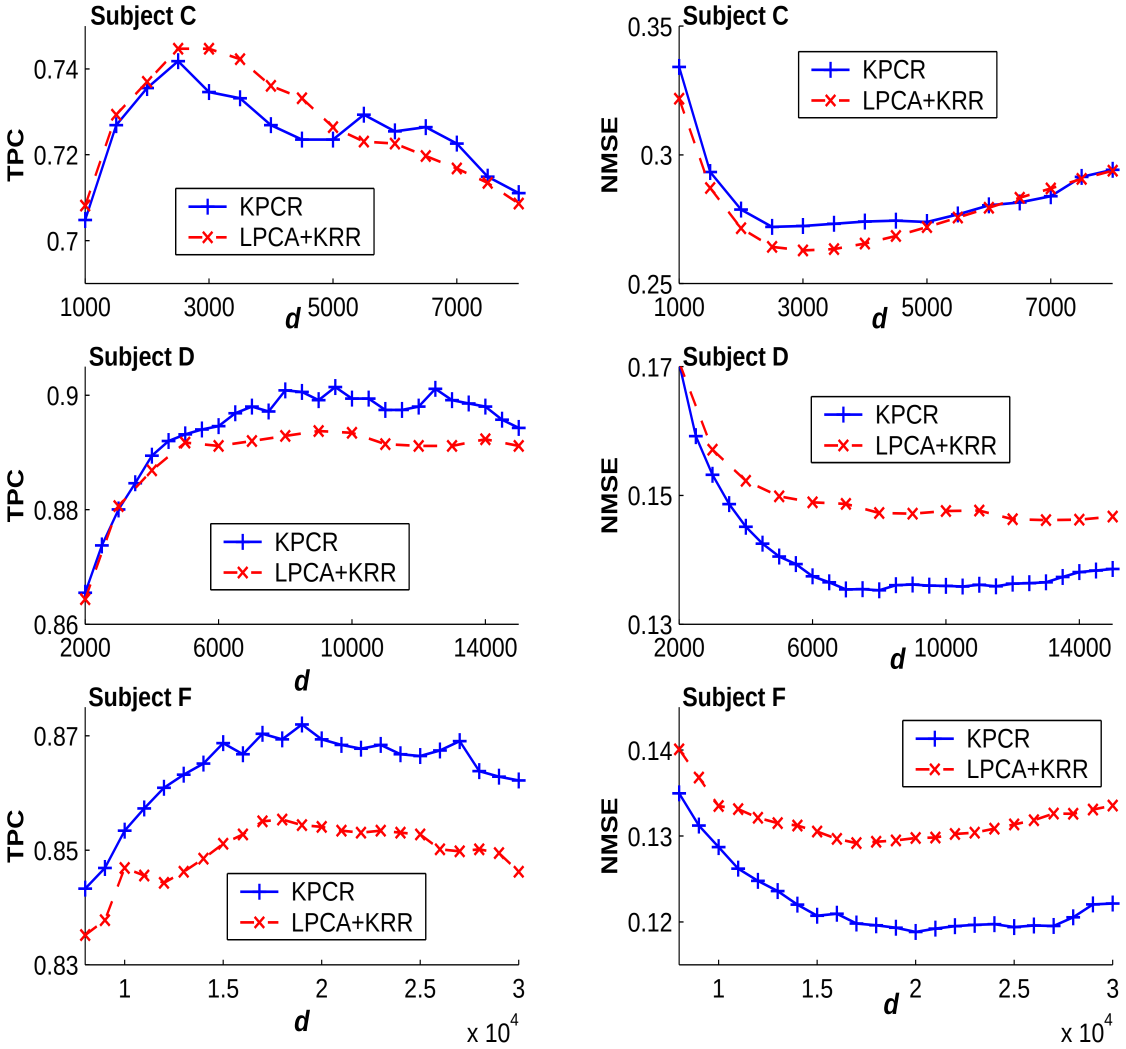

Fig. 5.6: Comparison of the results achieved on subjects C, D and F with Kernel PCR (KPCR) and Kernel RR (KRR) on data preprocessed by linear PCA (LPCA + KRR), respectively. In both cases the principal components describing 99% of variance were used. The performance for the different widths (d) of the Gaussian kernel is compared in terms of test proportion correct (TPC) and normalized mean squared error (NMSE).

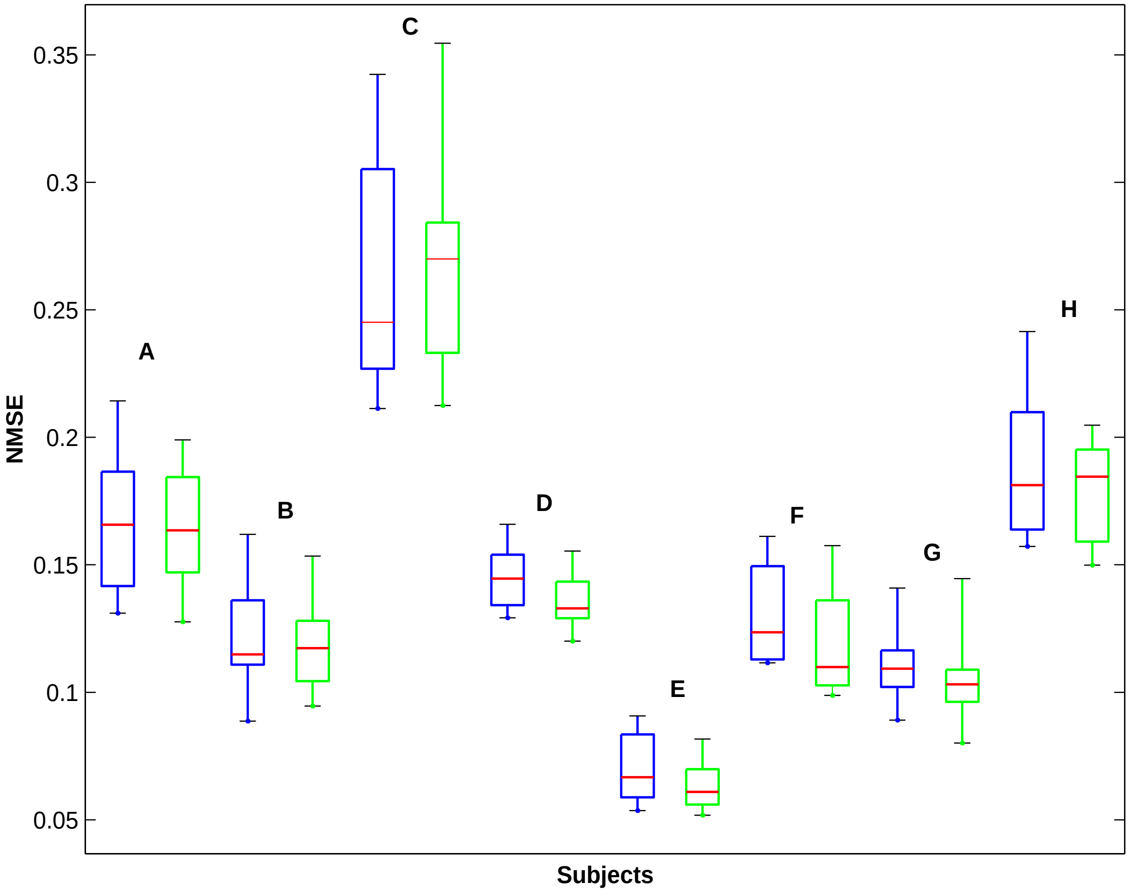

Fig. 5.7: Boxplots with lines at the lower quartile, median, and upper quartile values and whisker plot for subjects A to H. The performance of Kernel RR with linear PCA preprocessing step (left-hand (blue) boxplots) is compared with Kernel PCR on data preprocessed by Kernel PCA (right-hand (green) boxplots) in terms of normalized mean squared error (NMSE). The boxplots are computed on results from 10 different runs using the widths of the Gaussian kernel on which the methods achieved minimal NMSE on test set.

However, on subject C the performance with the features selected by linear PCA was slightly better. In the next step, for individual subjects, we selected the results for a Gaussian kernel width d on which Kernel RR (with linear PCA preprocessed data) and Kernel PCR (with Kernel PCA preprocessing) achieved the minimal NMSE on the test set. In Figure 5.7 a boxplot with lines at the lower quartile, median, and upper quartile values and a whisker plot for individual subjects is depicted. The boxplots indicate the differences between the results on subjects D to H. Using the sign test and the Wilcoxon matched-pairs signed-ranks test we tested the hypotheses about the *direction* and *size* of the differences within pairs. On subjects D to H the p-values < 0.03 indicate that the difference between the results achieved using linear PCA and Kernel PCA preprocessing steps is statistically significant. The alternative hypothesis regarding the superiority of linear PCA leads to p-values < 0.02. Although both tests on subjects A, B and C did not show a statistically significant difference between the results (p-values between 0.11 and 0.75), the alternative Wilcoxon test about the superiority of linear PCA leads to a higher p-value only on subject C (A - 0.12, B - 0.25, C - 0.88). Note that on subject C the smallest number of ERPs is available (417). Figure 5.7 also indicates the weakest results with the highest variance over individual runs. This result suggests that the number of ERPs from the subject C were insufficient to model the desired dependencies between ERPs and the subject performance. Moreover, in this case the dimension of matrix $\mathbf{K}$ in the feature space $\mathcal{F}$ is lower (209) than the input dimensionality (225) and so we cannot exploit the advantage of Kernel PCA to improve overall performance

Method	NMSE	TPC
Kernel PCR (*with EMKPCA*)	0.154	83.3
Kernel RR	0.155	83.5
SVR (*with SVMTorch*)	0.161	82.8

Tab. 5.2: The comparison of the NMSE and TPC prediction errors on the test set for the model based on all subjects ERPs. The values represent an average of 3 different simulations.

by using more components in the feature space than the number available in the input space.

In [110], comparing SVR with MLSVR, we have also demonstrated that without the Kernel PCA preprocessing step in the feature space $\mathcal{F}$ we did not increase the overall performance. On the contrary, on subjects A, B and H the performance using the MLSVR method was slightly superior. On the remaining subjects the difference was insignificant. Note that these results were based on fixing the $\epsilon = 0.01$ and $\xi = 0.01$ values to be the same in both SVR and MLSVR approaches. Further we selected 90% of all nonlinear principal components in the case of MLSVR. Thus, the better results achieved with MLSVR provide us with only an indication about the usefulness of MLSVR rather than a conclusive answer.

Further, in [111, 110], we have also reported the results when Kernel PCR, Kernel RR and SVR were used on all eight subjects. We split the overall data set (5594 ERPs) into three different training (2765 ERPs) and testing (2829 ERPs) data pairs. 20% of the training data set was used for cross-validation to estimate ϵ, ξ and ξ parameters in SVR and Kernel RR, respectively. In the case of SVR the direct solution of the quadratic optimization problem to find the $\boldsymbol{\gamma}, \boldsymbol{\gamma}^*$ and b coefficients (4.25) was replaced by using the *SVMTorch* [14] algorithm designed to deal with large-scale regression problems. In the case of Kernel PCR, the eigenvectors and eigenvalues were estimated using the EM approach to Kernel PCA (EMKPCA) with 30 EM steps. Based on the results reported in [111], we provide the results when 2600 main nonlinear principal components and a Gaussian kernel of width $d = 6000$ were used. Table 5.2 summarizes the performance of the individual methods. We can see the slightly superior performance achieved with the Kernel PCR and Kernel RR models in comparison to SVR.

In the last part of the section we are comparing Kernel PCR, Kernel PLS, Kernel RR and SVR techniques using more extensive, CV based, search for the optimal parameters. For each individual subject we split the data into 10 different 55% and 45% training and testing partitions. Eleven-fold CV to estimate the desired parameters was applied on each training partition. After CV a final model was tested on an independent testing partition. This was repeated 10 times for each training and testing data pair. The results achieved on individual subjects in our former studies [108, 110, 111] informed our choice of the Gaussian kernel. In the case of SVR the *SVMTorch* [14] algorithm was used.

Tables 5.3 and 5.4 summarize the results achieved on eight subjects (A to H). In terms of NMSE and TPC we may see approximately the same performance of all kernel regression models. The number of components used in the case of Kernel PLS is on average 10 times lower compared to Kernel PCR, however only on subject F a slightly superior performance of Kernel PLS in terms of NMSE was observed. Further we can see that in the case of SVR we did not achieve significant sparsification of the solution and usually more than 90% of the training data points (support vectors) were used in the final model.

Subject	**Kernel PLS**		**Kernel PCR**		**Kernel RR**	**SVR**	
	NMSE	# of C	NMSE	# of PC	NMSE	NMSE	# of SV
A	0.159	27.9	0.159	373.1	0.159	0.159	421.7
(891 ERPs)	(0.025)	(16.3)	(0.023)	(30.9)	(0.025)	(0.025)	(8.9)
B	0.117	33.9	0.118	224.4	0.117	0.117	285.1
(592 ERPs)	(0.027)	(12.3)	(0.026)	(32.2)	(0.026)	(0.026)	(3.3)
C	0.259	15.5	0.260	134.8	0.253	0.258	185.9
(417 ERPs)	(0.060)	(4.1)	(0.044)	(17.9)	(0.044)	(0.039)	(31.6)
D	0.130	24.8	0.130	241.2	0.126	0.128	337.7
(702 ERPs)	(0.011)	(6.0)	(0.010)	(50.3)	(0.010)	(0.010)	(7.5)
E	0.058	42.4	0.059	274.4	0.057	0.057	348.2
(734 ERPs)	(0.006)	(20.1)	(0.006)	(42.2)	(0.007)	(0.006)	(5.5)
F	0.116	24.6	0.125	186.6	0.114	0.117	290.4
(614 ERPs)	(0.023)	(9.9)	(0.025)	(61.4)	(0.024)	(0.025)	(15.8)
G	0.105	23.6	0.107	323.3	0.105	0.105	417.4
(868 ERPs)	(0.018)	(13.5)	(0.018)	(53.2)	(0.017)	(0.017)	(3.9)
H	0.173	19.7	0.175	280.4	0.173	0.176	359.3
(776 ERPs)	(0.022)	(7.0)	(0.022)	(49.0)	(0.022)	(0.020)	(18.2)

Tab. 5.3: The comparison of the normalized mean squared error (NMSE) and the number of used components (C) and principal components (PC), respectively, for subjects A to H. The values represent an average of 10 different simulations and corresponding standard deviation is presented in parentheses.

Subject	**Kernel PLS**	**Kernel PCR**	**Kernel RR**	**SVR**
	TPC	TPC	TPC	TPC
A	84.3	84.4	84.1	84.1
B	91.2	90.3	91.2	90.9
C	74.8	73.7	74.9	74.4
D	90.4	90.2	91.0	90.9
E	94.5	94.4	94.4	94.5
F	87.3	86.1	88.1	87.6
G	90.0	89.5	89.6	89.6
H	84.8	84.6	84.8	84.8

Tab. 5.4: The comparison of the results achieved on subjects A to H in terms of the test proportion correct (TPC). The values represent an average of 10 different simulations. The corresponding standard deviation was in all cases lower than 0.03.

6. SUMMING UP

Several different nonlinear, kernel-based regression methods have been studied. The SRM principle and Regularization Theory were the main principles for their construction. A straightforward connection between a RKHS and the corresponding feature space representation of the transformed input data allowed us to consider linear regression models in a possibly very high dimensional feature space. The computations in a feature space were achieved by using the 'kernel trick' which obviates the need to carry out explicit nonlinear mappings.

To compare these regression techniques two data sets were employed, the chaotic Mackey-Glass time series prediction and the data associated with the problem of estimating human signal detection performance from the Event Related Potentials. However, in the first step we concentrated on feature extraction by Kernel PCA with the aim of making a comparison with linear PCA which had been used in previous studies [145, 66].

Kernel PCA for Feature Extraction and De-noising

The Kernel PCA method for feature extraction has been investigated and the selected features were used in a regression problem. On the performance monitoring data set, in more than half of the cases (subjects D to H), we demonstrated that the kernel regression methods with a (nonlinear) Kernel PCA preprocessing step provide significantly superior results over those with data preprocessed by linear PCA. Only in one case was an indication of the superiority of linear PCA observed, however, the sufficiency of the data representation in this case is questionable.

Moreover, we have shown that a reduction of the overall number of nonlinear principal components can reduce the noise present in the data. Similar to the investigated Mackey-Glass time series prediction task, this can be exploited especially in the situation where the low-dimensional input data are spread in all directions and the noise reduction by projection to a lower number of linear principal components leads to information loss.

The solution of the eigenvalue problem (3.3) can be numerically difficult in the case of a high number of data samples. On the noisy Mackey-Glass time series we demonstrated that we can make a sufficiently precise estimate of the main eigenvalues and eigenvectors from a smaller data representation. This implies the possibility of significantly reducing the computation and memory requirements and of practically dealing with large-scale regression problems. In fact, in [111] we experimentally demonstrated that Kernel PCR with the principal components extracted by the proposed EM approach to Kernel PCA provides the same results in comparison to the case when the principal components are extracted by standard Kernel PCA; i.e. by diagonalization of the $\mathbf{K}$ matrix in (3.3). Moreover, we have shown that EMKPCA is computationally more attractive when extraction of $p \ll n$ main principal components is required.

On the performance monitoring data set, by employing Kernel PCR on the selected nonlinear principal components we demonstrated comparable performance with the SVR technique. Moreover, the reported results suggest that on this data set the regression models with a quadratic cost function may be preferable; i.e. a Gaussian type of noise is more likely. Thus, in the next step we concentrated on regularized kernel-based regression models with a quadratic cost function and we extended this family of models with the nonlinear Kernel PLS method.

Regularized Least Squares Regression Models in RKHS

There exists a large body of literature comparing standard OLS regression with PLS, PCR and RR (see e.g. [26, 131]). Assuming a construction of regularized linear regression models in a RKHS we can make some conclusions by using the analogy with the reported observations. First, in the situation when high multicollinearity among regressors exists OLS leads to an unbiased but high variance estimate of the regression coefficients. PLS, PCR and RR are designed to shrink the solution to the regression from the areas of low data spread resulting in biased but lower variance estimates. Second, there exist real world regression problems where the number of observed variables N significantly exceeds the number of samples (observations) n – a situation quite common in chemometrics. Moreover, we may usually also observe that the real rank of the matrix of regressors is significantly lower than n and N. The projection of the original regressors to the 'real' latent variables is the main advantage of methods such as PLS or PCR. This is also similar to the situation when the input variables are corrupted by a certain amount of noise (the situation with noisy Mackey-Glass time series and the ERPs data sets). By projecting the original data to the components with higher eigenvalues we may usually discard the noise component contained in the original data. We may hypothesize that both situations are also quite common when a kernel-type of regression is assumed. Usually we nonlinearly transform the original data to the high dimensional space whose dimension M is in many cases significantly higher than the number of observations $M \gg n$.

Applying regularized regression techniques in a feature space, we observed that on both data sets employed Kernel PLS provide the same results as Kernel PCR and Kernel RR. However, in comparison to Kernel PCR, the Kernel PLS method utilizes a significantly smaller number of qualitatively different components.

Future Work

In contrast to [145] where one training (odd-numbered blocks of trials)-testing (even-numbered blocks of trials) data pair was used, in our study we created (similar to [66]) the different training-testing data partitions by random sampling from all blocks of trials. By using the kernel regression models on these data partitions we achieved approximately twice the level of improvement in terms of TPC. This is a quite significant improvement on this biomedical application. However, as the within-block correlations may be much higher than between blocks in our future work the same data setting and representation (discrete wavelet transforms of ERPs) as reported in [145] will be used to investigate the possibility of improvement of the reported results.

