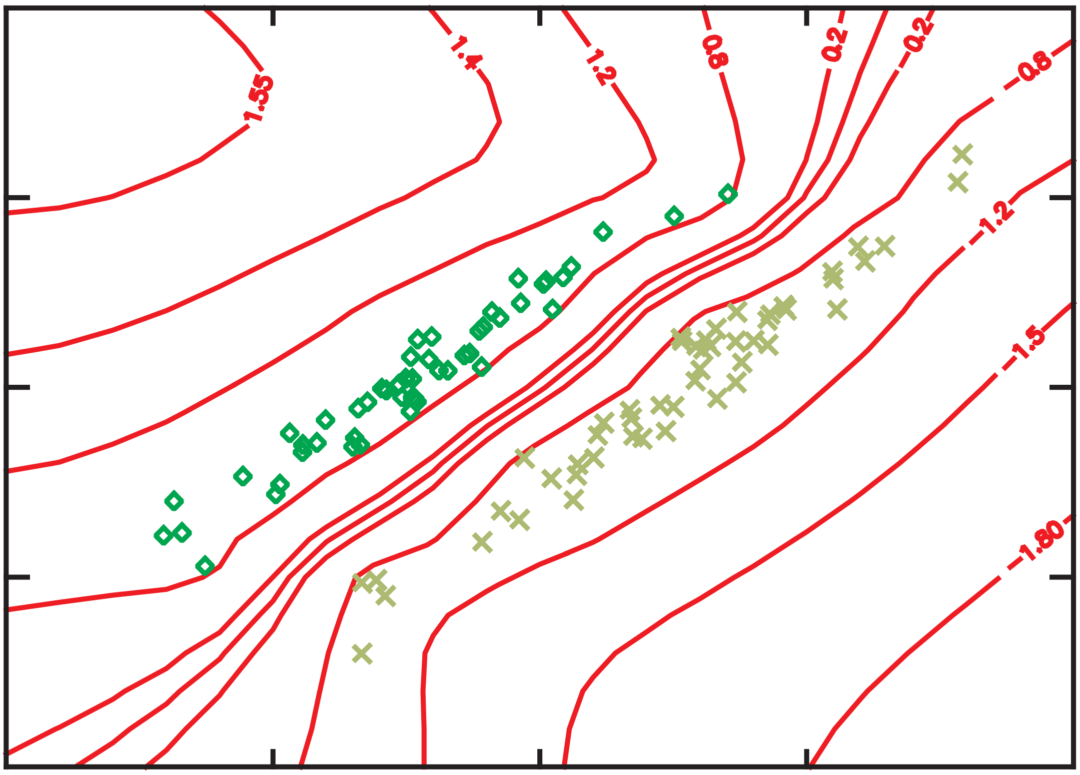

Helmut Alexander Kremper

Dimension-Reduction and Discrimination of Neuronal Multi-Channel Signals

Cover

The cover illustrates the two-class problem in two dimensions and the functioning of the dimension reduction approach based on radial basis functions (RBF). Randomly selected measurements (centres) serve as construction aids for a non-linear contour map. In a classification task, unlabeled measurements are mapped following the continuous hypersurface, as indicated by the contour lines.

Dimension-Reduction and Discrimination of Neuronal Multi-Channel Signals

Dimensionsreduktion und Trennung Neuronaler Multikanal-Signale

Dissertation

Presented in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Natural Sciences
(Dr. rer. nat.)

15 27

Helmut Alexander Kremper

Marburg/Lahn
Februar 2006

Vom Fachbereich Physik der Philipps-Universität als Dissertation angenommen am:	11.01.2006
Erstgutachter:	Prof. Dr. Reinhard Eckhorn
Zweitgutachter:	Prof. Dr. Ad Aertsen
Tag der mündlichen Prüfung:	08.02.2006

Preface

'We are making only the first, tentative steps in a long journey to the brain. Our progress so far might be compared to that of the Wright brothers, who flew the first aeroplane if their goal were to reach the moon.' (Vaadia, E. (2000) Nature,405:523)

In recent years, technological advances have lead to exciting developments in our understanding of how the brain performs neural computations, but most problems remain unsolved. Our ability to handle a vast abundance of sensory information in everyday situations, without any considerable effort, is of especially great interest. A detailed understanding of these complex processings would have significant impact and technical applications in many areas of research, e.g., in computer vision and robotics. The investigation and analysis of information transmitted by the nervous system after sensory stimulation is thus an undertaking that is widespread in neuroscience. In order to understand the spatio-temporal neural interaction, research groups have expended considerable effort in enhancing neurophysiological recording methods. At the neuron level, increasing numbers of micro-electrodes have been employed, allowing the simultaneous recording of cortical activity with a high temporal and spatial resolution. Due to these technological advances, the complexity and dimensionality of the recorded signals increases. The dimensionality of the observation space is determined by the number of electrodes and the sampling rate, amongst other things. In spite of these technical breakthroughs, there is still a major problem in getting consistent results across experiments designed to capture stimulus-response pairs under physiologically identical conditions. Reasons for the limited amount of stimulus-response samples (trials) are attributed to the experimental nature and instationarities in the signals.