Employing the CV model selection technique we observed approximately the same performance on all the studied kernel regression models. It is the aim of our further study to investigate and compare different model selection techniques. This may be more interesting in the case of selection of the appropriate (principal) components entering the Kernel PLS and Kernel PCR models. In practical situations splitting of the available data set into training and validation sets leads not only to less accurate estimates of the components but also has the potential to decrease their number when $n \ll M$. In [109] we have shown good performance when an *in-sample* Covariance Inflation Criterion was used [139]. However, a more extensive comparison between the *out-sample* and *in-sample* model selection procedures may be fruitful here.

PART B

We start this part of the thesis by introducing the concept of depth of anaesthesia monitoring and explaining why, despite several decades of research, this problem still remains unsolved and is the focus for a great deal of research. We will also provide a summary of existing results, methods and recent research on this topic. This review is not intended to be comprehensive but covers the main areas of interest and references more detailed literature. We conclude the first introductory section by describing the aim of our study, the motivation for the selection of new measures, and potential contributions to the topic.

The theoretical basis, assumptions and motivations for the selection of individual measures is highlighted at the beginning of the second chapter. The complexity measures employed in the study are individually described. We start with the wider family of entropy rate measures. The methodology of their derivation and final algorithms for their use in practice are provided. Existing connections among the measures are stressed. Other complexity measures used in the study are described in the second section of this chapter.

Following this, we describe the settings, conditions and devices required for the acquisition of anaesthetic EEG data. Several aspects of the individual measures are then compared by applying them to this data set. The main focus is a comparison of their abilities to discriminate moderate and light anaesthetic states.

The results and observations are summarized in the last chapter.

7. DEPTH OF ANAESTHESIA MONITORING PROBLEM

Depth of anaesthesia (DOA) during surgery has been widely investigated in recent years, perhaps prompted by the persistent occurrence of intraoperative cases of awareness. The Royal College of Anaesthetists in the UK estimates around 1000 cases of inadvertent awareness during general anaesthesia per annum. The current gold-standard measure of anaesthetic depth is an exercise in applied pharmacology by an experienced clinician who takes into consideration relevant aspects of the following information: patient medication, age, neurological status, blood electrolyte concentrations, thyroid status, temperature, premedicant medication, intravenous anaesthetic agent doses, inhaled anaesthetic concentration, computer pharmacokinetic estimates of blood agent levels, end-tidal measures of inhaled agent concentration, patient reactions to surgery, cardio-respiratory disease, and the methods and type of analgesia employed. Much of this mass of information is interdependent, and the clinician is also aware of sources of error, their recognition and magnitudes. The majority of anaesthetics are conducted with depth assessed in this way, but despite progressive improvements in these techniques, cases of awareness under general anaesthesia still occur and have profound clinical and medico-legal consequences.

This is reliable enough for the vast majority of clinical situations but a robust, direct monitor of anaesthetic depth invaluable to clinicians still does not exist. Most of the current descriptors of depth of anaesthesia are agent specific, not monotonic, and show a time lag behind clinical changes in anaesthetic depth. Several fundamentally different groups of drugs are routinely involved in anaesthesia – usually in combination. These are benzodiazepines, opioids, inhaled volatile agents, inhaled gases, and intravenous induction agents – which may also be used by infusion for maintenance of anaesthesia. Each group of drugs has characteristic effects on the EEG waveform, and there is also considerable variation within members of each group.

7.1 DOA monitoring: Review of Methods

There is a large corpus of literature on different DOA monitoring techniques. General methods suitable for a wide range of anaesthetic drugs do exist, but either suffer from a residual degree of interpatient variability, or lack precision. There are several methods for DOA monitoring:

7.1.1 Clinical Signs

The main attributes in this category are changes of heart rate (HR), blood pressure (BP), pupillary size, eye movement, temperature and some visual signs such as swallowing and sweating. These have the disadvantage of being significantly affected by factors other than depth of anaesthesia: patient medication, disease, surgical interventions and patient age. Many drugs alter the variables listed above and thus decrease the reliability of the DOA monitoring systems. On the other hand factors such as HR, BP, sweating or temperature changes are still important for anaesthesiologists who are able to take these confounding factors into account.

7.1.2 Isolated Forearm Technique (IFT)

The IFT method was introduced by Tunstall [147] to investigate conscious awareness during obstetric anaesthesia. This technique was designed to overcome the problem of measurement of cognitive functions in the cases of anaesthetized patients who also underwent neuromuscular blockade. A blood pressure cuff isolates an arm from the neuromuscular blocking drugs, and the anaesthesiologist can talk with a patient during surgery, the patient being able to respond using the unparalyzed arm. It is now accepted that this technique is a more useful indicator of consciousness than clinical signs. It is the only technique which permits evaluation of cognitive function during surgery where the anaesthetic regimen includes muscle relaxants [137].

7.1.3 EEG Monitoring

The EEG is a microvolt-level electrical signal read from the scalp. It is derived mainly from cerebral cortical post-synaptic potentials. Cellular activity and synchrony determine the magnitude and frequency of the signal. The raw EEG is difficult to interpret and various transforms and processed displays have been developed to aid clinicians. Classical observation has correlated behavior and frequency activity within broad bands: these cover 0.5Hz to 50Hz and have been labeled delta, theta, alpha, beta and beta 2 (in ascending frequency order).

The EEG can be considered to consist of an underlying background process (assumed stationary and ergodic on several segments lasting several seconds) with superimposed transient nonstationarities (e.g. eye-blinks, patient movement or electrode movement) and superimposed ‘unwanted’ continuous signals like electrocardiogram (ECG), electromyogram (EMG) or an AC power signal. Features include synchrony, spikes (sharp epileptiform waves) and burst suppression (brief periods of very low amplitude). Higher processes (conscious thought) are characterized by desynchronisation.

Anaesthesia typically – but not always – causes synchronisation and slowing (shift to lower frequencies) of the EEG. Deep anaesthesia is characterized by burst suppression and ultimately electrical silence.

Frequency Domain

The possibility of using the EEG for measuring DOA has been extensively explored during the last two decades. This was made feasible by the introduction of computer-based signal processing methods. Part of the difficulty in evaluating the raw EEG was addressed by using different representations especially in the frequency domain. The Fast Fourier Transform (FFT) is a method which transforms data from the time-domain into the frequency-domain. Display methods such as the Compressed Spectral Array (CSA) [58, 7, 6] and Density Spectral Array (DSA) [58, 25] are suitable for clinicians wishing to monitor the amplitude spectra obtained by FFT. Despite the specificity of the changes of EEG frequency amplitudes to anaesthetic agents [58][1],[102] and the clinical state of the patients [21], visualizing the EEG signal in a compressed form by CSA or DSA plots gives the anaesthetist the opportunity to complement traditional clinical signs with on-line EEG signal monitoring [135, 7, 6].

Indices derived from EEG spectral frequency changes are the spectral edge frequency (SEF), median frequency (MF), total power spectra and frequency band power ratio. These have been extensively investigated for DOA monitoring [20, 30, 68, 123, 144, 29]. In general, these measures seem to be less capable of measuring DOA due to their sensitivity to the anaesthetic drugs used and the variability of the results between different researchers [30]. A disadvantage of MF and SEF is hysteresis – the measure lagging behind changes of measured anaesthetic concentration.

[1] The effects on the EEG of some of the agents are summarized in this publication.

The spectral representation by autoregressive model coefficients (AR) was investigated in [124, 136]. Although the AR representation can lead to a better estimation of the spectral properties of the EEG we may conjecture that this method is unlikely to avoid the problems encountered by conventional spectral analysis following FFT.

Bispectral Index

If we accept the theory that the EEG is generated by nonlinear processes, the above mentioned spectral measures then have to fail to fully describe the EEG, because in the conventional power spectral EEG analysis, the phase information, which describe nonlinear connections between the elementary process of the EEG, is suppressed. Bispectral analysis (BA) [9, 53] used for EEG analysis tries to overcome the problem as BA investigates the phase relations between the fundamental frequencies – bicoherence[2].

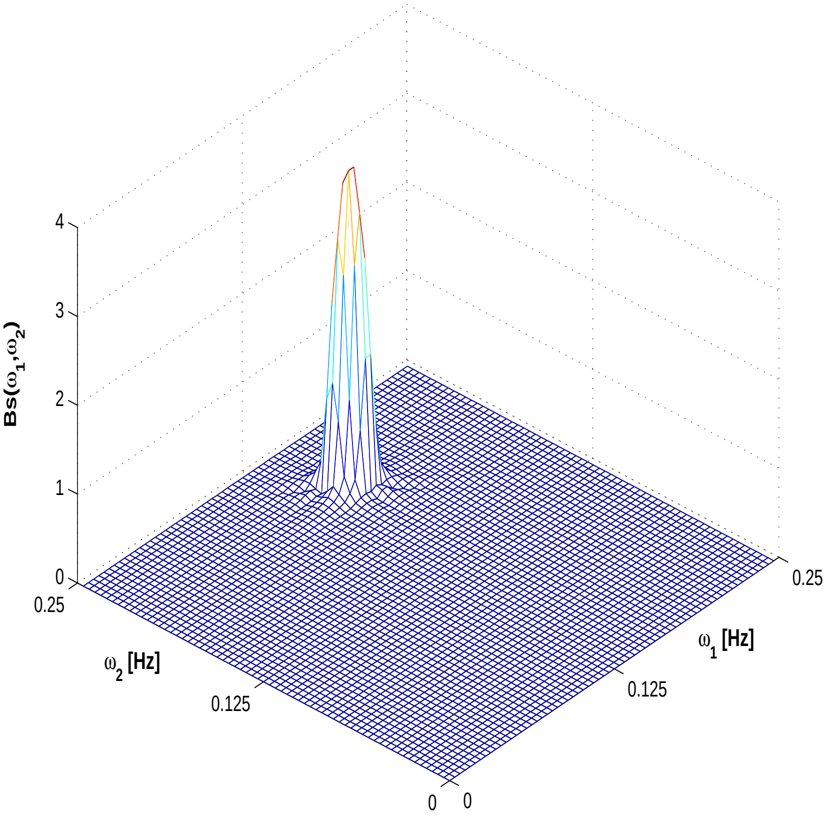

Fig. 7.1: The bispectrum computed for the signal linearly mixed from sines waves with frequencies ω_1, ω_2 and $\omega_3 = \omega_1 + \omega_2$. The phases ϕ_1 and ϕ_2 were randomly generated and the phase for third harmonic was set $\phi_1+\phi_2$. The peak indicates high magnitude of bispectrum $\mathrm{Bs}(\omega_1, \omega_2) = |f(\omega_1)f(\omega_2)f(\omega_1+\omega_2)|$ and reflects a strong frequency and phase relation between ω_1 and ω_2 and modulation frequency $\omega_1+\omega_1$. $f(.)$ represents complex spectral values given by the FFT of the original signal. Bicoherence between two fundamental frequencies ω_1 and ω_2 is then defined as $\mathrm{Bc}(\omega_1, \omega_2) = \mathrm{Bs}(\omega_1, \omega_2)/\sqrt{|f(\omega_1)|^2|f(\omega_1)|^2|f(\omega_1+\omega_2)|^2}$.

Combining *burst suppression ratio*, *relative alpha/beta ratio* and *bicoherence* between individual frequencies the bispectral index (BIS) was introduced as a measure of DOA based on a multivariate regression on a huge database of patients [63, 70]. BIS appeared to generalize well across a number of general anaesthetic agents, and a significant amount of clinical research in the use of BIS as a DOA measure has been carried out (see Aspect Medical Systems

[2] The bicoherence is computed by normalizing the bispectrum which reflects frequency and phase coupling between the individual fundamental frequencies making up the investigated signal – Figure 7.1

Inc.[3] for references regarding application of BIS). A comparison of the MF, SEF, BIS and an auditory evoked potentials (AEP) index (described below) was reported in [30]. It was concluded there, that the BIS and the AEP index distinguish the states of consciousness and unconsciousness with a higher accuracy than MF or SEF.

Complexity Measures

Recently, in [161, 10, 11, 104] a fundamentally different measure reflecting changes of the complexity of EEG during the different stages of anaesthesia was proposed. The complexity in this context is generally understood to be regularity or predictability of EEG patterns. Hence the periodic repetition of patterns in EEG provide an indication of the deterministic nature of the system generating the signal. Such systems are considered to have lower complexity compared to systems generating fully random signals which are understood to be highly complex. We discuss the concept of complexity in more detail in the next chapter and for the moment we only provide references to several recently reported results using this approach for DOA monitoring problem.

Nonlinear complexity measures – Approximate Entropy (section 8.1.1) and Nonlinear Correlation Index (section 8.2.1) – were used in [161, 10, 11]. Results reported in [10] provided an indication of the potential use of Approximate Entropy in the case of a single drug (desflurane) anaesthesia. In the case of classification of EEG patterns of burst suppression under single drug anaesthesia (sevoflurane or desflurane anaesthesia) the use of these complexity measures resulted in superior performance compared to BIS and frequency based measures [161, 11].

7.1.4 Evoked Potentials Monitoring

Evoked potentials are the responses of the central nervous system to specific stimuli in the form of electrical signals. Three types of stimuli were investigated for DOA monitoring; visual evoked potentials (VEP), somatosensory evoked potentials (SEP) and auditory evoked potentials (AEP). Averaging the EEG responses on repeating stimuli results in subtractive suppression of the ongoing EEG and reveals the evoked response. The requirement of many repetitions of the stimulus results in a time lag in DOA monitoring. In [41, 84], some improvements were proposed to overcome the problem of averaging allowing the use of evoked potentials in a quasi on-line regime.

Auditory evoked potentials

It is clear that from a practical point of view it is mainly AEP which can be used for DOA monitoring. In our study we did not have the opportunity to collect this type of data which would have required further specialized equipment and resulted in potential interference in the surgical process. However, for completeness we provide a short review of published work in this domain. The AEP can be divided into three time stages.

Brainstem auditory evoked potential (BAEP) are EEG wave responses with a latency shorter than 8-9ms which correspond to the brainstem response. In [137] the difference between the changes of the BAEP during inhalational anaesthesia (increase in the latency) and intravenous anaesthesia (little or no effect) was shown. They concluded that brain stem waves were unsuitable for measuring DOA as the effect of intravenous agents was insignificant.

Mid-latency auditory evoked potentials (MLAEP) are EEG responses of the primary auditory cortex with latencies from 8 to 60 msec. Several studies independently suggested that promising results in DOA monitoring were achieved by investigations of the changes

[3] `http://www.aspectms.com/`

in the latencies and amplitudes of the MLAEP [138, 137, 74, 84, 30, 58]. The advantage of using MLAEP changes is their relative independence to a wide range of general anaesthetic agents, however analysis requires an intact auditory pathway and MLAEP are therefore not universally applicable.

Late cortical AEP are responses of the frontal cortex with post-stimulus latencies greater than 50ms. In [138, 137], it was concluded that late cortical AEPs are too sensitive as a quantitative measure of the DOA due to the fact that they are attenuated or abolished by general anaesthesia [51], sedation [48] and sleep [97]. On the other hand, waves between 50 and 100ms may be usefully used to detect the transition from the anaesthetized to the awake state.

7.2 Aim of the Study

The main focus of this study is the investigation of different types of descriptors of depth of anaesthesia applicable to a wide range of anaesthetic drugs which are also insensitive to the patient's physical and psychological status. Assuming brain activity, in common with many biological systems, is a nonlinear dynamical system with irregular and short term predictable characteristics, only measures which reflect nonlinear characteristics would be valuable as descriptors of the EEG.

Recent mathematical results in the theory of nonlinear dynamical systems and systems with deterministic chaotic behavior are a potential source for the extraction of relevant dynamic complexity features of the signal. Thus, to extend the family of DOA descriptors we study different measures of signal complexity applicable to EEG and we investigate the possibility that the measures reflect different stages of DOA using different anaesthetic drugs. We have already mentioned in the previous section some very recent results published during our study which confirmed the promising nature of this approach [161, 10, 11]. As a result of these studies, we also assessed the described measures, comparing them with the measures we considered most suited to assessing DOA.

We hypothesize that convenient signal complexity measures will provide features that will form different clusters reflecting different stages of depth of anaesthesia. These features may act as inputs for a nonlinear classifier or nonlinear regression techniques.

In the next section we describe corrected conditional entropy [99] and coarse entropy rates [87] as new potential measures for DOA monitoring. Other complexity measures used in our experiments are discussed.

8. COMPLEXITY MEASURES

Before we describe the individual complexity measures used, we would like to briefly introduce the concept of complexity, as it is understood in our study, providing the motivation for using this approach to the DOA monitoring problem.

Currently, our knowledge of the genesis of the EEG waveforms during anaesthesia is far from complete, despite the past 20 years of research which have brought new results into this domain. It is mainly the postsynaptical potentials (PSP) of neo-cortical neurons which create the EEG signal as it is measured on the scalp. During the normal awake state the EEG is created by millions of PSP asynchronously firing over the cortex. Visual inspection of the measured EEG traces therefore shows virtually no repetitive patterns. The EEG signal under these conditions displays no deterministic origin making it difficult to detect and track underlying states of the brain. However, the theory of chaotical systems introduced in the 60's [71] gave rise to the question whether it is possible to distinguish between fully random processes and processes with deterministic origin but showing high levels of irregularity and usually having low levels of predictability. Early promising results describing different EEG stages by several absolute values characterizing a low-dimensional chaotical system generating the EEG often had to be re-examined. This was mainly due to the effects of the relatively high noise component present in EEG, nonstationarity and other factors violating the conditions required for the estimation of the desired descriptors of a chaotical system. Also, hypotheses suggesting a low-dimensional origin of the EEG signal were in many cases re-considered. In spite of these facts, many of these results, together with the results reported when classical linear (time or frequency domain) descriptors were used, permitted the extraction of characteristics of the EEG signal which quantitatively appeared to distinguish between different brain states.

The situation is dramatically altered when an anaesthetic drug is administered, or during sleep. The neuronal activity of the thalamus then produces an oscillating activity which synchronizes firing of neo-cortical neurons, coinciding with a decrease in the overall excitability of neo-cortical neurons. This leads to the EEG trace showing more regular behavior with dominant frequencies significantly shifted to lower frequency activities falling into the delta band. These changes are usually evident on simple visual inspection of the EEG. Increasing the concentration of anaesthetic drug produces a further decrease of excitability of neo-cortical neurons leading ultimately to the state of deep anaesthesia or coma during which neo-cortical neurons become inactive.

This suggests that measures reflecting changes in regularity and predictability of EEG patterns associated with the transition from awake stages to the stages of anaesthetized or deeply anaesthetized patients may serve as valuable indicators of DOA. Thus our concept of EEG complexity is based on the consideration that irregular, minimally predictable patterns in the EEG are associated with a system of high complexity generating the EEG while regular, more predictable traces of EEG are considered to be less complex. Measures associated with this concept of complexity are discussed in the next section.

A slightly different approach is used in the case of the Nonlinear Correlation Index discussed in section (8.2.1). Although this measure is derived from the correlation dimension of an attractor of a low dimensional chaotic system it is not understood as a measure of proper dimension. Further, it was suggested in [61, 91, 161] that, in situations where we cannot assume a low dimensional chaotical system generating the observed EEG signal, the measures derived in the context of chaos may be used to discriminate the differences between data recorded under different experimental conditions. This suggestion must be treated with caution. Higher values of the Nonlinear Correlation Index similar to the correlation dimension are associated with systems with a higher number of degrees of freedom and reflect more complex awake or light anaesthesia stages of the patients.