Therefore, new recording techniques as well as new data mining approaches are required. The development of adequate signal processing software is very difficult and an active area of research. Classical approaches, analyzing the signals from each recording site separately or averaging the different time series, make use of the spatio-temporal correlations in an unsatisfactory way. A pair-wise analysis, e.g. by the cross-correlation function suffers from its linear model assumption. Sometimes prior information about the distribution of the samples can be derived, and a Gaussian assumption might be a promising approach, but in most cases, the underlying distributions are complex and unknown. The amount of data to adjust the parameters and in order to get reasonably low variance estimators, becomes extremely high. The statistical assumptions of many proposed methods are often not fulfilled, since the signals in the recording channels are often statistical dependent. Furthermore, it is often the case that neural responses generated by different stimuli

are similar, creating a strong overlap in their response properties. Such responses also contain signal components which are not correlated with the original stimulus.

I propose to search for a simpler, lower-dimensional representation of the neural data before attempting to estimate the statistical dependency between the stimulus and response sets. Nevertheless, because of the high-dimensionality and the complexity of the data any dimension reduction method causes a loss of information, and affects the statistical dependencies. For an adequate treatment, it is important to take the properties of the underlying data into consideration and to adapt the dimension reduction approach to the properties of the data. In this work I investigate various projection methods in a classification context (supervised learning) with regard to their transformation properties and their application to neural population data. Out of the large number of different approaches, I concentrate on dimension reduction techniques which fulfill the following aspects:

i) applicable in high-dimensional spaces, as well as to small-data samples,
ii) robust against noise and outliers,
iii) of low computational effort, and,
iv) of an objective representation.

Besides two linear approaches, I investigate members from regularization, kernel machines, nearest neighbors, and radial basis functions. Each method requires at most the solution of a linear system of equations. To quantify the information loss after dimension reduction, I estimate the information by cross-validation in combination with Monte Carlo sampling for various artificial data sets. Emphasis is placed on the relationship between the training size and the dimensionality. Practical behavior is further examined by discriminating signals from small neural networks. These results are used to investigate the dependence on the signal-to-noise ratio and the influence of irrelevant signal components on the reduction process. At the end, micro-electrode recordings from the visual cortex of two monkeys performing a matching-to-sample experiment will be investigated.

The comparison of the six methods shows that there is no best method, and that in special situations linear projection is sufficiently accurate. With respect to the investigated cortical population signals (recorded from the visual cortex of awake monkeys), the radial basis function approach seems to be most reliable and robust in the high-dimensional small sample case. In contrast to single channel approaches, this dimension reduction approach makes it possible to investigate multi-channel data simultaneously without pooling or averaging. As a consequence, taking the spatio-temporal statistical dependencies of multiple, simultaneously recorded signals into account, leads to higher significance of the information values compared to single channel approaches. At the same time, dimension reduction offers signal processing with high temporal and spatial resolution as well as an objective representation. The improvement in information rate by simultaneously using the signals from multiple recording sites can be quantified and the channels and signal segments which transmit relevant information can be determined reliably.

Outline

This thesis combines results that were obtained by using methods and techniques from across several disciplines including mathematics, physics, biology and computer science. General statistical concepts are introduced in *Chapter 1*. Main emphasis is put on the two-class problem in a supervised learning context. For the estimation of the statistical dependence, Bayes error, Shannon information, and the receiver operating characteristic (ROC) will be discussed. After an overview of experimental restrictions in cortical multi-channel recordings, the dimension reduction approach will be described. Finally, different numerical techniques for the quantization of the three distance measures will be compared.

In *Chapter 2*, six projection methods (performing a projection to the one dimensional space), adapted to the high-dimensional small sample case are described. Further, their relationship with each other is examined.

Chapter 3 serves as an empirical benchmark of the six projection methods. The information loss is quantified for four different uniformly distributed samples.