8.1 Entropy Rates

Entropy rates are measures designed to quantify regularity of a time series or predictability of the new samples based on previous observations. Consider a time series represented by samples from a complex dynamic process evolving in time. The complexity of such a process can be evaluated in terms of how quickly the system loses information about previous states. Entropy rates will tend to zero values for processes with a periodic repetition of the same pattern and conversely will lead to high values for processes with aperiodic or random behavior. On the other hand, for a dynamical system evolving in some measurable state space the entropy rates are related to Kolmogorov-Sinai entropy (KSE) [61]. This connection allows one to apply entropy rates not only when a general linear or nonlinear stationary stochastic process is assumed but also when an observed time-series is considered as a projection of a trajectory of a dynamical system (e.g. low-dimensional chaotic system).

Consider a discrete stochastic process $\{X_i\}$, i.e. an indexed sequence of discrete random variables characterized by the joint probability distribution function $p^m(x_1, \cdots, x_m) = \Pr[(X_1, \ldots, X_m) = (x_1, \ldots, x_m)]$, where $x_i \in \mathcal{X}$ to be a realization of X_i drawn from the set of all possible values $\mathcal{X}$ (alphabet). Then we have the following definition of the *entropy rate* [15]

Definition: The entropy rate of a stochastic process $\{X_i\}$ is defined by

$$\lim_{m\to\infty} \frac{H(X_1, X_2, \ldots, X_m)}{m} = h, \tag{8.1}$$

where

$$H(X_1, X_2, \ldots X_m) = -\sum_{x_1,\cdots,x_m} p^m(x_1, \cdots, x_m) \ln p^m(x_1, \cdots, x_m) \tag{8.2}$$

is the entropy of the random vector $\mathbf{X}_m = (X_1, \ldots, X_m)$.

The entropy rate is a quantity giving the *average amount of uncertainty per one random variable.* An alternative way to express limit (8.1) follows from the fact that for stationary random processes we also have (for the proof see [15])

$$\lim_{m\to\infty} H(X_m/X_1, \ldots, X_{m-1}) = h, \tag{8.3}$$

where the conditional entropy $H(X_m/X_1, \cdots, X_{m-1})$ is defined by the relation

$$H(X_m/X_1, \ldots, X_{m-1}) = -\sum_{x_1,\ldots,x_m} p^m(x_1, \ldots, x_m) \ln p^m(x_m/x_1, \ldots, x_{m-1}) \tag{8.4}$$

in which

$$p^m(x_m/x_1, \ldots, x_{m-1}) = \frac{p^m(x_1, \ldots, x_m)}{p^{m-1}(x_1, \ldots, x_{m-1})} \tag{8.5}$$

is the conditional probability distribution of $X_m = x_m$ given $x_1, \ldots, x_{m-1}$. Existence of limits (8.1) and (8.3) belongs to basic properties of a stationary process.

In practical situations we usually assume the time series $\{x(t), t = 1, 2, \ldots, N\}$; i.e the observed measurements at time points t, to be a realization of a stationary and ergodic stochastic process $\{X_i\}$. The ergodicity condition allows us to estimate the statistics of $\{X_i\}$ from a single realization (time series); i.e we can replace ensemble averages by equal time averages. Thus, we can assume the variables X_i to be

$$X_i = x(t + (i-1)\tau),$$

where τ is a time delay. But, even in such a case, in practice we can not compute exact entropy rates from a finite number of measurements. The exact estimate of entropy rates is restricted only to some specific cases [15]. If the observed time-series is a realization of a zero-mean stationary Gaussian process we can estimate entropy rates through its spectral density function $f(\omega)$. This is given by the fact that a stationary Gaussian process can be fully described by its spectrum. In such a case the entropy rate can be expressed up to the constant term as [54, 3, 90]

$$h_{GP} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln f(\omega) d\omega. \tag{8.6}$$

Now, consider a time series $\{x(t)\}$ to be a finite projection of a dynamical system evolving in some continuous measurable state space. Using the time delays technique [85, 133] we can construct an m dimensional embedding vector $\mathbf{x}_m(j) = [x(j), x(j-\tau), \dots, x(j-(m-1)\tau)]^T$. However, now we assume each vector $\mathbf{x}_m(j)$ to be a sample of m variables $X_1, \dots, X_m$ determined by the dimensionality of the evolving dynamical system. We can define the joint probability distribution $p^m(x_1, \cdots, x_m) = \Pr[(X_1, \dots, X_m) = (x_1, \dots, x_m)]$ and consider (8.1) as the KSE of a dynamical system [95]. This connection is given by the fact that each measure preserving dynamical system[1] corresponds to a stationary stochastic process and vice versa [95, 89]. KSE is a topological invariant related to the sum of positive Lyapunov exponents [94] and may be of use as an appropriate measure to characterize dynamical systems and their states. Lyapunov exponents are quantities characterizing the strength of chaos in terms of the rate of divergence of trajectories of a dynamical system [61]. Higher values indicate fast exponential divergence of the close trajectories over the course of time and the loss of predictability despite a deterministic origin of the system.

Again, a difficulty arises when exact KSE has to be estimated from a finite number of observations usually containing a relatively high noise component. Thus, in many practical situations alternative approximations of KSE have to be used.

In the next subsections we present three possible algorithms constructed with the aim of approximating KSE or entropy rates where a stationary stochastic process generating the observed data is assumed. However, rather than estimate exact values of KSE or entropy rates the algorithms are used to measure regularity and predictability of time series with the aim of distinguishing different states and/or determining the characteristics of the dynamical systems or stochastic processes generating the observed data.

8.1.1 Approximate entropy

The concept of approximate entropy (ApEn) was proposed by Pincus et. al [98]. Assuming an observed time-series of length N from which the set of m-dimensional vectors $\{\mathbf{x}_m(j) = [x(j), x(j-1), \dots, x(j-(m-1))]\}_{j=1}^{N-m+1}$ was constructed by the previously mentioned time-delay embedding technique we can define ApEn as

$$\text{ApEn} = \Phi^m(r) - \Phi^{m+1}(r), \tag{8.7}$$

where

$$\Phi^m(r) = \frac{1}{N-m+1}\sum_{i=1}^{N-m+1} \ln C_i^m(r), \tag{8.8}$$

$C_i^m(r) = \sum_{j=1}^{N-m+1} \theta(r - \|\mathbf{x}_m(i) - \mathbf{x}_m(j)\|)$ is a correlation sum, θ stands for the Heaviside function

$$\theta(u) = \begin{cases} 0 & : \quad u < 0 \\ 1 & : \quad u \geq 0 \end{cases}$$

[1] Assume that all states of a dynamical system create a space A and define the probability space $(A, \mathcal{B}, \mu)$ where $\mathcal{B}$ is a σ-algebra of measurable subsets of A and μ is a probability measure such that $\mu(A) = 1$. Further, let $T : A \to A$ be a measurable mapping (i.e. a mapping satisfying $T^{-1}\mathcal{B} = \mathcal{B}$ and $\forall B \in \mathcal{B} : \mu(T^{-1}B) = \mu(B)$) representing the evolution of the dynamical system under study. Then $(A, \mathcal{B}, \mu, T)$ is called the *measure preserving transformation system.*

$\|.\|$ represents a norm in a phase space of embedded vectors (usually the maximum norm is utilized) and the parameter r is the diameter of the phase space partition (grain). Rather than giving the rigorous mathematical description of ApEn let us provide a more heuristic insight into the definition of ApEn. ApEn measures the (logarithmic) probability that m dimensional patterns that are close to each other will also stay close when their dimension increases. The frequency of the m dimensional patterns similar to the pattern $\mathbf{x}_m(i)$ is measured by the correlation sum $C_i^m(r)$ where the parameter r defines the level of proximity of the patterns. The decrease of $C_i^{m+1}(r)$ in relation to $C_i^m(r)$ indicates the 'diversity' of the patterns compared with the pattern $\mathbf{x}_m(i)$ when the length of the pattern is increased to $m+1$ and therefore leads to the increase of ApEn. It is clear, that for data generated randomly without any regular structure ApEn will tend to higher values. In contrast, for a rigorously periodic pattern $\mathbf{x}_m(i)$, whose period can be 'captured' by the m dimensional embedding vector, the increase of the embedding dimension to $m+1$ will not decrease $C_i^{m+1}(r)$ and ApEn will tend to zero. The final ApEn value is then taken as the average over 'regularity' of all possible patterns $\mathbf{x}(i)$.

Several aspects of ApEn are now summarized. First, in the case of noise-free data there is a clear relation between ApEn and KSE which (under some assumptions) can be estimated from a time series as [39, 61]

$$h_{KS} = \frac{1}{N-m+1} \lim_{r \to 0} \lim_{m \to \infty} \lim_{N \to \infty} \sum_{i=1}^{N-m+1} \ln \frac{C_i^m(r)}{C_i^{m+1}(r)}.$$

The estimate of KSE in the presence of low-magnitude noise leads to incorrect results [98]. In contrast to KSE, ApEn does adapt the limit process but fixes the m, r parameters. This has several consequences:

- ApEn may be also used in the presence of a significant noise component in the observed data. Noise smaller than the value r is filtered out.
- ApEn is finite. This is in contrast to KSE which for stochastic processes tends to infinity. Thus, ApEn may be also used to distinguish among different stochastic processes.
- ApEn ≥ 0 but is not necessarily zero even for periodic processes.
- ApEn is a relative measure appropriate for discriminating different states of a given system corresponding to different parameter values.
- ApEn is translation and scaling invariant under the condition that the adequate change of the r parameter is provided.

When ApEn is computed from N input data points we need to adjust the parameter m until the correlation sums in (8.8) are correctly estimated. In fact, Pincus et. al. [98] had to restrict the range of embedding dimension to $m \leq 2, 3$ when approximately 1000 data points were used. This may be disadvantageous if the periodic structure of the patterns exceeds these values. In the next subsection we discuss corrected conditional entropy; i.e. another complexity measure closely related to ApEn in which the problem of a possible increase in the embedding dimension m when short data sequences are used was addressed.

8.1.2 Corrected conditional entropy

Similar to section 8.1 we define the conditional entropy as

$$H(X_{m+1}/\mathbf{X}_m) = -\sum_{x_1,\cdots,x_{m+1}} p^{m+1}(x_1,\cdots,x_{m+1}) \mathrm{ln} p^{m+1}(x_{m+1}/x_1,\cdots,x_m),$$

where $p^{m+1}(x_{m+1}/x_1, \ldots, x_m)$ is the conditional probability distribution given by (8.5) and $\mathbf{X}_m$ denotes the random vector $\mathbf{X}_m = (X_1, \ldots, X_m)$. Using the chain rule [15]

$$H(X_m, X_{m-1}, \ldots, X_1) = \sum_{i=1}^{m} H(X_i/X_1, \ldots, X_{i-1})$$

after simple manipulation we can also write

$$H(X_{m+1}/\mathbf{X}_m) = H(\mathbf{X}_{m+1}) - H(\mathbf{X}_m). \tag{8.9}$$

In fact we can see ApEn as an approximation of conditional entropy where correlation sums are used instead of empirical probability distribution functions. Moreover, a similar concept of measuring the regularity of patterns when the dimension of patterns is increased by one can be used here [99]. In the following we denote the estimate of conditional entropy using empirical probability distributions as CEn.

The individual probability distribution functions are in practice estimated by the empirical probability distributions, i.e. by the computation of the frequencies of individual samples. To this end the original time-series is divided into several quantization levels (or bins) Q based on the level of amplitude of the individual points $\{x(j)\}_{j=1}^{N}$. The number of points inside individual bins divided by the overall number of samples defines the empirical probabilities of the points.

Problems occur when the probabilities have to be estimated from a finite (usually small) number of observed points. With increasing dimensionality m and quantization level Q the number of bins will increase as Q^m. CEn will then tend to zero values even in those cases where a clearly random signal (e.g. white noise) is investigated [99]. This effect will also apply to the estimation of ApEn and as we pointed out in the previous subsection we need to restrict ourselves to fixing the embedding dimension to lower values when a small amount of data is available. Hence for higher values of m and Q, Porta et. al. [99] proposed a correction term whenever CEn has to be computed from smaller numbers of observations. Corrected conditional entropy (CCEn) is then defined as

$$\text{CCEn}(X_{m+1}/\mathbf{X}_m) = \text{CEn}(X_{m+1}/\mathbf{X}_m) + perc(X_{m+1})\hat{H}(X_1), \tag{8.10}$$

where $perc(X_{m+1})$ is the percentage of single points in the $m+1$ dimensional phase space (i.e. number of $m+1$ dimensional bins containing one sample) and $\hat{H}(X_1)$ stands for an estimate of the Shannon entropy for $m = 1$. This rather heuristic correction is proposed based on the fact that after finding 100% of single points CEn will tend to zero and we prefer to select randomness; i.e. $H(X_1)$ representing the theoretical value of white noise with the same probability distribution as the investigated time series. Thus we can see that (8.10) consists of two terms, the first decreasing with m whilst the second increasing with m. Finally, CCEn as a function of the parameter m is measured and the minimum is taken as the estimate of the conditional entropy. For further more detailed explanation and experimental evaluation of CCEn we refer the reader to [99].

8.1.3 Coarse-grained entropy rates

Coarse-grained entropy rates (CER) were proposed and successfully used in several applications when complexity or regularity of physiological signals were investigated [87, 92]. To describe CER we will first introduce the term of marginal redundancies, i.e. measures quantifying the average amount of information about the variable X_{m+1} contained in the vector of variables $\mathbf{X}_m = (X_1, \ldots, X_m)$

$$\rho_\tau^{m+1}(X_{m+1}; \mathbf{X}_m) = \sum_{x_1,\ldots,x_{m+1}} p^{m+1}(x_1, \ldots, x_{m+1}) \ln \frac{p^{m+1}(x_1, \ldots, x_{m+1})}{p^m(x_1, \ldots, x_m)p(x_{m+1})},$$

where the subscript τ was used to stress the fact that the marginal redundancies are also functions of the time-delay τ used to construct the m-dimensional embedding vectors. We can also write [89]

$$\rho_\tau^{m+1}(X_{m+1}; \mathbf{X}_m) = H(\mathbf{X}_m) - H(\mathbf{X}_{m+1}) + H(X_{m+1})$$

and see that

$$\rho_\tau^{m+1}(X_{m+1}; \mathbf{X}_m) = -H(X_{m+1}/\mathbf{X}_m) + H(X_{m+1}) \tag{8.11}$$

giving us the relation of CER to CEn and ApEn, respectively.

It has been shown in [27, 100] that for some range of τ parameter there exists an asymptotic relation between marginal redundancies and KSE of the dynamical system

$$\lim_{m\to\infty} \rho_\tau^m(X_m; \mathbf{X}_{m-1}) = H(X_1) - \tau h_{KS}$$

This asymptotic relation provides the possibility of estimating KSE using marginal redundancies [100]; i.e.

$$h_{KS} \approx \lim_{m\to\infty} \frac{\rho_{\tau_1}^m(X_m; \mathbf{X}_{m-1}) - \rho_{\tau_2}^m(X_m; \mathbf{X}_{m-1})}{\tau_2 - \tau_1}$$

Thus, for a deterministic system we may obtain the estimate of KSE from the slope of the marginal redundancy versus the time delay τ. However, in practice, Paluš proposed computing the CER rather than estimating the exact entropy rates or equivalent KSE of dynamical systems [87, 92]. He defined CER as

$$h^{(0)} = \frac{\rho_{\tau_0}^{m+1}(X_{m+1}; \mathbf{X}_m) - \rho_{\tau_1}^{m+1}(X_{m+1}; \mathbf{X}_m)}{\tau_1 - \tau_0}$$

or alternatively

$$h^{(1)} = \frac{\rho_{\tau_0}^{m+1}(X_{m+1}; \mathbf{X}_m) - \|\rho^{m+1}\|}{\|\rho^{m+1}\|}\,; \qquad \|\rho^{m+1}\| = \frac{\sum_{\tau=\tau_0}^{\tau_{max}} \rho_\tau^{m+1}(X_{m+1}; \mathbf{X}_m)}{\tau_{max} - \tau_0}.$$

Paluš[2] suggests setting τ_{max} to the value that for $\tau \geq \tau_{max} : \rho_\tau^{m+1}(X_{m+1}; \mathbf{X}_m) \approx 0$ and also setting τ_0 to zero. On several EEG epochs we observed that $\rho_\tau^{m+1}(X_{m+1}; \mathbf{X}_m)$ (for different m and Q values) tends to zero for the value $\tau_{max} \approx 150$ and we therefore used this value in all our future analyses. By setting $\tau_0 = 0$ it is easy to see that $h^{(0)}$ is in fact identical to the estimate of CEn (Appendix B.1). In the following, when we refer to CER the $h^{(1)}$ measure is considered.

The fact that we do not estimate limit values providing the estimates of exact entropy rates means that CER are similar to ApEn relative measures of regularity and predictability of the investigated systems. Higher values of CER indicate less predictable and more irregular behavior of the underlying processes.

8.1.4 Quantization effect

The algorithms discussed in the last two subsections assume the knowledge of joint and/or conditional probability distribution functions of individual constructed embedding vectors. In our study these probabilities were estimated based on the computation of empirical probability distribution functions using the technique of histograms. Consider that the observed time series may have K distinct values. Then we may divide the values of time series $\{x(t)\}$ into Q categories (bins) for any Q in the range $1 \leq Q \leq K$. However, as we already noted in subsection 8.1.2 the number of bins increases as Q^m and when Q and m increases without

[2] Paluš also defines CER as $\rho_{\tau_0}^{m+1}(X_{m+1}; \mathbf{X}_m) - \|\rho^{m+1}\|$, however, it is reported that those estimates which do not depend on absolute values of $\rho_\tau^m(.)$ are more stable and less influenced by the noise component contained in the observed data [87].

a corresponding increase of data samples many empty bins occur. This may significantly influence the estimates of probability distribution functions resulting in incorrect estimates of entropies or marginal redundancies in the case of CCEn or CER, respectively. In practice we should ensure the number of data points must be a minimum of five times the number of bins [149][3]

$$N \geq 5Q^m$$

to correctly estimate m-dimensional entropies or marginal redundancies. On the other hand for many practical problems we can not presume that a sufficiently long time series will be available. Specifically, in the case of DOA monitoring only on-line or almost on-line systems have practical meaning. Thus an appropriate trade off between both requirements has to be found.

Another problem which arises with the quantization of the original data into Q levels is to define the way in which the data are merged. Consider a decision to quantize data into a predefined number of levels Q. One obvious method is to split the range of the variables into Q equally spaced segments. Although this procedure will roughly preserve the information about the distribution of individual variables it may be unreasonably sensitive to extreme values. Further, in the case of the estimation of m-dimensional entropies or marginal redundancies we are not interested in the distribution of individual variables rather we are interesting in intervariable relations; i.e. the structure of the system. It was discussed in [149] that the quantization of the variables into equally (or almost equally) populated bins – *equiquantization* – may be profitable mainly due to this method being more sensitive to the internal details of the distribution and embodies more information about it. Thus equiquantization seems to be more optimal as it may preserve more structural information in comparison to standard quantization into bins of equal length. The equiquantization will generally provide a more 'dense' m-dimensional histogram in the sense of a lower number of zero bins. Further, the equiquantization approach was successfully applied in the case of CER estimation [87, 92] and also in our former study [105] where the method was used for the estimation of several information-theoretic functionals. In the current study we also applied equiquantization in the case of CCEn estimation. Note, that in this case the correction term in (8.10) is considered to reflect a uniform distribution rather than the actual distribution of the observed time-series. To make the correction term equal to the original proposal [99] we have to compute the estimate $\hat{H}(X_1)$ based on the quantization preserving actual distribution of the time series as much as possible.

We experimentally compared the results when different values of Q, m, N and both quantization approaches were used. However before we report these results we will briefly describe other complexity measures considered in our experiments on the DOA monitoring task.

8.2 Other Complexity Measures

8.2.1 Nonlinear Correlation Index

The correlation dimension (CD) of an attractor is one of the most fundamental quantities of low dimensional chaotic systems that can be computed from a time series [38, 40]. Assuming a low dimensional chaotic origin of brain signals the CD estimated from EEG signals was used to discriminate different dynamical states of the brain. However, spurious results revealed problems in estimating CD from short or nonstationary EEG recordings corrupted by noise, and also gave rise to the question of the validity of assuming a low dimensional, chaotic origin of brain signals (see e.g. [56, 101, 88] and ref. therein). In spite of this fact, there is a belief that measures based on the estimate of the CD may still provide a reasonable discrimination between different dynamical states of the brain when correctly applied to the observed data [61, 161].

[3] For higher values of the Q parameter Paluš suggests an even more strict condition: $N \geq Q^{m+1}$ [86].

In [67, 161, 160] the nonlinear correlation index (NCI) was used to quantify depth of anaesthesia and to predict epileptic seizures, respectively. Although NCI is derived from the algorithm used to estimate CD it does not provide an absolute value estimating exactly the CD, rather it is designed to serve as a more robust, discriminative measure of changes in the dynamical system under study. First the correlation sum

$$C_m(r) = \frac{2}{(N-\Delta n)(N-\Delta n-1)} \sum_{i=1}^{N} \sum_{j=1}^{i-\Delta n} \theta(r - \|\mathbf{x}_m(i) - \mathbf{x}_m(j)\|) \tag{8.12}$$

is computed using m dimensional embedding vectors $\{\mathbf{x}_m(j)\}_{j=1}^{N-(m-1)\tau}$. The term Δn (*Theiler window*) is used to exclude temporal correlations [134]. This makes the correlation sum slightly different from the sum computed in the case of ApEn described in subsection 8.1.1. Next we need to look for the *scaling region*, that is, a range of radius r values where

$$C'_m(r) = \frac{d \log C_m(r)}{d \log r}$$

is constant. For an appropriately chosen embedding dimension m this constant will be the estimate of CD. The authors then compute the averages of $C'_m(r)$ over the range of embedding dimensions $[m_1, m_2]$ and N_r radius values in $[r_l, r_u]$

$$d = \frac{1}{N_r} \sum_{r=r_l}^{r_u} \frac{1}{m_2 - m_1} \sum_{m=m_1}^{m_2} C'_m(r)$$

and define NCI as

$$\mathrm{NCI} = \begin{cases} d & : \text{ if } d \leq D_{max} \text{ and } N_r \geq 5 \\ D_u & : \text{ else} \end{cases}$$

where $D_{max} \approx 2\log_{10} N$ is a maximum resolvable dimension as proposed in [115] and D_u is an arbitrary but fixed threshold value. The upper bound r_u is defined as $C'_1(r_u) > r^*$ where r^* is a small value approaching 1. The authors define the lower bound r_l as

$$r_l = \min_r \{|C'_{m^*}(r_u) - C'_{m^*}(r)| \leq 0.05 C'_{m^*}(r_u) \wedge r < r_u\}$$

where m^* is chosen to be a high embedding dimension. In [67] the authors provide a straightforward heuristic reasoning for the selection of these parameters. In our case, we discuss the actual selection of the τ, r^*, m^*, m_1, m_2 and D_u values in the next chapter.

8.2.2 Spectral Entropy

Spectral Entropy (SpEn) was introduced and defined as Shannon entropy [104]

$$\mathrm{SpEn} = -\sum_{i=1}^{k} p(\omega_i) \mathrm{ln} p(\omega_i)$$

where $p(\omega_i)$ is the probability density function (pdf) value at frequency ω_i. The pdf is obtained by normalization of the power spectral density function given by the Fourier Transform. It is an entropic measure which can be used as a measure of system complexity and is therefore included in this study. However, here the complexity of the system is understood as the number of different processes making up the time series [104]. High SpEn will be due to a large number of processes, while lower values of SpEn will indicate a smaller number of dominating processes creating the time series. Regular, periodic processes with a single dominant frequency will lead to zero values of SpEn whilst random white noise will provide maximum values of SpEn due to a ‘flat’ power spectral density function.

Significant SpEn changes would be grossly visible in graphical displays of the EEG frequency spectrum, and would certainly have been described in early studies of the EEG effects of anaesthesia. SpEn can be seen as a single value measure to quantify these changes. Further, we would like to note that SpEn is a linear measure and its use to fully describe dynamics of a stochastic process is limited to the case of a stationary Gaussian process fully determined by its spectrum. We have already pointed out that in the case of a zero-mean stationary Gaussian processes we can similarly estimate the entropy rate of the process through its power spectral density function (8.6).

9. EXPERIMENTS

The following measures were assessed: Approximate Entropy (ApEn); Conditional Entropy (CEn); Corrected Conditional Entropy (CCEn); Coarse-grained Entropy rates (CER); Gaussian Process Entropy rates (GPER); Spectral Entropy (SpEn); Nonlinear Correlation Index (NCI); Spectral Edge 95 (SEF95) and Bispectral Index (BIS).

Although we might first provide the comparison of these algorithms on more simple examples or artificially generated data we directly 'jump' into the domain of anaesthesiology; i.e. the domain of our interest. This decision is prompted by the fact that it is difficult to simulate EEG data measured under the real surgical conditions and also by the fact that the evaluation of the individual measures on several sets of artificially generated data has already been reported in [87, 90, 99].

9.1 Data Description & Collection

This study used EEG waveform data measured in eight adult patients undergoing routine elective surgery. The patients were all between the ages of 30 and 75 years old, and were ASA I-III (the American Society of Anaesthesiologists widely used scale of fitness for anaesthesia). The surgical specialties represented were orthopaedics, gastro-intestinal surgery, and ophthalmology. The anaesthetic technique was not standardised and including target controlled infusions (TCI) of propofol as well as inhalational anaesthesia. EEG data were tagged with relevant clinical data which included premedicant and sedative medication, hypo or hyperthermia, thyroid disease, target propofol concentration, end-tidal agent concentration, and clinical events such as gagging or eye-opening. The study met the requirements of the local Ethical Committee.

The EEG was recorded continuously with a below-hairline bifrontal montage (Fp1-Fpz, Fp2-Fpz, international 10-20 system, Aspect A-2000 monitor). This has been considered suitable for many studies of anaesthetic depth, and the simple montage reflects the global nature of EEG changes caused by sedation. The raw EEG data were manually cleared of artifacts which were unphysiological, and data identified by the BIS monitor as corrupted were removed. This left approximately 500 minutes of detailed tagged EEG data. The raw EEG was digitized at 128Hz and then filtered between 0.5 - 30Hz. EEG epochs of 1204 data points (8sec) were used for computation of individual measures employed in the study. The epochs were taken in steps of 128 data points (1sec). The BIS values are computed from 60 second EEG segments and averaged over 15sec intervals (Aspect Medical Systems Inc.).

9.2 Results

Although we discuss the selection of parameters of individual measures later in the section, in the case of spectral based measures the 512 point FFT was used. The power spectrum was then computed for the frequency range 0.5-30Hz. Unit delay ($\tau = 1$) was used to construct embedded vectors in the case of ApEn, CEn , CCEn and NCI. In the case of NCI the correlation sum (8.12) and the local slopes of its logarithm were computed using $d2$ function of TISEAN2.1 software package [46]. Similar to [161] in all the cases the Theiler window was set to $\Delta n = 10$, $m^* = 25$ and the range of embedding dimension $[m_1 = 10, m_2 = 25]$ was used. We observed that the values of r^* parameter in the range $[0.8, 0.9]$ provided more stable

	BIS	SE95	SpEn	ApEn	CCEn	eCCEn	CEn	eCEn	CER	eCER	NCI	GPER
BIS	1	0.48	0.52	0.52	0.51	0.51	0.51	0.51	0.51	0.5	0.6	0.52
SE95	0.48	1	0.89	0.95	0.85	0.96	0.84	0.95	0.89	0.9	0.52	0.93
SpEn	0.52	0.89	1	0.93	0.82	0.93	0.82	0.93	0.97	0.97	0.69	0.98
ApEn	0.52	0.95	0.93	1	0.91	0.98	0.91	0.98	0.93	0.93	0.59	0.97
CCEn	0.51	0.85	0.82	0.91	1	0.86	1	0.86	0.82	0.82	0.57	0.86
eCCEn	0.51	0.96	0.93	0.98	0.86	1	0.86	1	0.93	0.95	0.59	0.96
CEn	0.51	0.84	0.82	0.91	1	0.86	1	0.86	0.82	0.82	0.57	0.86
eCEn	0.51	0.95	0.93	0.98	0.86	1	0.86	1	0.93	0.95	0.61	0.95
CER	0.51	0.89	0.97	0.93	0.82	0.93	0.82	0.93	1	0.98	0.68	0.97
eCER	0.5	0.9	0.97	0.93	0.82	0.95	0.82	0.95	0.98	1	0.69	0.96
NCI	0.6	0.52	0.69	0.59	0.57	0.59	0.57	0.61	0.68	0.69	1	0.66
GPER	0.52	0.93	0.98	0.97	0.86	0.96	0.86	0.95	0.97	0.96	0.66	1

Tab. 9.1: Spearman's ranked correlation coefficient computed from individual values averaged over 15sec intervals. EEG data recorded during general anaesthesia. CER($Q = 5, m = 3$), CEn($Q = 5, m = 3$), CCEn($Q = 8, m = 5$), ApEn($r = 0.5\text{SD}, m = 2$), NCI($r^* = 0.9$). Measures prefixed with e were computed using equiquantization.

results. The values of r^* greater than 0.9 usually resulted in $N_r < 5$ on several epochs. The NCI values greater than $D_{max} \approx 6$ were observed only on some awake stage EEG epochs.

We did not observe significant differences between the results computed from two EEG traces recorded. In the following we report the results from the trace on which a smaller number of artifacts was detected.

General assessment of the used measures

Visually comparing measures, in the light of the clinical data, showed some level of correlation with the perceived depth of anaesthesia in all cases. Figure 9.1 shows a typical plot of measures computed over the first 35min of a general anaesthetic.

Existing correlations between individual measures were quantified by Spearman's ranked correlation coefficient (SRC). Selecting sub-parameters for each measure, and varying the averaging of the observed values provided a wide range of different settings for estimating correlation among the measures. However, generally we observed high correlations between the entropy rates measures (SRC > 0.8). We also observed high correlation between particular entropy rates measures when equiquantization or standard quantization was used (SRC > 0.9). Spectral measures (SE95, SpEn and GPER) showed a high correlation with nonlinear entropy rates measures (SRC 0.7 - 0.95). The correlation between NCI and other measures was in some cases lower (SRC 0.6 - 0.9) but still showing significant statistical dependence among the measures. Finally, a slightly lower correlation was usually found between BIS and other measures (SRC 0.5 - 0.9). This may simply be due to mis-synchronization between the recorded BIS values and the measures computed from raw EEG data. As we do not have the exact formula for computation of BIS we cannot reliably answer the question about the level of correlation between BIS and other measures. The example of SRC for the anaesthesia case depicted in Figure 9.1 is provided in Table 9.1.

In the next step, the ability of each measure to discriminate stages of anaesthesia was assessed. This was based on two examples which provided traces with two clinically different stages of anaesthesia and therefore were suitable for quantifying the discriminative power of individual measures. The initial trace was considered as a reference level and individual measures were plotted and investigated as the difference from the mean divided by standard deviation (SD) computed from the reference part. This provides a relative scale in units of standard deviation. During each period, anaesthesia and surgical stimulus were stable, and an ideal measure would contain little fluctuation. Discriminative power is therefore reflected by the difference between the means, expressed in units of reference standard deviation.

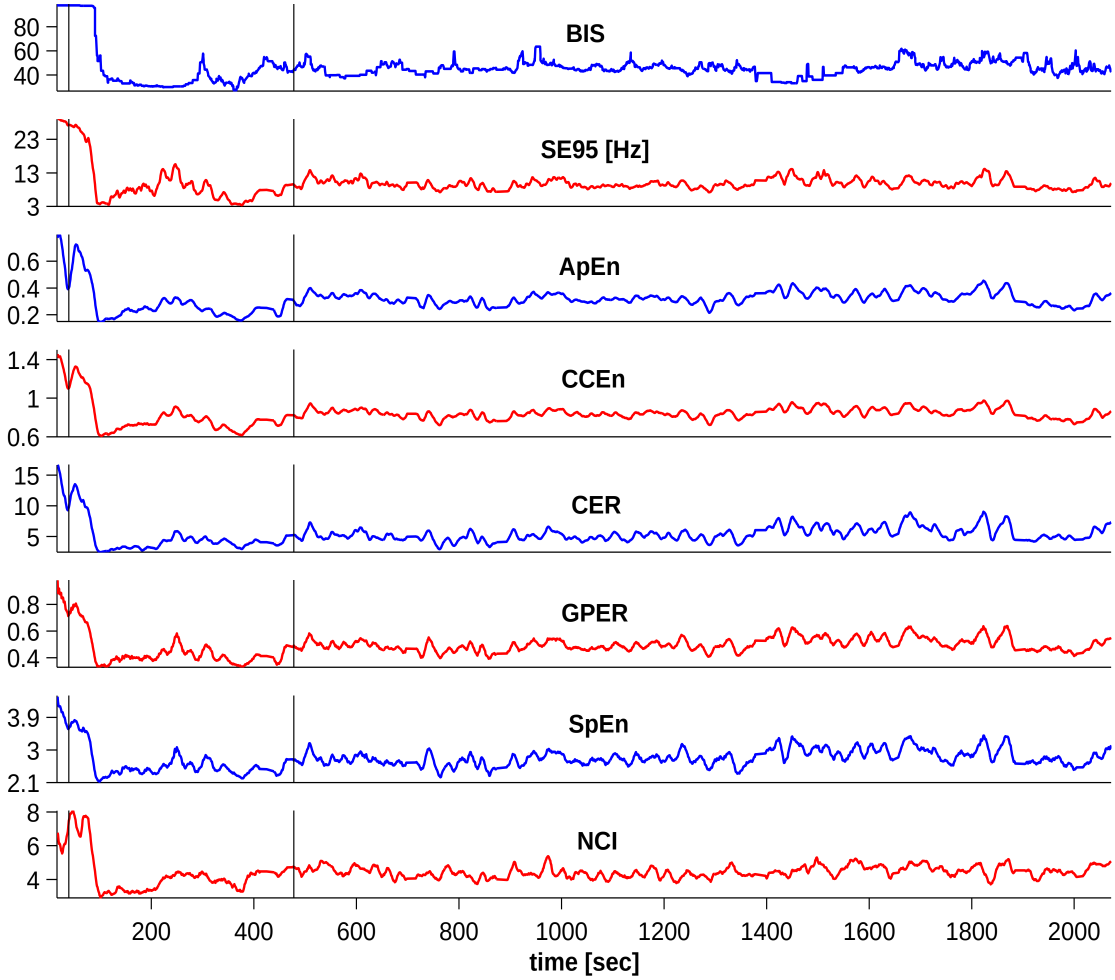

Fig. 9.1: Traces of individual measure over approximately 35min covering induction and maintenance of general anaesthesia. First vertical line: i.v. propofol 2mg/kg. From second vertical line to end: end-tidal desflurane concentration maintained between 3.4% and 4.5%. ApEn($r = 0.5\text{SD}, m = 2$), CCEn($Q = 6, m = 4$), CER($Q = 5, m = 3$), NCI($r^* = 0.9$).

9.2.1 Transition from moderate to light anaesthesia

Each measure's ability to discriminate moderate (propofol level 6mg/L) and light (propofol level 4mg/L) anaesthetic depth is shown in Figure 9.2 where the individual measures are averaged over 15sec intervals (BIS is internally averaged over the same time interval). The graph shows good separation of stages without significant overlapping of the values computed during the individual anaesthesia periods. This data is presented as boxplots in Figures 9.3 and 9.4. These show the lower quartile, median, and upper quartile values and a whisker plot and are plotted in pairs computed over the first period (propofol 6mg/L) (left) and over a second period started 4min after the reaching the desired level of propofol 4mg/L (right). First in Figure 9.3 the boxplots computed from unaveraged values are depicted (BIS is not included as the unaveraged values are not available from the BIS monitor). In all cases overlapping values are seen, however non-overlapping of the upper and lower quartiles indicates a relatively high discriminative power. The boxplots show that in this case entropy rates measures provide higher discrimination between the two levels of anaesthesia with CER performing best. Figure 9.4 shows the same boxplot using 15sec averaged values. Again as in Figure 9.2 no overlaps between the values from different anaesthesia stages indicate high discrimination of all the measures. Although BIS clearly discriminates the two stages we observed paradoxical behavior of BIS in this case and in Figure 9.5 we plotted BIS and

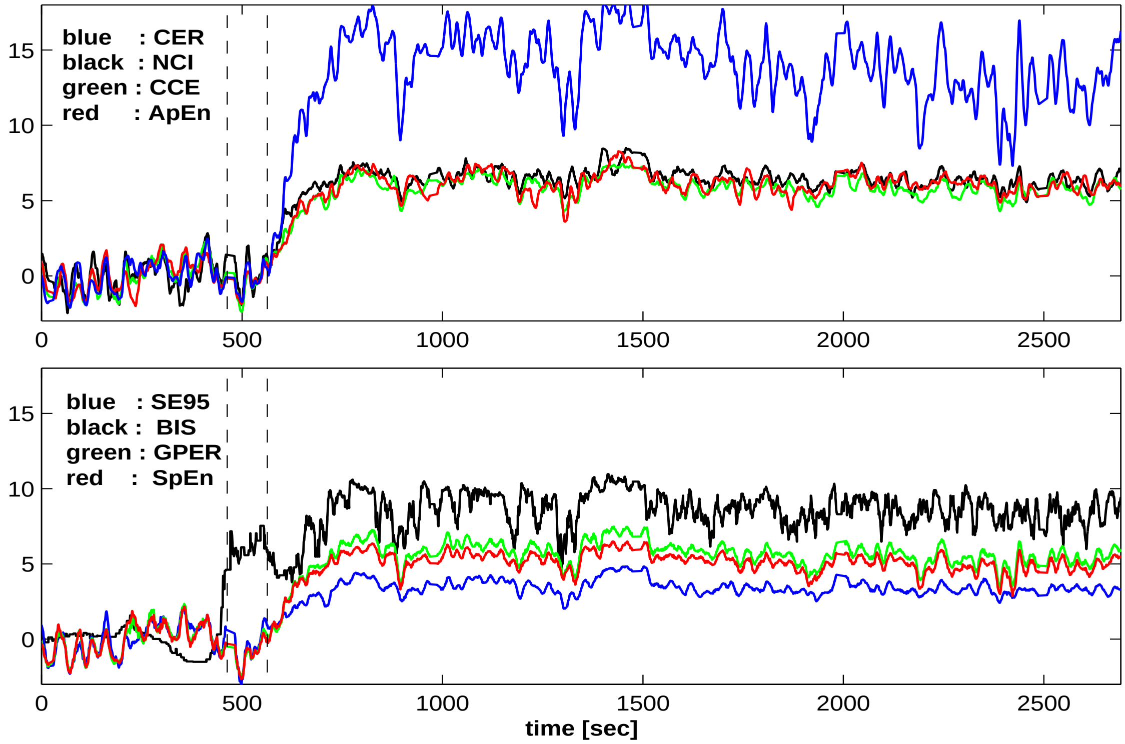

Fig. 9.2: ApEn($r = 0.5\text{SD}, m = 3$), CCEn($Q = 6, m = 4$), CER($Q = 5, m = 3$), NCI($r^* = 0.8$), BIS, GPER, SPEn and SE95 during the transition from moderate to light anaesthesia. The measures are plotted in a relative scale reflecting the difference between the means during moderate and light anaesthesia (propofol levels 6mg/L and 4mg/L respectively) in units of standard deviation (SD) computed during the stage of moderate anaesthesia. Prior to the first vertical line the TCI propofol level was 6mg/L; the first vertical line shows where TCI propofol target was set to 4mg/L; the second vertical line shows the point when 4mg/L was achieved. The values of individual measures were averaged over 15sec intervals. Equiquantization was used in the case of CCEn and CER.