In *Chapter 4*, the behavior of the six methods will be investigated, applied to discretely sampled amplitude-continuous neuronal like signals. The performance of the projection methods is tested by varying the internal uncorrelated signal components systematically. An important result of Chapter 4 is that radial basis functions (RBF) reveal superior results in contrast to the other methods, in connection with continuous neural network signals.

In *Chapter 5*, I examine the application of the RBF method to multi-channel local field potentials and multi-unit activity recorded from the visual cortex of two awake monkeys, during a matching-to-sample experiment. Furthermore, I show how to combine the projection approach with other signal processing techniques in order to get further insight into cortical interaction and visual signal processing. Besides, the results of the multi-channel approach and classical single channel methods are compared.

Chapter 6 summarizes the findings and conclusions of my research. Beside the limits of the two-class dimension reduction approach, various generalizations are discussed. *Chapter 6* ends with a view to future research in this area.

Abbreviations

- **CDF** – **C**umulative **D**istribution **F**unction
- **EDF** – **E**mpirical **D**istribution **F**unction
- **ECG** – **E**lectro**C**ardio**G**ram
- **EEG** – **E**lectro**E**ncephalo**G**ram
- **KDE** – **K**ernel **D**ensity **E**stimation
- **KFD** – **K**ernel **F**isher **D**iscriminant
- **kNN** – **k**-**N**earest **N**eighbor
- **LCC** – **L**inear **C**orrelation **C**lassifier
- **LIE** – **L**inear **I**mplicit **E**uler Method
- **LFD** – **L**inear **F**isher **D**iscriminant
- **LFP** – extracellularly recorded **L**ocal **F**ield **P**otentials (1 - 140 Hz)
- **LS-SVM** – **L**east **S**quares **S**upport **V**ector **M**achines
- **MDS** – **M**ulti**D**imensional **S**caling
- **MEG** – **M**agneto**E**ncephalo**G**ram
- **MI** – **M**utual **I**nformation (Shannon)
- **MUA** – **M**ulti-**U**nit **A**ctivity (Action Potentials)
- **PCA** – **P**rincipal **C**omponent **A**nalysis
- **pdf** – **P**robability **D**ensity **F**unction
- **RBF** – **R**adial **B**asis **F**unction
- **RNG** – **R**andom **N**umber **G**eneration
- **ROC** – **R**eceiver **O**perating **C**haracteristics
- **SNR** – **S**ignal-to-**N**oise **R**atio
- **SOM** – **S**elf-**O**rganizing **M**ap
- **SUA** – **S**ingel-**U**nit **A**ctivity (Action Potentials)
- **SVM** – **S**upport **V**ector **M**achines

Notation

- P_i or $P(c_i)$ (a priori class probability, probability mass function)
- $p(x|c_i)$ (class likelihood)
- $p(c_i|x) = \frac{p(x|c_i)P(c_i)}{p(x)}$ (a posteriori probability)
- $p(x) = \sum_i P(c_i)p(x|c_i)$ (marginal probability density function)
- $p(x, y) = p(x|y)p(y) = p(y|x)p(x)$ (joint density)
- $p(x, y) = p(x) \cdot p(y|x) = p(y) \cdot p(x|y)$ (Bayes theorem)
- $p(x_1, ..., x_p) = p(x_p|x_1, .., x_{p-1}) \cdot p(x_{p-1}|x_1, .., x_{p-2}) \cdot ... \cdot p(x_2|x_1) \cdot p(x_1)$ (chain rule)
- $\bar{x}$ (mean response vector)
- I (identity matrix)
- M^T (transposed matrix)
- M^{-1} (inverse matrix)
- $\frac{df(x)}{dx} = f'(x)$ (derivation of a function)

Contents

1 Statistical Pattern Recognition **1**

1.1 Two-Class Problem 1
1.2 Bayes Error 4
1.3 Information Theory 5
1.4 Curse of Dimensionality 7
1.5 Dimension Reduction 9
1.6 Supervised Learning 11
1.7 Kernel Density Estimation 13
1.8 Differential Entropy Estimation 19
1.9 Significance Analysis 23

2 Projection Methods **25**

2.1 Demands on the Discriminant Function 25
2.2 Linear Correlation Classifier (LCC) 25
2.3 Linear Fisher Discriminant (LFD) 27
2.4 Regularized Discriminant Analysis (RDA) 29
2.5 Kernel Fisher Discriminant (KFD) 30
2.6 Approximate Distance Classifier (ADC) 32
2.7 Radial Basis Function (RBF) 33