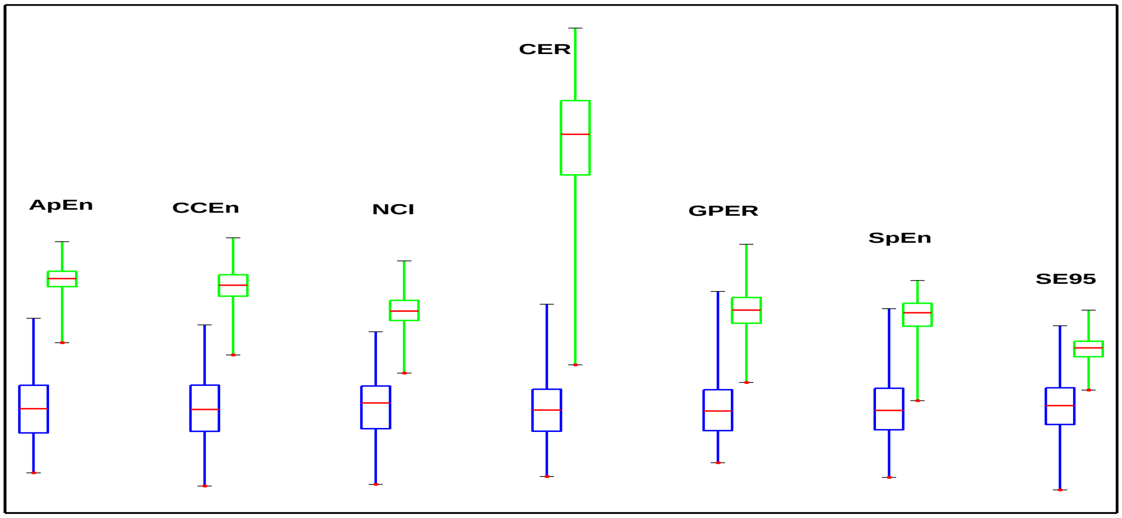

Fig. 9.3: Boxplots of the data presented in Figure 9.2. The left-hand (blue) boxplot of each pair represents a TCI propofol level of 6mg/L, the right-hand (green) boxplot represents a TCI propofol level of 4mg/L. ApEn($r = 0.5\text{SD}, m = 3$), CCEn($Q = 6, m = 4$), CER($Q = 5, m = 3$), NCI($r^* = 0.8$). The values of individual measures were not averaged. Equiquantization was used in the case of CCEn and CER.

Signal Quality Index (SQI)[1] as recorded from the monitor. The SQI changes approximately 6min after the beginning and this might be the cause of the small change in BIS at that time. However the values of SQI are greater than 70% more than 3min before the target level of propofol was decreased from 6mg/L to 4mg/L (first vertical line) the sudden increase of BIS preceding this event therefore cannot be readily explained by a deterioration in signal quality.

9.2.2 Detecting awakening from general anaesthesia

Figure 9.6 shows the measures applied to emergence from anaesthesia. Entropy rates increase progressively as anaesthesia lightens. The increase in BIS is not as dramatic and shows a paradoxical decrease after the stage of eye-opening in response to speech. The spectral measures SE95 and GPER clearly reflect anaesthetic emergence.

9.2.3 Parameters and type of quantization selection

The parameters of all entropy rates measures were tuned using appropriate ranges and compared on the two examples investigated in detail. For measures computed from empirical probabilities (CEn, CCEn, CER), when the restriction to the minimum number of data points was kept in mind, we did not see any significant changes in the discriminative power of the measures. Using the correction term in CCEn allowed us to increase the number of quantization levels Q to 8 when embedding dimension $m = 5$ or $m = 6$ was assumed, however the results were similar to the results provided by CEn using lower values of Q and m parameters (typically $Q = 5, m = 3$). As predicted, increasing the embedding dimension to 5 or 6 degraded the performance of CEn. In all cases equiquantization produced superior results to standard quantization, Figure 9.7.

We observed that in the case of NCI the change of r^* parameter in the range of $[0.75, 0.9]$ did not influence the discrimination power, however, as we discussed above for r^* values greater than 0.9 we observed a higher number of epochs with $N_r < 5$ (in such a case the average of two nearest epochs for which NCI values were determined was taken) and also lower discriminative power of NCI.

[1] The values of SQI between 50%-100% indicate a good quality signal and hence reliable values of BIS. BIS values are not provided if SQI is less than 15% (Aspect Medical Systems Inc.).

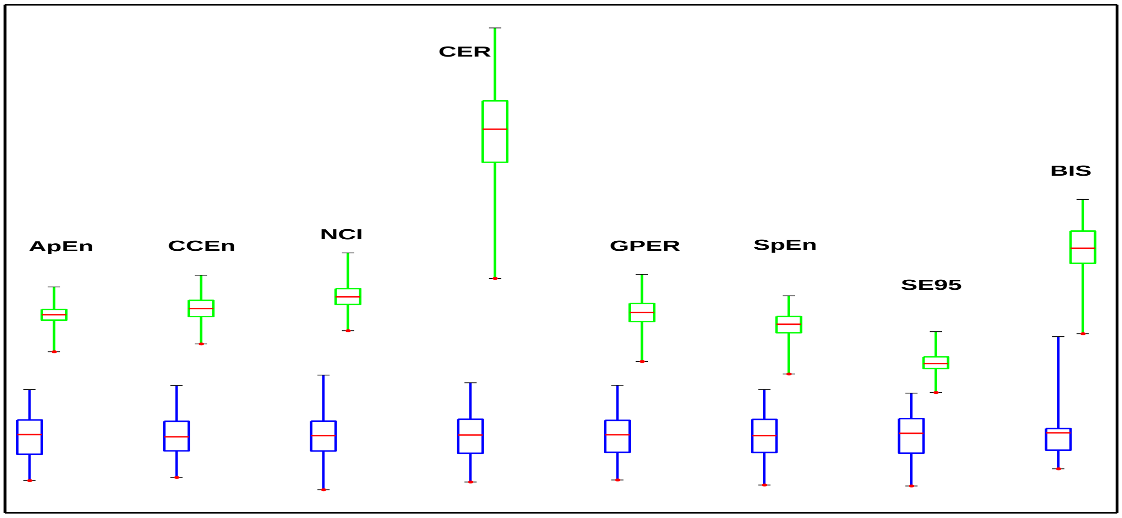

Fig. 9.4: Boxplots of the data presented in Figure 9.2. The left-hand (blue) boxplot of each pair represents a TCI propofol level of 6mg/L, the right-hand (green) boxplot represents a TCI propofol level of 4mg/L. ApEn($r = 0.5\text{SD}, m = 3$), CCEn($Q = 6, m = 4$), CER($Q = 5, m = 3$), NCI($r^* = 0.8$). The values of individual measures were averaged over 15sec intervals. Equiquantization was used in the case of CCEn and CER.

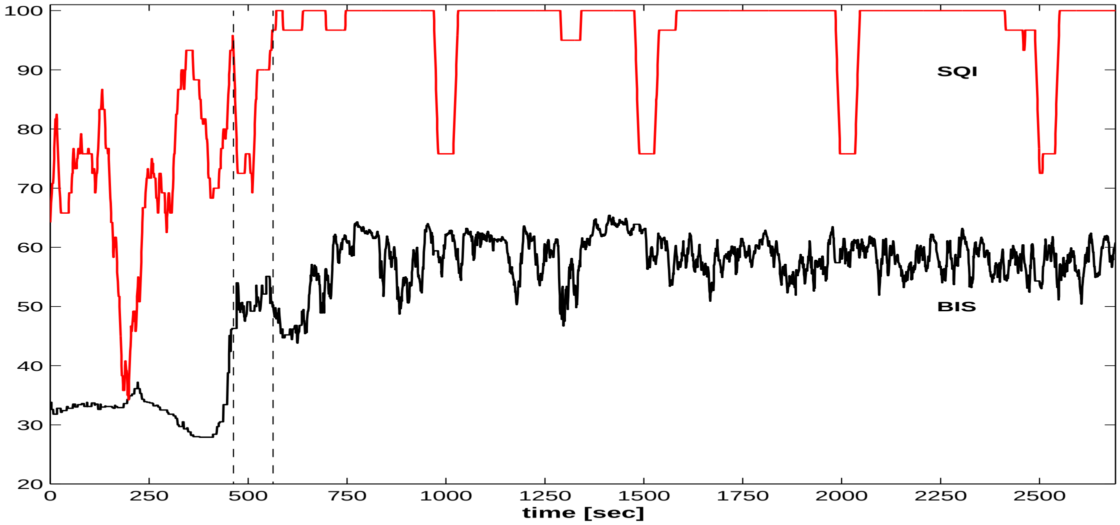

Fig. 9.5: BIS (lower trace) and Signal Quality Index (upper trace) recorded during the period of transition from moderate to light anaesthesia. Prior to the first vertical line the TCI propofol level was 6mg/L; the first vertical line shows where TCI propofol target was set to 4mg/L; the second vertical line shows the point when 4mg/L was achieved.

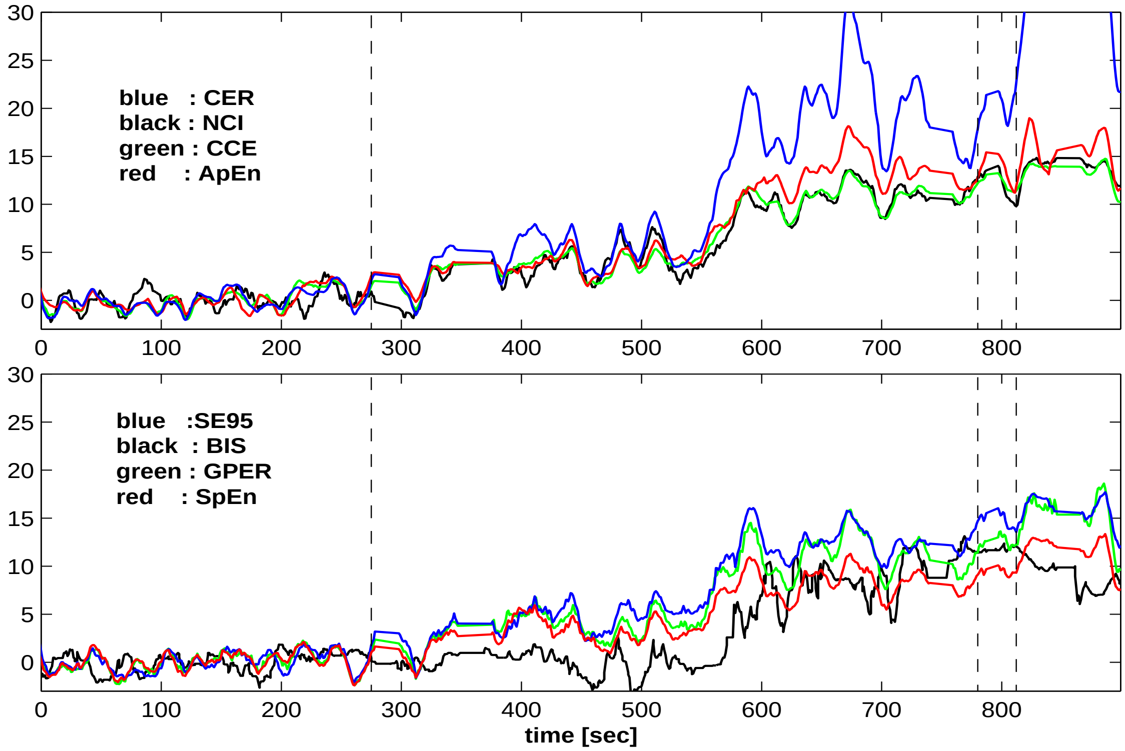

Fig. 9.6: ApEn($r = 0.5$SD$, m = 2$), CCEn($Q = 6, m = 4$), CER($Q = 5, m = 3$), NCI($r^* = 0.8$), BIS, GPER, SpEn and SE95 during emergence from general anaesthesia. The measures are plotted in a relative scale established prior to the first vertical line. The baselines were set to the mean values, and the values then charted in units of standard deviation (SD). The first vertical line indicates when TCI propofol level of 4mg/L was set to 0mg/L. The second vertical line shows when the patient began to gag with the Laryngeal Mask Airway in situ, and the third vertical line denotes eye-opening in response to speech. The values of individual measures were averaged over 15sec intervals. Equiquantization was used in the case of CCEn and CER.

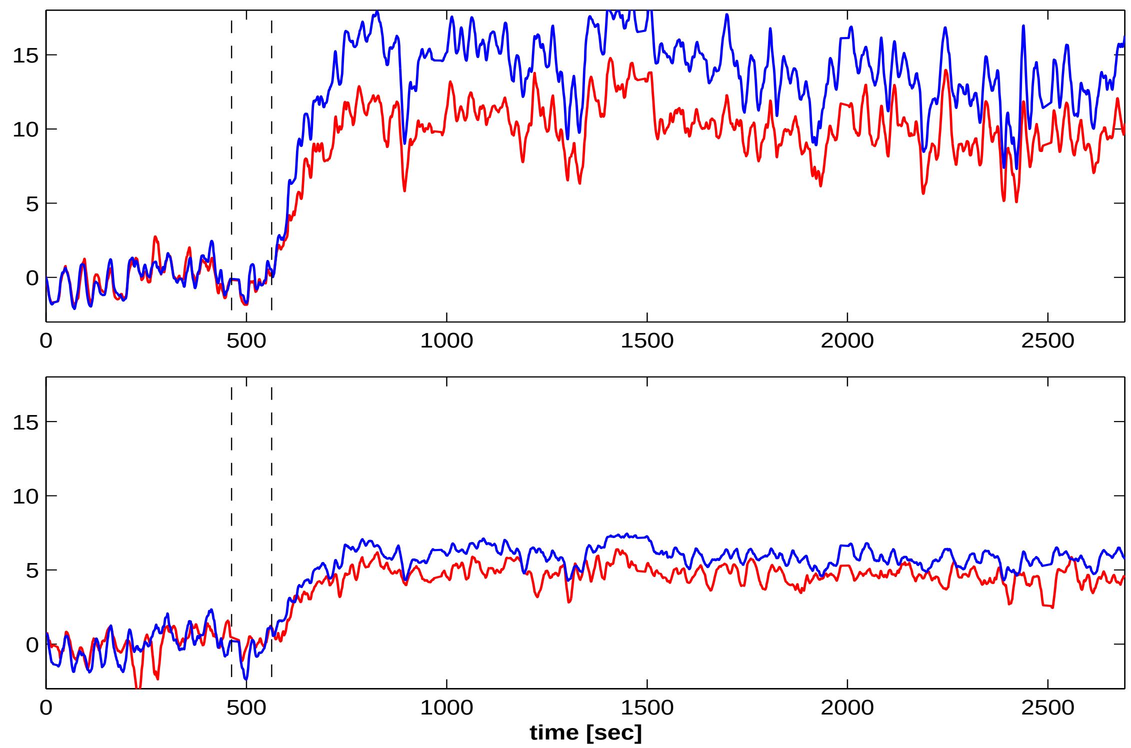

Fig. 9.7: A comparison of two quantization methods – equiquantization (blue) and standard quantization (red) – used to compute CER($Q = 5, m = 3$) (left) and CCEn($Q = 6, m = 4$) (right). The graphs show the transition from moderate to light anaesthesia. For the description of the plotted values and the vertical lines see Figure 9.2.

Finally, in the case of ApEn, assuming a fixed number of data points, we may influence the performance of the measure by varying the sub-parameters grain r and embedding dimension m. As we have already noted in the case of approximately 1000 data points the embedding dimension should not be significantly high, and we confirmed that $m = 2$ or $m = 3$ provided good results. The grain parameter is usually selected in proportion to the standard deviation of the investigated EEG epoch. We also observed that increasing r values resulted in smoother estimates in agreement with the theoretical assumption of inherent noise filtering. High r values, however, may lead to the loss of important system information. We investigated this trade off in the case of transition from moderate to lighter anaesthesia as described above. In Figure 9.8 we plotted the averages of the ApEn values in dependence on the r parameter selected. ApEn values from the second stage (values computed over final 30 minutes were used) referenced to the mean value of the first stage were used. The corresponding variance of ApEn values over the first stage is also depicted. It can be readily seen that values around 0.5SD may provide a good trade off between discriminative power and loss of detailed system information due to the selection of too high values of r.

9.2.4 Surrogate Data Test & Nonlinearity

The results reported on the two TCI cases indicate that the CER, CCEn, ApEn and NCI may provide a better discrimination between the two different stages of anaesthesia monitored during surgery compared to linear measures; i.e. SpEn, SE95 and GPER. However, it is still not clear whether this is due to the nonlinear character of the investigated EEG or simply due to better numerical properties of the nonlinear measures used. In agreement with neurological observations that deeper anaesthesia is characterized by more regular behavior compared to light or awake stages we assume that the transition from deeper to light anaesthesia is also associated with the increase of nonlinearity. Five different surrogate data sets from nonfiltered

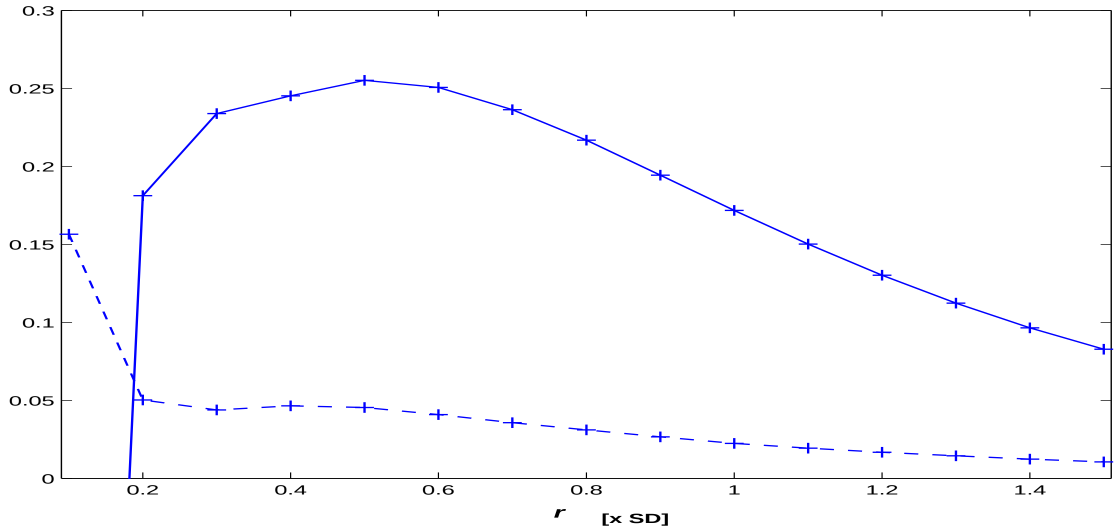

Fig. 9.8: ApEn performance showing the effects of varying the grain parameter r. Source data was from the transition between moderate and light anaesthesia (as Figure 9.2). The solid line shows the difference between the means for the two stages of anaesthesia (TCI propofol 6mg/L; and 4mg/L). Following 6 minutes for equillibration, 30 minutes of data were used to compute stage 2 mean values. The dashed line shows the standard deviation measured during the baseline stage (TCI propofol 6mg/L).

raw EEG were generated. Surrogate data reflects linear properties of the original data (sample autocorrelation, sample amplitude distribution), however, the nonlinear structure is destroyed by phase randomization [122]. The *surrogates* function of the TISEAN2.1 software package was used to generate iterative FFT surrogates [46]. CER and GPER from nonfiltered EEG and surrogate CER (sur-CER) from the stochastic surrogate data were computed. However, rather than investigate the discrimination power of these measures we compared the ratio between the median computed over the reference part and the median computed a) over 30 minutes starting 1min after reaching the desired level of propofol 4mg/L in the case of transition from moderate to light anaesthesia b) over the last 4.5min (the onset of awakening stage) in the case of emergence from anaesthesia. We assume that the higher ratio indicates a better ability of a particular measure to reflect an increase in nonlinearity. In Figure 9.9 we plotted the individual ratios in dependence on averaging interval used to smooth the CER, sur-CER and GPER values. Although we may see a difference between ratios corresponding to CER and GPER there is only a small increase of CER ratios in comparison to sur-CER. Thus, we cannot provide a conclusive answer to the question whether the improved discrimination of CER observed on the two propofol cases is caused by the CER reflecting the nonlinear character of the investigated time-series or simply by better numerical properties of CER compared to spectral based measures.

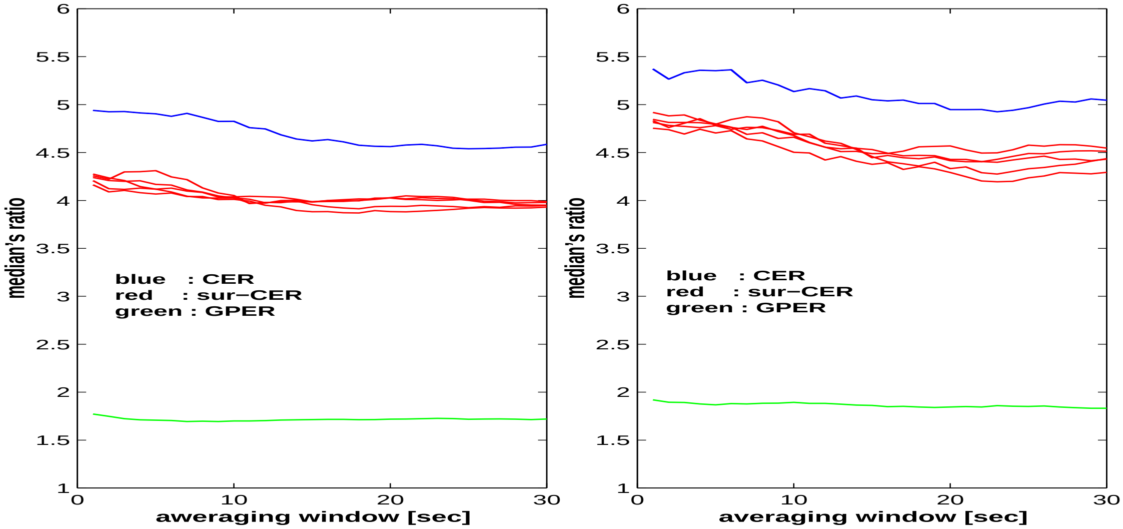

Fig. 9.9: Dependence of ratio between the medians computed from two stages of TCI of propofol anaesthesia on smoothing window used. Five different surrogate data sets were used. CER$(Q = 5, m = 3)$, sur-CER$(Q = 5, m = 3)$. *Left:* Transition from moderate to light anaesthesia. *Right:* Awakening from general anaesthesia.

10. SUMMING UP

Many different measures have been proposed for the purpose of quantifying depth of anaesthesia from analysis of the EEG. We have mainly focused on measures related to the concept of entropy rates estimation and assessed their ability to discriminate different anaesthetic stages. The nonlinear correlation index as an alternative complexity measure was also considered. Particular emphasis was put on each measure's suitability for assessing limited amounts of noisy data; i.e. EEG traces recorded in the typical operating theatre environment. These nonlinear complexity measures were compared with established and widely used spectral measures BIS and SE95. SpEn was considered as a measure quantifying changes in the frequency domain which correspond to varying EEG data characteristics. We pointed out the close relationship between SpEn and entropy rates computed from the periodogram when a stationary Gaussian process is considered to generate the EEG. Assessing any nonlinear component in the EEG would however require one of the other techniques studied here.

The presented work has confirmed nonlinear complexity measures to be useful for determining depth of anaesthesia, and has shown that different complexity measures differ in their ability to identify levels of light anaesthesia. These differences have been explored in some depth in order to identify the factors involved and their significance. The limited size of this study does not permit definite and detailed conclusions about the applicability of these techniques to the wide range of conditions encountered in clinical practice.

Physiological justification

During the transition from awake to anaesthetised states the EEG becomes increasingly influenced by thalamic oscillators, and this synchronisation may provide a physiological relevance for measuring EEG entropy rates. The function of these oscillators is unknown, but EEG regularity and predictability may change as a result of this process and entropy rates can be used to measure these changes. Entropy rates tend to zero for processes with periodic repetition and conversely tend to high values for processes with aperiodic or random behavior.

One estimate of EEG entropy rates (ApEn) has been shown to decrease progressively as anaesthesia deepens [10]. It has also been demonstrated that ApEn applied to burst suppressed data showed a low entropy rate – tempting speculation that entropy rates might be a unified measure of anaesthetic depth [11]. We considered whether other estimates of entropy rates might be more suited to measuring depth of anaesthesia.

Quantifying differences between complexity measures

Of the seven measures assessed, four had been previously applied to physiological data, but not to EEG data recorded during anaesthesia (CEn, CCEn, CER, GPER); the other three had been investigated individually as measures of anaesthetic depth (ApEn, NCI, SpEn). Although differences between these measures were obvious on visual inspection of the traces, quantifying these differences is difficult since there is no benchmark gold-standard measure that can be applied where the anaesthetic technique is not standardized. Unfortunately standardizing anaesthetic technique then provides no information for variability due to differences in technique. We therefore quantified the differences between these measures using traces where the only varying factor was the blood propofol concentration. This is a circumstance where BIS is known to perform reliably, and would provide a useful comparison. All the traces studied reflected moderate to light depth of anaesthesia – less than 3 percent showed burst-suppression.

Similarity and discriminative power of complexity measures

Although the individual complexity measures are derived from similar theoretical assumptions they have different numerical properties. A high level of correlation among the measures was observed suggesting similar behavior of the measures. This correlation was seen across all the cases studied here, despite no standardization of the anaesthetic technique.

The detailed results reported here indicate that the investigated nonlinear complexity measures CCEn, CEn, CER, ApEn and NCI may provide better discrimination between two different stages of anaesthesia than spectral measures (SpEn, SE95 and GPER). This is not necessarily due to some nonlinear component of the EEG but may simply reflect better numerical properties of the measures. Other researchers have also reported a significant difference in discriminative power between GPER and CER when used to assess the pre/ictal EEG [92]. However when EEG surrogate data were used to compute CER a smaller difference was observed. Regarding the numerical properties of individual methods, it has been suggested [92] that the superior performance of nonlinear entropy rates measures is based on a filtering element which can be identified in all these methods (quantization into Q levels in the case of CEn, CCEn and CER and the selection of grain parameter r in the case of ApEn).

This is in contrast to the measures derived from frequency characteristics where noise present in the original recordings of EEG is necessarily incorporated in the final estimates. Further confirmation of this source of error is that 'optimal' values for the r parameter in ApEn are higher (0.5SD) than the recommended values 0.1-0.25SD [98]. Similarly smaller values of the r^* parameter (0.8-0.9) in the case of NCI in comparison to [160, 161] ($r^* > 0.9$) indicate the selection of a wider scaling region. Our findings suggest that this loss of detailed system information and higher sensitivity to detect spikes or data periods of higher absolute values may not matter when the objective is discrimination between different anaesthetic states.

Finally, the non-stationary nature of the EEG time-series limits measurement to less than 15-30sec windows which may contain significant amounts of noise. This limits the data available for analysis and it would seem appropriate to use the correction term included in CCEn and to use equiquantization rather than standard quantization.

Multiple component strategies

Multiple component strategies have been used. For example BIS uses a proprietary combination of three different measures (burst suppression ratio, relative alpha/beta ratio, and bicoherence between individual frequencies) optimized using multivariate regression on a large clinical database (Aspect Medical Systems Inc.). Combining spectral and nonlinear complexity measures in this way may suggest solutions which combine the relative merits of both these approaches.

Further research

Simple spectral measures are known to provide useful information about anaesthetic depth and are less sensitive to the values of the parameters used. This contrasts with investigated complexity measures which present the user with the problem of selecting appropriate parameter values. For example, the most appropriate parameters for ApEn identified in this study differed from those suggested by Bruhn et al. [10]. Further and larger clinical studies will be required before definitive statements can be made about the optimum values of these parameters. This study suggests that, in isolation, nonlinear complexity measures provide superior discrimination of anaesthetic depth over spectral measures. It is still unknown how well these measures will generalize across widely differing anaesthetic drug regimens.

11. CONCLUSION

Two different problems of reflecting brain functioning were addressed. This involved human performance monitoring during the signal detection task and depth of anaesthesia monitoring. Before we will discuss common aspects of both parts of the thesis and how the methods investigated in the individual parts may be combined, we first summarize the main results separately.

- New statistical learning theory based on SRM Inductive Principle gave rise to a number of powerful kernel-based algorithms which were experimentally shown to provide in many research areas the same or superior results in comparison to existing techniques. These kernel-based techniques were found very efficient when observed data are mapped to a high dimensional feature space where usually algorithms as simple as their linear counterparts in input space are used. A good example are Support Vectors Classifiers which were shown to provide in many fields superior results to the existing classification techniques (see [82] and refs. therein). Moreover, these algorithms are independent on the dimension of input data sets and thus in the case of physiological measurements usually associated with a high dimensional representation their use may by highly profitable. In the thesis we focused on kernel-based regression techniques. We extended the family of regularized, kernel-based least squares regression models and provided theoretical and experimental comparisons among the models. Promising results on one artificially generated and one real world data sets provided an indication of applicability of the approaches into the domain of physiological data analysis. Finally, we have shown, that the extraction of the nonlinear principal components by Kernel PCA or by the EM approach to Kernel PCA may provide a better structural representation of the investigated ERP.

- We have found that complexity measures investigated in the thesis may be as good or better indicators of depth of anaesthesia in comparison to the existing, mainly spectral based techniques. These findings are even more valuable from the point that during our study several similar observations and results were reported when some of the investigated measures were used on different EEG data sets recorded during the anaesthesia. This opens a new area of more detailed and extensive research into this very important medical problem of depth of anaesthesia monitoring.

Although the individual parts of the thesis were treated in a parallel way we may find further problems where the methodologies described are applicable together. We have already mentioned that in the case of anaesthesia only measures or their combination which are extensively evaluated on a large clinical database may be more widely acceptable by anaesthesiologists and result in new depth of anaesthesia monitors. This will inevitably bring us to the problem of the construction of appropriate classifiers or in the case of continuous response variable to the construction of appropriate regression models. An example of this is BIS where individual measures were optimized using a multivariate regression technique. Thus we may hypothesize that the use of the kernel-based regression techniques discussed in the part A of the thesis may be also profitable here. On the other hand we have demonstrated that the complexity measures presented in the second part of the thesis may provide us features which represent the changes of the investigated EEG traces under different brain states. We believe that these results may be also applicable to the domain of monitoring of human performance under different cognitive tasks where the measures may provide better representation of ERP or EEG recordings. We also believe that the reported theoretical and practical results will be applicable to different areas of physiological data analysis and will attract the attention of different researchers.

Further work

We have partially reported possible future extensions of the individual methodologies in the end of particular parts of the thesis. The author of the thesis will be working with the research group at NASA Ames Research Center and will concentrate on possibilities to improve reliability of the detection of ERP and other electroencephalographic recordings time-locked to specific stimulus or cognitive activity. The proposal for this project is highly motivated by the results achieved on ERP processing reported in the thesis. The application of the studied nonlinear regression techniques with the aim of signal de-noising and the possibility of using a priori knowledge about the desired signal to the construction of more appropriate kernel functions will be investigated. Future proposed collaboration with Dr Alan Hope will be oriented to the confirmation of usefulness of the investigated complexity measures for a wider range of anaesthetic agents, different surgical condition and groups of patients. Finally we believe that this will bring us to the possibility of the combination of both approaches as suggested above.

BIBLIOGRAPHY

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):615–631, 1997.

[2] S. Amari and S. Wu. Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, 12(6):783–789, 1999.

[3] V.V. Anh and K.E. Lunney. Parametric estimation of random fields with long-range dependence. *Mathematical and Computer Modelling*, 21(9):67–77, 1995.

[4] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[5] P. Bartlett, P.M. Long, and R.C. Williamson. Fat-shattering and the learnability of real-valued functions. *Journal of Computer and Systems Sciences*, 53(2):434–452, 1996.

[6] R.G. Bickford, J. Brimm, L. Berger, and M. Aung. Application of compressed spectral array in clinical EEG. In B. Kellaway, I. Petersen, editor, *Automation of Clinical Electroencephalography*, pages 55–64. Raven Press, New York, 1973.

[7] R.G. Bickford, N.I. Flemming, and T.W. Billinger. Compression of the EEG data by isometric power spectral plots. *Electroencephalography and Clinical Neurophysiology*, 31:632, 1971.

[8] C.M. Bishop. *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford, 1995.

[9] D.R. Brillinger. An introduction to polyspectra. *Annals of Mathematical Statistics*, 36:1351–1374, 1965.

[10] J. Bruhn, H. Röpcke, and A. Hoeft. Approximate Entropy as an Electroencephalographic Measure of Anesthesia Drug Effect during Desflurane Anesthesia. *Anesthesiology*, 93(3):715–726, 2000.

[11] J. Bruhn, H. Röpcke, B. Rehberg, T. Bouillon, and A. Hoeft. Electroencephalogram Approximate Entropy Correctly Classifies the Occurrence of Burst Suppression Pattern as Increasing Anesthetic Drug Effect. *Anesthesiology*, 93(4):981–985, 2000.

[12] A. Buja, T. Hastie, and R. Tibshirani. Linear smoothers and additive models. *Annals of Statistics*, 17:453–555, 1989.

[13] C.J.C. Burges. Geometry and invariance in kernel based methods. In B. Schölkopf, C.J.C. Burges, and A.J. Smola, editor, *Advances in Kernel Methods - Support Vector Learning*, pages 89–116. The MIT Press, Cambridge, MA, 1999.

[14] R. Collobert and S. Bengio. SVMTorch: Support Vector Machines for Large-Scale Regression Problems. *Journal of Machine Learning Research*, 1:143–160, 2001.

[15] T.M. Cover and J.A. Thomas. *Elements of Information Theory.* J. Wiley & Sons, New York, 1991.

[16] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines.* Cambridge University Press, 2000.

[17] S. de Jong. SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18:251–263, 1993.

[18] L.M. Delves and J. Walsh. *Numerical Solution of Integral Equations*. Clarendon Press, Oxford, 1974.

[19] A.P. Dempster, N.M. Laird, and D.R. Rubin. Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, series B*, 39:1–38, 1977.

[20] J.C. Drummond, C.A. Brann, D.E. Perkins, and D.E. Wolfe. A comparison of median frequency, spectral edge frequency, a frequency band power ratio, total power and dominance shift in the demonstration of depth of anaesthesia. *Acta Anaesthesiologica Scandinavica*, 35:693–699, 1991.

[21] Editorial. *Lancet*, 1:553–580, 1986.

[22] T. Evgeniou and M. Pontil. On the V_γ dimension for regression in Reproducing Kernel Hilbert Space. Technical Report A.I. Memo No. 1656, Artificial Intelligence Lab, MIT, 1999.

[23] T. Evgeniou, M. Pontil, and T. Poggio. Regularization Networks and Support Vector Machines. *Advances in Computational Mathematics*, 13:1–50, 2000.

[24] R.A. Fisher. Theory of statistical estimation. In *Proceedings of the Cambridge Philosophical Society*, volume 22, pages 700–725, 1925.

[25] R.A. Flemming and N.T. Smith. Density modulation: A technique for the display of three-variable data in patient monitoring. *Anesthesiology*, 50:543–546, 1979.

[26] I.E. Frank and J.H. Friedman. A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, 35(2):109–147, 1993.

[27] A.M. Fraser. Information and entropy in strange attractors. *IEEE Transactions on Information Theory*, 35(2):245–262, 1989.

[28] J.H. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76(376):817–823, 1981.

[29] L. Gaitini, C. Vaida, G. Collins, M. Somri, and E. Sabo. Awareness detection during caesarean section under general anaesthesia using EEG spectrum analysis. *Canadian Journal of Anaesthesia*, 42:377–381, 1995.

[30] R.J. Gajraj, M. Doi, H. Mantzaridis, and G.N.C. Kenny. Analysis of the EEG bispectrum, auditory evoked potentials and the EEG power spectrum during repeated transitions from consciousness to unconsciousness. *British Journal of Anesthesiology*, 80:46–52, 1998.

[31] P.H. Garthwaite. An Interpretation of Partial Least Squares. *Journal of the American Statistical Association*, 89(425):122–127, 1994.

[32] T. Gasser and H.G. Müller. Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics*, 11:171–185, 1985.

[33] F. Girosi. An equivalence between sparse approximation and Support Vector Machines. *Neural Computation*, 10(6):1455–1480, 1998.

[34] F. Girosi, M. Jones, and T. Poggio. Priors, stabilizers and basis functions: From regularization to radial, tensor and additive splines. Technical Report A.I. Memo No. 1430, MIT, 1993.

[35] F. Girosi, M. Jones, and T. Poggio. Regularization Theory and Neural Network Architectures. *Neural Computation*, 7:219–269, 1995.

[36] G.H. Golub and Ch.F. van Loan. *Matrix Computations*. The John Hopkins University Press, London, 1996.

[37] Y. Grandvalet and S. Canu. Outcomes of the Equivalence of Adaptive Ridge with Least Absolute Shrinkage. In M.S. Kearns, S.A. Solla and D.A. Cohn, editor, *Advances in Neural Information Processing Systems 11*, pages 445–451. The MIT Press, 1999.

[38] P. Grassberger and I. Procaccia. Characterization of strange attractors. *Physical Review Letters*, 50:346–349, 1983.

[39] P. Grassberger and I. Procaccia. Estimation of the Kolmogorov entropy from a chaotic signal. *Physical Review A*, 28:2591–2593, 1983.

[40] P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. *Physica D*, 9:189, 1983.

[41] M. Hansson, T. Gansler, and G. Salomonsson. A System for Tracking Changes in Mid-Latency Evoked Potential During anesthesia. *IEEE Transactions on Biomedical Engineering*, 45(3):323–334, 1998.

[42] W. Härdle, P. Hall, and J.S. Marron. How far are automatically chosen regression smoothing parameters from their optimum? *Journal of the American Statistical Association*, 83(401):86–101, 1988.

[43] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall, 1990.

[44] D. Haussler. Convolution kernels on discrete structures. Technical Report UCSC-CRL-99-10, Computer Science Department, University of California, Santa Cruz, 1999.

[45] S. Haykin. *Neural Networks: A comprehensive Foundation*. Prentice-Hall, 2nd edition, 1999.

[46] R. Hegger, H. Kantz, and T. Schreiber. Practical implementation of nonlinear time series methods: The TISEAN package. *Chaos*, 9(2):413–435, 1999.

[47] I.S. Helland. On structure of partial least squares regression. *Communications in Statistics – Elements of Simulation and Computation*, 17:581–607, 1988.

[48] W.M. Herrmann, W. Hofmann, and W. Kubicki. Psychotropic drug induced changes in auditory averaged evoked potentials: results of a double blind trial using an objective fully automated AEP analysis method. *International Journal of Clinical Pharmacology and Therapeutics*, 19:56–62, 1981.

[49] A.E. Hoerl and R.W. Kennard. Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12:69–82, 1970.