3 Numerical Comparison **37**

3.1 Uniform Data 37
3.2 Difference in One Component 39
3.3 Difference in All Components 41
3.4 Differences in the Noise Level 43
3.5 Nonlinear Separable 45
3.6 Discussion 47

4 Neural Network Modeling **49**

4.1 Historical Overview 49
4.2 Mathematical Description 50
  4.2.1 Neural Output Signal 51
  4.2.2 Membrane Voltage 51
  4.2.3 Noise 53

4.2.4 Threshold . . . 53
4.3 Numerical Computation of the Neuron Dynamic . . . 54
4.3.1 Threshold Computation . . . 54
4.3.2 Membrane Voltage Computation . . . 56
4.3.3 Noise Generation . . . 57
4.4 Small Neural Networks . . . 58
4.4.1 Wilson-Cowan Oscillator . . . 58
4.4.2 Time-Resolved Discriminant Analysis . . . 62
4.4.3 Nonlinear Component Discrimination . . . 65
4.5 Summary . . . 68

5 Binocular Rivalry **71**
5.1 Experimental Setup . . . 71
5.2 Decoding of Neural Signals by Dimension Reduction . . . 73
5.2.1 Congruent - Non-Rivalrous Perception . . . 74
5.2.2 Incongruent Rivalrous Perception . . . 75
5.3 Feature Extraction . . . 78
5.3.1 Centre and Power Analysis . . . 78
5.3.2 Time-Frequency Discriminant Analysis . . . 81
5.3.3 Cross-Correlation Analysis . . . 85
5.3.4 Parameter Dependency of the Discriminant Analysis . . . 85
5.4 Stimulus Invariant Components . . . 89
5.4.1 Perception-Related Co-Modulation . . . 89
5.4.2 Multidimensional Scaling . . . 91
5.5 Discussion . . . 94
5.5.1 Binocular Rivalry in V1 . . . 94
5.5.2 Information Transmission by LFP and MUA Signals . . . 95
5.5.3 Single versus Multi-Channel Approaches . . . 96

6 General Discussion **99**
6.1 Multi-Class Problem . . . 99
6.1.1 Multi-Class Projection . . . 99
6.1.2 Multi-Class Test Statistic . . . 101
6.1.3 Single Trial Prediction . . . 102
6.2 Feature Extraction . . . 102
6.2.1 Wrapper Strategy . . . 103
6.2.2 Embedded Strategy . . . 104
6.2.3 Filter Strategy . . . 104
6.2.4 Surrogate Test . . . 105
6.3 Projection Approach . . . 106
6.4 Multi-Channel Population Data . . . 109
6.5 Outlook . . . 110

A Appendix **111**
A.1 111
A.2 111
A.3 112
A.4 113

Bibliography **115**

Abstract **129**

Zusammenfassung **131**

Danksagung **133**

Wissenschaftlicher Werdegang **135**

1 Statistical Pattern Recognition

'I have little more to say. I merely repeat, remember always your duty of enmity towards Man and all his ways. Whatever goes upon two legs is an enemy. Whatever goes upon four legs, or has wings, is a friend. And remember also that in fighting against Man, we must not come to resemble him. Even when you have conquered him, do not adopt his vices. No animal must ever live in a house, or sleep in a bed, or wear clothes, or drink alcohol, or smoke tobacco, or touch money, or engage in trade. All the habits of Man are evil. And, above all, no animal must ever tyrannize over his own kind. Weak or strong, clever or simple, we are all brothers. No animal must ever kill any other animal. All animals are equal.' (George Orwell, Animal Farm. A fairy story, 1945)

In this Chapter, I give a short introduction to statistical pattern recognition and define the basic notation. Two approaches are proposed to quantify the relation between two multivariate data sets. This will be the Bayes error measurement and the Shannon information measurement. I exemplify some problems with the practical computation of these measures. Further, I declare how to circumvent these problems for the most part by dimension reduction. The implementations in this work are restricted to two-class problems. The generalization to multi-class problems will be discussed in Chapter 6.

1.1 Two-Class Problem

Suppose an experiment leads to two data sets: $C_1 = \{x_i \in \mathbb{R}^d |\ i = 1, ..., m\}$ and $C_2 = \{y_j \in \mathbb{R}^d |\ j = 1, ..., n\}$, overall $m + n = N$ (training) samples (Fig. 1.1). The vectors x_i and y_j may be identified with the cortical activity of a person looking repeatedly at one of two different visual stimuli labeled C_1 and C_2. The indices $i = 1, .., m$ and $j = 1, .., n$ represent separate stimulus-response repetitions also called trials. The dimensionality d is determined by the number of features that have been collected.