[50] A.E. Hoerl and R.W. Kennard. Ridge regression: bias estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

[51] E.C. Hosic and M.I. Mendel. Effects of secobarbital on the late components of the auditory evoked potentials. *Review of Laryngology*, 96:185–191, 1975.

[52] P.J. Huber. *Robust Statistics*. John Wiley and Sons, New York, 1981.

[53] P.J. Huber, B. Kleiner, T. Gasser, and G. Dumermuth. Statistical methods for investigating phase relations in stationary stochastic processes. *IEEE Transactions on Audio and Electroacoustics*, 19:78–86, 1971.

[54] S. Ihara. *Information theory for continuous systems.* World Scientific, Singapore, 1993.

[55] V.V. Ivanov. On linear problems which are not well-posed. *Soviet Mathematical Doklady* (in Russian), pages 981–983, 1962.

[56] J. Theiler and P.E. Rapp. Re-examination of the evidence for low-dimensional, nonlinear structure in human electroencephalogram. *Electroencephalography and Clinical Neurophysiology*, 98:213–222, 1996.

[57] T.S. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In M.S. Kearns, S.A. Solla, and D.A. Cohn, editor, *Advances in Neural Information Processing Systems 11*, pages 487–493, Cambridge, 1999. The MIT Press.

[58] J.L. Jenkinson. The Monitoring of Central Nervous System function in Anaesthesia and Intensive Care. *Current Anaesthesia and Critical Care*, 1:115–121, 1990.

[59] I.T. Jolliffe. A Note on the Use of Principal Components in Regression. *Applied Statistics*, 31:300–302, 1982.

[60] I.T. Jolliffe. *Principal Component Analysis.* Springer-Verlag, New York, 1986.

[61] H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis.* Cambridge Nonlinear Science Series 7. Cambridge University Press, 1997.

[62] M. Kearns and R.E. Shapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and Systems Sciences*, 48(3):464–497, 1994.

[63] L. Kearse, V. Saini, F. deBros, and N. Chamoun. Bispectral analysis of EEG may predict anesthetic depth during narcotic induction (Abstract). *Anesthesiology*, 75:175, 1991.

[64] G. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. *Annals of Mathematical Statistics*, 41:495–502, 1970.

[65] G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33:82–95, 1971.

[66] M. Koska, R. Rosipal, A. König, and L.J. Trejo. Estimation of human signal detection performance from ERPs using feed-forward network model. In *Computer Intensive Methods in Control and Signal Processing, The Curse of Dimensionality.* Birkhauser, Boston, 1997.

[67] K. Lehnertz and C.E. Elger. Can Epileptic Seizures be Predicted? Evidence from Nonlinear Time Series Analysis of Brain Electrical Activity. *Physical Review Letters*, 80(22):5019–5022, 1998.

[68] K. Leslie, D.I. Sessler, M. Schroeder, and K. Walters. Propofol blood concentration and the bispectral index predict suppression of learning during propofol/epidural anesthesia in volunteers. *Anesthesia and Analgesia*, 81:1269–1274, 1995.

[69] P.J. Lewi. Pattern recognition, reflection from a chemometric point of view. *Chemometrics and Intelligent Laboratory Systems*, 28:23–33, 1995.

[70] C.A. Lien, M. Berman, V. Saini, and at. al. The accuracy of the EEG in predicting hemodynamic changes with incision during isoflurane anesthesia (Abstract). *Anesthesia and Analgesia*, 74:187, 1992.

[71] E.N. Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20:130–141, 1963.

[72] D. Mackey and L. Glass. Oscillation and Chaos in Physiological Control Systems. *Science*, 197, 1977.

[73] R. Manne. Analysis of Two Partial-Least-Squares Algorithms for Multivariate Calibration. *Chemometrics and Intelligent Laboratory Systems*, 2:187–197, 1987.

[74] H. Mantzaridis and G.N.C. Kenny. Auditory evoked potential index : a quantitative measure of changes in auditory evoked potentials during general anaesthesia. *Anaesthesia*, 52:1030–1036, 1997.

[75] H. Martens and T. Naes. *Multivariate Calibration*. John Wiley, New York, 1989.

[76] J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions Royal Society London*, A209:415–446, 1909.

[77] S. Mika, B. Schölkopf, A.J. Smola, K.R. Müller, M. Scholz, and G. Rätsch. Kernel pca and de-noising in feature spaces. In M.S. Kearns, S.A. Solla, and D.A. Cohn, editor, *Advances in Neural Information Processing Systems 11*, pages 536–542, Cambridge, 1999. The MIT Press.

[78] G.F. Miller. Fredholm equations of the first kind. In L.M. Delves and J. Walsh, editor, *Numerical Solution of Integral Equations*, pages 175–188. Clarendon Press, Oxford, 1974.

[79] P. Moerland. *Mixture Models for Unsupervised and Supervised Learning*. PhD thesis, École Polytechnique Fédérale de Lausanne, Computer Science Department, 2000.

[80] D.C. Montgomery and E.A. Peck. *Introduction to Linear Regression Analysis*. John Wiley & Sons, 2nd edition, 1992.

[81] S. Mukherjee, E. Osuna, and F. Girosi. Nonlinear Prediction of Chaotic Time Series Using Support Vector Machines. In *Proceedings of IEEE NNSP'97*, 1997.

[82] K.R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. An Introduction to Kernel-Based Learning Algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–201, 2001.

[83] E.A. Nadaraya. On estimating regression. *Theory of Probability and its Application*, 9:141–142, 1964.

[84] A. Nayak and R.J. Roy. Anesthesia Control Using Midlatency Auditory Evoked Potentials. *IEEE Transactions on Biomedical Engineering* , 45(4):409–421, 1998.

[85] N.H. Packard, J.P. Crutchfield, J.D. Farmer, and R.S. Shaw. Geometry from a time series. *Physical Review Letters*, 45:712–716, 1980.

[86] M. Paluš. Identifying and quantifying chaos by using information-theoretic functionals. In A.S. Weigand and N.A. Gershenfeld, editor, *Time Series Prediction: Forecasting the Future and Understanding the Past*, pages 387–413. Addison-Wesley, Reading, Mass, Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. XV edition, 1993.

[87] M. Paluš. Coarse-grained entropy rates for characterization of complex time series. *Physica D*, 96:64–77, 1996.

[88] M. Paluš. Nonlinearity in Normal Human EEG: Cycles, temporal asymmetry, nonstationarity and randomness, not chaos. *Biological Cybernetics*, 75:389–396, 1996.

[89] M. Paluš. Kolmogorov entropy from time series using information-theoretic functionals. *Neural Network World*, 3:269–292, 1997.

[90] M. Paluš. On entropy rates of dynamical systems and Gaussian processes. *Physics Letters A*, 227:301–308, 1997.

[91] M. Paluš. Nonlinear Dynamics in the EEG Analysis: Disappointments and Perspectives. In N. Pradhan, P.E.Rapp, R. Sreenivasan, editor, *Nonlinear Dynamics and Brain Functioning*, pages 201–216. Nova Science, NY, 1999.

[92] M. Paluš, V. Komárek, Z. Hrnčir, and T. Procházka. Is nonlinearity relevant for detecting changes in EEG? *Theory in Biosciences*, 118:179–188, 1999.

[93] A. Pázman. *Nonlinear Statistical Models*, volume 254 of *Mathematics and Its Applications.* Kluwer Academic Publishers, 1993.

[94] Y.B. Pesin. Characteristic Lyapunov exponents and smooth ergodic theory. *Russian Mathematical Surveys*, 32:55–114, 1977.

[95] K. Petersen. *Ergodic theory.* Cambridge University Press, Cambridge, 1983.

[96] D.Z. Phillips. A technique for numerical solution of certain integral equation of the first kind. *Journal of the ACM*, 9:84–96, 1962.

[97] T.W. Picton, S.A. Hillyard, H.I. Kraus, and R. Galambos. Human auditory evoked potentials. I. Evaluation of components. *Electroencephalography and Clinical Neurophysiology*, 36:179–190, 1974.

[98] S.M. Pincus, I.M. Gladstone, and R.A. Ehrenkranz. A Regularity Statistic for Medical Data Analysis. *Journal of Clinical Monitoring*, 7(4):335–345, 1991.

[99] A. Porta, G. Baselli, D. Liberati, N. Montano, C. Cogliati, T. Gnecchi-Ruscone, A. Malliani, and S. Cerutti. Measuring regularity by means of a corrected conditional entropy in sympathetic outflow. *Biological Cybernetics*, 78:71–78, 1998.

[100] D. Prichard and J. Theiler. Generalized redundancies for time series analysis. *Physica D*, 84:476–493, 1995.

[101] W.S. Pritchard, D.W. Duke, and K.K. Krieble. Dimensional analysis of resting human EEG II: surrogate-data testing indicates nonlinearity but not low-dimensional chaos. *Psychophysiology*, 32:486–491, 1995.

[102] R.A.F. Pronk, A.J.R. Simons, R.G.A. Ackerstaff, and E.H.J.F. Boezeman. Intraoperative EEG monitoring. In *Proceedings of the Ninth Annual Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3:1250–1251, 1987.

[103] S. Rännar, F. Lindgren, P. Geladi, and S. Wold. A PLS kernel algorithm for data sets with many variables and fewer objects. Part 1: Theory and algorithm. *Chemometrics and Intelligent Laboratory Systems*, 8:111–125, 1994.

[104] I.A. Rezek and S.J. Roberts. Stochastic Complexity Measures for Physiological Signal Analysis. *IEEE Transactions on Biomedical Engineering*, 44(9):1186–1191, 1998.

[105] R. Rosipal. Analysis of relations in stochastic systems. Master's thesis, DCE, FEE, Czech Technical University, Prague, 1993.

[106] R. Rosipal and M. Girolami. An Adaptive Support Vector Regression Filter: A Signal Detection Application. In *International Conference on Artificial Neural Networks*, volume 2, pages 603–607, Edinburgh, Scotland, 1999.

[107] R. Rosipal and M. Girolami. An Expectation-Maximization Approach to Nonlinear Component Analysis. *Neural Computation*, 13(3):505–510, 2001.

[108] R. Rosipal, M. Girolami, and L.J. Trejo. Kernel PCA for Feature Extraction of Event-Related Potentials for Human Signal Detection Performance. In *Proceedings of ANNIMAB-1 Conference*, pages 321–326, Götegorg, Sweden, 2000.

[109] R. Rosipal, M. Girolami, and L.J. Trejo. On Kernel Principal Component Regression with Covariance Inflation Criterion for Model Selection. Technical Report 13, Computing and Information Systems, University of Paisley, Scotland, 2001.

[110] R. Rosipal, M. Girolami, L.J. Trejo, and A. Cichocki. Kernel PCA for Feature Extraction and De-Noising in Non-Linear Regression. *Neural Computing & Applications,* forthcoming, 2001.

[111] R. Rosipal, L. J. Trejo, and A. Cichocki. Kernel Principal Component Regression with EM Approach to Nonlinear Principal Components Extraction. Technical Report 12, Computing and Information Systems, University of Paisley, Scotland, 2000.

[112] R. Rosipal and L.J. Trejo. Kernel Partial Least Squares Regression in RKHS. Technical Report 14, Computing and Information Systems, University of Paisley, Scotland, 2001.

[113] S. Roweis. EM Algorithms for PCA and SPCA. In M. Jordan, M. Kearns and S. Solla, editor, *Advances in Neural Information Processing Systems 10*, pages 626–632, Cambridge, MA, 1998. The MIT Press.

[114] S. Roweis and Z. Ghahramani. A unifying Review of Linear Gaussian Models. *Neural Computation*, 11:305–345, 1999.

[115] D. Ruelle. Deterministic chaos: the science and the fiction. In *Proceedings of Royal Society of London*, volume 427 of *A*, pages 241–248, 1990.

[116] C. Saunders, A. Gammerman, and V. Vovk. Ridge Regression Learning Algorithm in Dual Variables. In *Proceedings of the 15th International Conference on Machine Learning*, pages 515–521, Madison, Wisconsin, 1998.

[117] I. Schoenberg. On interpolation by spline functions and its minimum properties. *International Series of Numerical Analysis*, 5:109–129, 1964.

[118] I. Schoenberg. Spline functions and the problem of graduation. In *Proceedings of the National Academy of Science*, volume 52, pages 947–950, 1964.

[119] B. Schölkopf, S. Mika, C.J.C. Burges, P. Knirsch, K.R. Müller, G. Rätsch, and A.J. Smola. Input Space vs. Feature Space in Kernel-Based Methods. *IEEE Transactions on Neural Networks*, 10(5):1000–1017, 1999.

[120] B. Schölkopf, A.J. Smola, and K.R. Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Technical Report 44, Max Planck Institut für biologische Kybernetik, Tübingen, 1996.

[121] B. Schölkopf, A.J. Smola, and K.R. Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10:1299–1319, 1998.

[122] T. Schreiber and A. Schmitz. Surrogate time series. *Physica D*, 142:356–382, 2000.

[123] H. Schwilden, H. Stoeckel, and J. Shuttler. Closed-loop feedback control of propofol anaesthesia by quantitative EEG analysis in humans. *British Journal of Anesthesiology*, 62:290–296, 1989.

[124] A. Sharma and R.J. Roy. Design of a Recognition System to predict Movement During Anesthesia. *IEEE Transactions on Biomedical Engineering*, 44(6):505–511, 1997.

[125] B. Silverman. Some Aspects of the Spline Smoothing Approach to Nonparametric Regression Curve Fitting. *Journal of the Royal Statistical Society, series B*, 47:1–52, 1985.

[126] L. Sirovich. Turbulence and the dynamics of coherent structures; Parts I-III. *Quarterly of Applied Mathematics*, 45:561–590, 1987.

[127] A.J. Smola. *Learning with Kernels*. PhD thesis, Technische Universität Berlin, 1998.

[128] A.J. Smola and B. Schölkopf. A Tutorial on Support Vector Regression. Technical Report NC2-TR-1998-030, NeuroColt2, Royal Holloway College, 1998.

[129] A.J. Smola, B. Schölkopf, and K.R. Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11:637–649, 1998.

[130] M. Stone. Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, series B*, 36:111–147, 1974.

[131] M. Stone and R.J. Brooks. Continuum Regression: Cross-validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression. *Journal of the Royal Statistical Society, series B*, 52(2):237–269, 1990.

[132] J.A.K. Suykens, L. Lukas, and J. Vandewalle. Sparse approximation using least squares support vector machines. In *IEEE International Symposium on Circuits and Systems ISCAS'2000*, 2000.

[133] F. Takens. Detecting strange attractors in fluid turbulence. In D. Rand and L.S. Young, editor, *Lecture notes in mathematics Volume 898*, pages 336–381. Springer, 1981.

[134] J. Theiler. Estimating fractal dimension. *Journal of Optical Society of America*, 7:1055, 1990.

[135] C.E. Thomsen, K.N. Christensen, and A. Rosenflack. Computerized monitoring of depth of anaesthesia with isoflurane. *British Journal of Anesthesiology*, 63:36–43, 1989.

[136] C.E. Thomsen, A. Rosenflack, and K.N. Christensen. Assessment of anaesthetic depth by clustering analysis and autoregressive modelling of electroencephalograms. *Computer Methods and Programming in Biomedicine*, 34:125–138, 1991.

[137] C. Thornton and J.G. Jones. Evaluating Depth of Anesthesia: Review of Methods. *International Anesthesiology Clinics*, 31:67–88, 1993.

[138] C. Thornton and D.E.F. Newton. The auditory evoked response: a measure of depth of anaesthesia. *Bailliere's Clinical Anesthesiology*, 3(3):559–585, 1989.

[139] R. Tibshirani and K. Knight. The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society, series B*, 61(3):529–546, 1999.

[140] R.J. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, series B*, 58(1):267–288, 1995.

[141] A.N. Tikhonov. On solving ill-posed problem and method of regularization. *Doklady Akademii Nauk USSR*, 153:501–504, 1963.

[142] A.N. Tikhonov and V.Y. Arsenin. *Solution of Ill-Posed Problems.* W.H. Winston, Washington, DC, 1977.

[143] M.E. Tipping and C.M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, series B*, 61:611–622, 1999.

[144] H.S. Traast and C.J. Kalkman. Electroencephalographic characteristics of emergence from propofol/sufentanil total intravenous anesthesia. *Anesthesia and Analgesia*, 81:366–371, 1995.

[145] L. J. Trejo and M. J. Shensa. Feature Extraction of ERPs Using Wavelets: An Application to Human Performance Monitoring. *Brain and Language*, 66:89–107, 1999.

[146] L.J. Trejo, A.F. Kramer, and J.A. Arnold. Event-related Potentials as Indices of Display-monitoring Performance. *Biological Psychology*, 40:33–71, 1995.

[147] M.E. Tunstall. Detecting wakefulness during general anesthesia for caesarian section. *British Journal of Anesthesiology*, 1:1321, 1977.

[148] M. Unser and A. Aldroubi. Polynomial Splines and Wavelets - A Signal Perspectives. In C.K. Chui, editor, *Wavelets - A Tutorial in Theory and Applications*, pages 99–122. Academic Press, Inc., 1992.

[149] R.E. Valdes-Perez and R.C. Conant. Information loss due to data quantization in reconstructability analysis. *International Journal of General Systems*, 9:235–247, 1983.

[150] R.J. Vanderbei. LOQO: An interior point code for quadratic programming. Technical Report SOR-94-15, Princeton University, NJ, 1994.

[151] V.N. Vapnik. *Estimation of Dependencies Based on Empirical Data.* Springer-Verlag, Berlin, 1982.

[152] V.N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.

[153] V.N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, 2nd edition, 1998.

[154] V.N. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16:264–280, 1971.

[155] V.N. Vapnik, S. Golowich, and A.J. Smola. Support vector method for function approximation, regression estimation, and signal processing. In *Advances in Neural Information Processing Systems 9*, pages 281–287, 1997.

[156] G. Wahba. *Splines Models of Observational Data*, volume 59 of *Series in Applied Mathematics.* SIAM, Philadelphia, 1990.

[157] G. Wahba. Support Vector Machines, Reproducing Kernel Hilbert Spaces, and Randomized GACV. In B. Schölkopf, C.J.C. Burges, and A.J. Smola, editor, *Advances in Kernel Methods - Support Vector Learning*, pages 69–88. The MIT Press, Cambridge, MA, 1999.

[158] C. Watkins. Dynamic alignment kernels. In A.J. Smola, P.Bartlett, B. Schölkopf and C. Schuurmans, editor, *Advances in Large Margin Classifiers*, pages 39–50. The MIT Press, 2000.

[159] G.S. Watson. Smooth regression analysis. *Sankhyā: The Indian journal of Statistica A*, 26:359–372, 1964.

[160] G. Widman, D. Bingmann, K. Lehnertz, and C.E. Elger. Reduced signal complexity of intracellular recordings: a precursor for epileptiform activity? *Brain Research*, 836:156–163, 1999.

[161] G. Widman, T. Schreiber, B. Rehberg, A. Hoeft, and C.E. Elger. Quantification of depth of anesthesia by nonlinear time series analysis of brain electrical activity. *Physical Review E*, 62:4898–4903, 2000.

[162] J.H. Wilkinson. *The Algebraic Eigenvalue Problem.* Clarendon Press, Oxford, 1965.

[163] C.K.I. Williams. Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In M.I. Jordan, editor, *Learning and Inference in Graphical Models*, pages 599–621. Kluwer, 1998.

[164] C.K.I. Williams and M. Seeger. The Effect of the Input Density Distribution on Kernel-based Classifiers. In P. Langley, editor, *Proceedings of the Seventeenth International Conference on Machine Learning.* Morgan Kaufmann, 2000.

[165] C.K.I. Williams and M. Seeger. Using the Nyström Method to Speed Up Kernel Machines. In T. K. Leen, T. G. Dietterich and V. Tresp, editor, *Advances in Neural Information Processing Systems 13.* The MIT Press, 2001.

[166] R.C. Williamson, A.J. Smola, and B. Schölkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. Technical Report NC-TR-98-019, NeuroCOLT, Royal Holloway College, 1998.

[167] H. Wold. Estimation of principal components and related models by iterative least squares. In P.R. Krishnaiah, editor, *Multivariate Analysis*, pages 391–420. Academic Press, New York, 1966.

[168] H. Wold. Soft Modeling by Latent Variables; the Nonlinear Iterative Partial Least Squares Approach. In J. Gani, editor, *Perspectives in Probability and Statistics, Papers in Honour of M.S. Bartlett*, pages 520–540. Academic Press, London, 1975.

[169] S. Wold. Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, 20(4):397–405, 1978.

[170] S. Wold, H. Ruhe, H. Wold, and W.J. Dunn III. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverse. *SIAM Journal of Scientific and Statistical Computations*, 5:735–743, 1984.

[171] W. Wu, D.L. Massarat, and S. de Jong. The kernel PCA algorithms for wide data. Part I: theory and algorithms. *Chemometrics and Intelligent Laboratory Systems*, 36:165–172, 1997.

[172] W. Wu, D.L. Massarat, and S. de Jong. The kernel PCA algorithms for wide data. Part II: Fast cross-validation and application in classification of NIR data. *Chemometrics and Intelligent Laboratory Systems*, 37:271–280, 1997.

[173] H. Zhu, C.K.I. Williams, R. Rohwer, and M. Morciniec. Gaussian Regression and Optimal Finite Dimensional Linear Models. In C. M. Bishop, editor, *Neural Networks and Machine Learning*, NATO ASI Series, pages 167–184. Springer-Verlag, 1997.

APPENDIX

A. APPENDIX

A.1 Implementation of the EM to KPCA algorithm

First assume the E-step; i.e equation (3.18) $\mathbf{Y} = (\mathbf{\Gamma}^T\mathbf{K}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{K}$. In first step we have to allocate and compute the $(n \times p)$ $\mathbf{\Gamma}$ and $(p \times p)$ $\mathbf{\Gamma}^T\mathbf{K}\mathbf{\Gamma}$ matrices. The computation of the latter matrix does not require storing of the $(n \times n)$ Gram matrix $\mathbf{K}$ because the procedure can be done by the elements of $\mathbf{K}$. In next step we can compute the right hand side $(p \times n)$ $\mathbf{\Gamma}^T\mathbf{K}$ matrix. However this will increase the memory requirements only by allocating one n dimensional vector into which we need to temporarily store the results of multiplication of one particular row of $\mathbf{\Gamma}^T$ matrix with the columns of $\mathbf{K}$. It is clear that this procedure can significantly slow down the algorithm. So, if we have enough memory we can perform the procedure for a couple of rows of $\mathbf{\Gamma}$ matrix at the same time. Moreover, if we can allocate additional $(n \times p)$ matrix the whole algorithm can be significantly faster as we will need to compute $\mathbf{\Gamma}^T\mathbf{K}$ only one time. In the next step we need to compute the $\mathbf{Y}$ matrix, however this will not increase the memory requirements because now we can overwrite the $\mathbf{\Gamma}^T\mathbf{K}$ matrix. Up till now we assumed we are dealing with a 'centralized' $\mathbf{K}$ matrix. As we noticed in subsection 3.1.1 the centralization is given by $\mathbf{K} \leftarrow \mathbf{K} - \mathbf{1}_n\mathbf{K} - \mathbf{K}\mathbf{1}_n + \mathbf{1}_n\mathbf{K}\mathbf{1}_n$ where $\mathbf{1}_n$ is a $(n \times n)$ matrix of $1/n$ elements. More detailed look will reveal that the individual columns $\{[\mathbf{1}_n\mathbf{K}]^j\}_{i=j}^n$ of the $\mathbf{1}_n\mathbf{K}$ matrix consist of the same numbers $\frac{1}{n}\sum_{i=1}^n \mathrm{K}_{ij}$ and that the $\mathbf{K}\mathbf{1}_n$ is just the transpose of $\mathbf{1}_n\mathbf{K}$. Similarly we can see that the elements of $\mathbf{1}_n\mathbf{K}\mathbf{1}_n$ are $\frac{1}{n^2}\sum_{i,j=1}^n \mathrm{K}_{ij}$. To speed up the centralization procedure this $n+1$ values can be computed in advanced and stored during execution of the EM algorithm. To compute the new $\mathbf{\Gamma}$ matrix in M-step $\mathbf{\Gamma}^{new} = \mathbf{Y}^{\mathrm{T}}(\mathbf{Y}\mathbf{Y}^{\mathrm{T}})^{-1}$ we can use the same approach as described for E-step. We can see that again we need to allocate the $(n \times p)$ and $(p \times p)$ matrices plus an n-dimensional vector.

Summarizing the both steps we can see that the memory requirements can be reduced to the $\mathcal{O}(p^2) + \mathcal{O}((p+1)n)$, however, as we discussed above, this will be done at the expense of the speed the algorithm. We conjecture that a good compromise can be achieved by allocating extra $(p \times n)$ space for storing of the $\mathbf{\Gamma}^T\mathbf{K}$ matrix.

A.2 Two effects of multicollinearity

To demonstrate two effects of multicollinearity we adopted a similar example as described in [80]. Consider the linear regression model

$$y = b_1x_1 + b_2x_2 + \eta,$$

where $x_1, x_2, y \in \mathcal{R}$ and $\eta \sim N(0, \sigma^2)$. Further assume that the variables x_1, x_2 and y are scaled to unit length and denote the correlation between x_1 and x_2 as r_{12} and by r_{jy} the correlation between x_j and $y, j = 1, 2$. The least squares estimate of the regression coefficients b_1, b_2 is then given by the solution of normal equations

$$(\mathbf{X}^T\mathbf{X})\hat{\mathbf{b}} = \mathbf{X}^T\mathbf{y}$$
$$\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix}$$

The inverse of $\mathbf{X}^T\mathbf{X}$ will be

$$\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 1/(1-r_{12}^2) & -r_{12}/(1-r_{12}^2) \\ -r_{12}/(1-r_{12}^2) & 1/(1-r_{12}^2) \end{bmatrix}$$

Now, recall that in the considered regression model we may express the variance and covariance of the estimates $\hat{b}_1, \hat{b}_2$ as $var(\hat{b}_j) = C_{jj}\sigma^2;\ j = 1,2$ and $cov(\hat{b}_1, \hat{b}_2) = C_{12}\sigma^2$, respectively, where $C_{ij}; i,j = 1,2$ are elements of $\mathbf{C}$ [80]. The strong multicollinearity between x_1, x_2 reflected by $|r_{12}| \rightarrow 1$ will result in $var(\hat{b}_j) \rightarrow +\infty$ and similar in $cov(\hat{b}_1, \hat{b}_2) \rightarrow \pm\infty$ in dependence on the sign of the correlation coefficient r_{12}.

Similarly it may be shown that in the case of p regressor variables the diagonal elements of $\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}$ are

$$C_{jj} = \frac{1}{1-R_j^2} \quad j = 1,2,\ldots,p$$

where R_j^2 is a coefficient of multiple determination between x_j and the rest of regressor variables. If there exists a strong multicollinearity between x_j and any subset of the other $p-1$ regressor variables this coefficient will be close to unity.

Consider now the squared distance from $\mathbf{b}$ to $\hat{\mathbf{b}}$

$$L^2 = (\hat{\mathbf{b}} - \mathbf{b})^T(\hat{\mathbf{b}} - \mathbf{b}).$$

The expected squared distance is

$$\begin{aligned} E[L^2] &= E[(\hat{\mathbf{b}} - \mathbf{b})^T(\hat{\mathbf{b}} - \mathbf{b})] = \textstyle\sum_{i=1}^p E[(\hat{b}_i - b_i)^2] = \sum_{i=1}^p var(\hat{b}_i) = \\ &= \sigma^2 trace(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2 \textstyle\sum_{i=1}^p 1/\lambda_i, \end{aligned}$$

where we used the facts that $E[\hat{\mathbf{b}}] = \mathbf{b}$ and that trace of matrix is equal to the sum of its eigenvalues. Thus if the matrix $\mathbf{X}^T\mathbf{X}$ is ill-conditioned at least one of its eigenvalues $\{\lambda_i\}_{i=1}^p$ will be small implying large distance between $\hat{\mathbf{b}}$ and $\mathbf{b}$ [80]. It may be also seen from

$$E[L^2] = E[(\hat{\mathbf{b}} - \mathbf{b})^T(\hat{\mathbf{b}} - \mathbf{b})] = E[\hat{\mathbf{b}}^T\hat{\mathbf{b}} - 2\hat{\mathbf{b}}^T\mathbf{b} + \mathbf{b}^T\mathbf{b}]$$

or equivalently from

$$E[\hat{\mathbf{b}}^T\hat{\mathbf{b}}] = \mathbf{b}^T\mathbf{b} + \sigma^2 trace(\mathbf{X}^T\mathbf{X})^{-1}$$

that the vector $\hat{\mathbf{b}}$ will be generally longer than the vector $\mathbf{b}$. This implies the estimate $\hat{\mathbf{b}}$ having coefficients too large in absolute value.

A.3 Kernel Functions

First, we describe several types of more frequently used kernel functions satisfying Mercer's theorem. Then several rules to create new kernel functions are provided.

A.3.1 Examples of Kernel Functions

Polynomial Kernels

Polynomial kernels of dth ($d \in N$) order have the form $K(\mathbf{x}, \mathbf{y}) = ((\mathbf{x}.\mathbf{y}) + c)^d$ and correspond to a dot product in the space of dth order monomials of the input coordinates. Assume a nonlinear mapping

$$\begin{aligned} \Phi : \mathcal{R}^2 &\rightarrow \mathcal{R}^6 \\ (x_1, x_2) &\rightarrow (x_1^2, x_2^2, x_1\sqrt{2c}, x_2\sqrt{2c}, x_1x_2\sqrt{2}, c) \end{aligned}$$

then for the product in a feature space we may write

$$(\Phi(\mathbf{x}).\Phi(\mathbf{y})) = x_1^2y_1^2 + x_2^2y_2^2 + 2cx_1y_1 + 2cx_2y_2 + 2x_1x_2y_1y_2 + c^2 = ((\mathbf{x}.\mathbf{y}) + c)^2$$

Usually we consider $c = 1$ or $c = 0$ corresponding to inhomogeneous and homogeneous polynomial kernels, respectively.

Translation Invariant Kernels

The kernels belonging to this group are of the form $K(\mathbf{x}, \mathbf{y}) = K(\mathbf{x} - \mathbf{y})$ and since the inner product in a feature space corresponding to $K(\mathbf{x} - \mathbf{y})$ is unchanged if the input vectors are translated by the same vector they are translation invariant.

The well-known representatives of translation invariant kernels are *radial* kernels[1] $K(\|\mathbf{x} - \mathbf{y}\|)$ into which category the widely used Gaussian kernel

$$K(\mathbf{x}, \mathbf{y}) = e^{-(\|\mathbf{x}-\mathbf{y}\|^2/d)}$$

belongs. The parameter d controls the width the bell-shaped Gaussian kernels. The conditions for the use of different type of translation invariant kernels as admissible kernels were discussed in [129].

Hyperbolic Tangent Kernels

The hyperbolic tangent kernels are of the form $K(\mathbf{x}, \mathbf{y}) = \tanh(\kappa(\mathbf{x}.\mathbf{y}) + \theta)$ however they satisfy Mercer's condition only for some range of the κ, θ parameters (e.g. for $\kappa < 0$ or $\theta < 0$ the condition is not satisfied [13]). The SVM models with this type of function correspond to the two-layer perceptron (artificial neural networks) learning machines.

Trigonometric Polynomials (Dirichlet Kernels)

Considering a $2d + 1$ dimensional feature space spanned by the Fourier expansion (trigonometric monomials)

$$\frac{1}{\sqrt{2}}, \cos x, \sin x, \ldots, \cos dx, \sin dx$$

the Dirichlet kernels

$$K(x, y) = \frac{1}{2} + \sum_{k=1}^{d} (\cos kx \cos ky + \sin kx \sin ky) = \frac{\sin{(d + 1/2)(x - y)}}{\sin \frac{(x-y)}{2}}$$

were constructed in [155] for interpolating data. For simplicity we assumed one-dimensional input data. In the case of higher dimensions the kernel can be computed as the sum over kernels computed using individual components of the input vectors [81].

However, it has been shown in [129] that this type of kernel corresponds to a regularization operator which suppresses only a finite band of frequencies and may lead to less smooth function estimates.

Kernels Generating Splines

In [155] the kernels generating the spline functions; i.e. piecewise polynomial functions of the form

$$f_d(x) = \sum_{r=0}^{d} a_r x^r + \sum_{s=1}^{n} w_s (x - t_s)_+^d \quad \text{where} \quad (x - t)_+ = \max\{(x - t), 0\}$$

defined on the interval $[0, a], 0 < a < \infty$ were considered. $t_1, \ldots, t_n \in [0, a]$ are the nodes, a_r, w_s are the real values and $d \in N$ represents the order of the spline function.

Assuming the $d + n + 1$ dimensional feature space spanned by

$$1, x, \ldots, x^d, (x - t_1)_+^d, \ldots, (x - t_n)_+^d$$

the inner product that generates the splines of order n in one dimension was derived

$$K(x, y) = \sum_{r=0}^{d} x^r y^r + \sum_{s=1}^{n} (x - t_s)_+^d (y - t_s)_+^d$$

[1] Radial kernels are also rotation invariant.

The generating kernel for the N dimensional splines is then the product of N one-dimensional generating kernels [155]. Further, in the case of SVM applications the authors considered an infinite number of nodes leading to the inner product of the form

$$K(x,y)=\sum_{r=0}^{d}x^r y^r+\int_0^a (x-t_s)_+^d(y-t_s)_+^d$$

Thus for linear splines we have the generating kernel

$$K(x,y)=1+xy+xy\min(x,y)-\frac{(x+y)}{2}(\min(x,y))^2+\frac{(\min(x,y))^3}{3}$$

Finally, in the case of assuming in practical applications B_d splines [148]

$$B_d(x)=\sum_{r=0}^{d+1}\frac{(-1)^r}{d!}\binom{d+1}{r}\left(x+\frac{d+1}{2}-r\right)_+^d$$

the corresponding kernel function has the form [155]

$$K(x,y)=\int_{-\infty}^{\infty}B_d(x-t)B_d(y-t)dt=B_{2d+1}(x-y).$$

A.3.2 Constructing Kernels from Kernels

To create new kernels function from existing kernels the following proposition (for the proof see [16]) may be used

Proposition: Let K_1 and K_2 be kernels defined over $\mathcal{X}\times\mathcal{X},\mathcal{X}\subseteq\mathcal{R}^N,a\in\mathcal{R}^+,f(.)$ a real valued function on $\mathcal{X}$, Φ a mapping $\mathcal{X}\rightarrow\mathcal{R}^M$ associated with a kernel K_3 defined on $\mathcal{R}^M\times\mathcal{R}^M$, $\mathbf{B}$ a symmetric positive semi-definite $N\times N$ matrix and p(.) a polynomial with positive coefficients. Then the following functions are kernels

1. $K(\mathbf{x},\mathbf{y})=K_1(\mathbf{x},\mathbf{y})+K_2(\mathbf{x},\mathbf{y})$
2. $K(\mathbf{x},\mathbf{y})=aK_1(\mathbf{x},\mathbf{y})$
3. $K(\mathbf{x},\mathbf{y})=K_1(\mathbf{x},\mathbf{y})K_2(\mathbf{x},\mathbf{y})$
4. $K(\mathbf{x},\mathbf{y})=f(\mathbf{x})f(\mathbf{y})$
5. $K(\mathbf{x},\mathbf{y})=K_3(\Phi(\mathbf{x}),\Phi(\mathbf{y}))$
6. $K(\mathbf{x},\mathbf{y})=\mathbf{x}^T\mathbf{B}\mathbf{y}$
7. $K(\mathbf{x},\mathbf{y})=\mathrm{p}(K_1(\mathbf{x},\mathbf{y}))$
8. $K(\mathbf{x},\mathbf{y})=e^{(K_1(\mathbf{x},\mathbf{y}))}$

Further forms for making kernels were discussed in [57, 44, 158, 2, 16].

A.4 Translation invariant kernels - regularization property

The regularization properties of a translation invariant kernel $K(\mathbf{x}-\mathbf{y})$ in connection to regularization networks and SVM were discussed in [35, 33, 129]. Similar to [33] we consider a continuous, symmetric and periodic function $K(x)$ whose Fourier coefficients α_n are positive. For simplicity we also assume K to be the function of one variable defined over $[0,2\pi]$. The extension to the case when K is defined over $\mathcal{R}^N$ can be also found in [33]. We can expand K into a uniformly convergent Fourier series

$$K(x)=\sum_{n=0}^{\infty}\alpha_n\cos(nx)=\alpha_0+\sum_{n=1}^{\infty}\alpha_n\cos(nx)$$

and using the elementary trigonometric formula $\cos(x-y) = \cos x \cos y + \sin x \sin y$ we can write

$$K(x-y) = \alpha_0 + \sum_{n=1}^{\infty} \alpha_n \sin(nx)\sin(ny) + \sum_{n=1}^{\infty} \alpha_n \cos(nx)\cos(ny). \tag{A.1}$$

Thus, using the fact that any kernel function may be expanded into the form (2.17) we can see that the function (A.1) defines RKHS $\mathcal{H}$ over $[0, 2\pi]$ with orthogonal basis

$$\{\psi_i(x)\}_{i=1}^{\infty} \equiv (1, \sin(x), \cos(x), \sin(2x), \cos(2x), \ldots, \sin(nx), \cos(nx), \ldots)$$

Thus, any function in $\mathcal{H}$ may be expressed in the form $f(x) = \sum_{i=1}^{\infty} b_i \psi_i(x)$ where b_i are Fourier coefficients of f. Using the norm in $\mathcal{H}$ as defined in section 2.4.1 we can see that

$$\|f\|_{\mathcal{H}}^2 = \sum_{i=1}^{\infty} \frac{b_i^2}{\lambda_i} < +\infty$$

and since the sequence λ_i is decreasing the constraint on the norm to be finite implies a decrease to zero of the Fourier coefficients corresponding to higher frequencies. The rate of decrease depends on the selected kernel function and defines the smoothness properties of the kernel.

B. APPENDIX

B.1 On connection of $h^{(0)}$ and CEn *measures*

Using the equation (8.11) we can write

$$h^{(0)} = \frac{-H_{\tau_0}(X_{m+1}/\mathbf{X}_m) + H_{\tau_0}(X_{m+1}) + H_{\tau_1}(X_{m+1}/\mathbf{X}_m) - H_{\tau_1}(X_{m+1})}{\tau_1 - \tau_0},$$

where the subscript τ was used to stress the dependence of the individual entropies on time delay used in construction of embedding vectors. In practice $(m+1)$th components of embedding vectors constructed from observed time-series using τ_0 and τ_1 will differ in $(m-1)(\tau_1 - \tau_0)$ data points. However, due to the assumed stationarity of the process we may consider $H_{\tau_0}(X_{m+1}) \equiv H_{\tau_1}(X_{m+1})$ and write

$$h^{(0)} = \frac{H_{\tau_1}(X_{m+1}/\mathbf{X}_m) - H_{\tau_0}(X_{m+1}/\mathbf{X}_m)}{\tau_1 - \tau_0}.$$

Finally, for $\tau_0 = 0$ we have $H_{\tau_0}(X_{m+1}/\mathbf{X}_m) = 0$ and $h^{(0)}$ is simply the estimate of CEn.