A feature could be the voltage recorded at some position of the brain or some other metric measurement. In the following, I assume that the dimensionality and with it the number of investigated features stays constant during the whole experiment.

From a statistical point of view, we are dealing with random variables drawn from different classes (or categories). Each data set is characterized by a probability density function (pdf). This pdf is called the class likelihood or conditional density of class C_k, and is expressed as $p(z|C_k)$ or in short $p(z|k)$, with $\int p(z|k)dz = 1$ and $k \in \{1, 2\}$ [89]. Strictly speaking the pdf $p(z|k)$ should be written as $p_{Z|C}(Z = z|C = C_k)$ to indicate that this notation corresponds to a realization of a particular conditional density function of two random variables X and C. The probability for the presentation of stimulus C_k is given by $P(C_k)$ (in short $P(k)$), which is determined by the

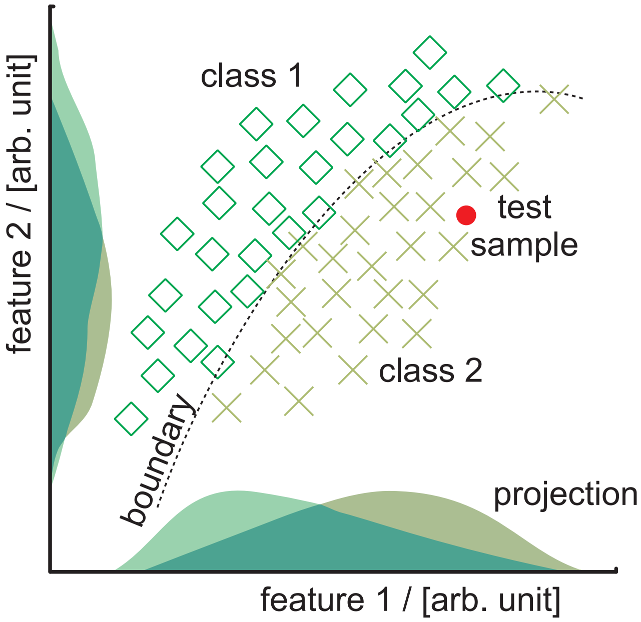

Figure 1.1: Schematic depiction of the multivariate two-class problem restricted to a two-dimensional feature space [89, pp.54]. Given are the samples of two different categories indicated by rhombus ◇ and cross ×. The two data sets are nonlinear separable. The dashed line symbolizes the boundary, necessary for perfect classification. In contrast, a separate investigation of each feature space (identical to a linear projection) would result in a large overlap.

frequency with which the stimulus has been presented during the experiment. In doing so, I assume that the samples x_i, y_j and their class labels C_1, C_2 are drawn identically and independently (i.i.d) from these distributions. Under the assumption that there are no stimuli from other classes we have: $P(1) + P(2) = 1$. The unconditional probability density function of the outcome, which is sometimes called the mixture density function, is given by: $p(z) = p(z|1) \cdot P(1) + p(z|2) \cdot P(2)$, with $\int p(z)dz = 1$. The joint probability density will be defined by: $p(c, z) := p(z|c) \cdot P(c)$, over the space $C \times \mathbb{R}^d$. It represents the probability of finding a pattern that is in class $c \in \{C_1, C_2\}$ having a feature vector $z \in \mathbb{R}^d$ and determines the process generating the data completely.

After an experiment has been done, investigating the (statistical) dependency or relation of the underlying measurements is a common problem in neuroscience but also in many other areas. At the moment there exists a vast amount of methods for statistical pattern recognition. A systematic description lies out of this scope. Excellent and established introductions into this field can be found, e.g., in [57; 63; 113; 262]. In short, two perspectives can be isolated:

i) the descriptive point of view, and
ii) the discriminative point of view.

In the first case, the main emphasis is put on an accurate description of the properties of the class likelihood. The centres and the forms of the two data sets are of special interest. In the second case, particular attention will be payed on the differences. So, the structures are not so important, but the overlap and the boundary are important. Although, this is not a stringent assignment of pattern recognition methods, it should be clear that a method from one of these groups, in general, is not appropriate to fulfill the properties of the other group [113; 170]. To say it in other words, I like to give an example:

It is well known that principal component analysis (PCA) has nothing to do with discriminative features optimal for classification, since it is only concerned with the covariance of all data regardless of the class. However, it may be very useful in reducing noise in the data [246].

In the following, I will concentrate myself on the discriminative aspects of the data. For this, I propose two distance or discrimination measurements. The first discriminative characteristic, in order to quantify the separability, will be the overlap between the two data sets. The second distance measurement quantifies the statistical dependency between the two random variables C and X. Both of them have some outstanding qualities, e.g., they are always positive: $d(C_1, C_2) \geq 0$. A precise definition of the two measurements will be given in the next two Sections.

1.2 Bayes Error

Among the possible distance measurements, Bayes error is the fundamental quantity in many scientific research areas [57]. For example, in radar technology people are confronted with the problem of assigning an unlabeled sample z to one class. If the probability density function of the outcome $p(z)$ and the posterior probabilities $p(c|z)$ are known, the lowest classification error that can be reached is given by the Bayes error, choosing the class with the highest posterior probability. For the two-class problem, Bayes error can be defined as:

$$d_{Bayes} := E = \int_z p(z) \cdot \min[p(1|z), p(2|z)]dz \ , \tag{1.1}$$

which is the mean over the minimum between the posterior probabilities [117]. This measurement can be reformulated by the Bayes theorem: $p(c|z) = \frac{P(c) \cdot p(z|c)}{p(z)}$. Inserting this formula into Equation (1.1) results in:

$$E = \int_z \min[p(z|1) \cdot P(1), p(z|2) \cdot P(2)]dz \ . \tag{1.2}$$

In this context, Bayes error quantifies the overlap between the prior weighted class conditional distributions. As I said before, Bayes error is always positive. In addition, Bayes error is symmetric and invariant against bijective transformations. If the conditional probabilities are completely separable, Bayes error will be zero. For equal prior probabilities $P(1) = P(2) = 0.5$, Bayes error will be between zero and 0.5. Given the prior probabilities, Bayes error can be easily normalized. The Bayes error measurement is related to other distance measurements, e.g., the total variation (L_1-distance). Since the minimum over two values can be expressed by: $\min(a, b) = \frac{1}{2}(a + b - |a - b|)$, Bayes error can be written in the form:

$$E = \frac{1}{2}(1 - d_{TV}) \ , \tag{1.3}$$

with $d_{TV} := \int |p(z|1) \cdot P(1) - p(z|2) \cdot P(2)|dz$. In some cases, it can be useful to compute an upper bound for the Bayes error. For example, if the samples are drawn from two multivariate Gaussian distributions, some error bounds are much easier to compute. I will review three of them.

1.) The Chernoff bound: $d_{Chernoff} := P(1)^s \cdot P(2)^{1-s} \int p(z|1)^s \cdot p(z|2)^{1-s}dz$, uses the inequality: $\min(a, b) \leq a^s \cdot b^{1-s}$ with $0 \leq s \leq 1$ and $a, b \geq 0$ [43].

2.) The Bhattacharyya bound: $d_{1/2} := P(1)^{1/2} \cdot P(2)^{1/2} \int p(z|1)^{1/2} \cdot p(z|2)^{1/2}dz$, is a special case of the Chernoff bound with $s = 1/2$, which is also related to the Hellinger distance: $d_{Hellinger} := \int (\sqrt{p(z|1) \cdot P(1)} - \sqrt{p(z|2) \cdot P(2)}\ )^2 dz$ [23].

3.) The asymptotic nearest neighbor error: $d_{NN} := 2 \int \frac{P(1) \cdot p(z|1) \cdot P(2) \cdot p(z|2)}{p(z)}dz$, considers the inequalities: $\min(a, b) \leq 2\frac{a \cdot b}{a+b} \leq \sqrt{a} \cdot b$ [46; 249].

1.3 Information Theory

Another pronounced measurement to quantify the relation between two or more data sets is the Shannon or mutual information (MI), which can be defined as [220]: [1]

$$d_{Shannon} := I(C;Z) = \sum_{c \in C} \int p(z,c) log \Big[\frac{p(z,c)}{p(z) \cdot P(c)} \Big] dz \ , \tag{1.4}$$

with a discrete stimulus set and a continuous response set. From the quotient (Equ. 1.4) you can see that Shannon information measures the average statistical dependence between a stimulus set (transmitter) and its responses (receiver). In contrast to the Bayes error measurement Shannon's information measurement quantifies the relation between the transmitted and the received signal. Using Bayes theorem to reformulate the joint probability function: $p(c,z) = p(z|c) \cdot P(c) = p(z) \cdot p(c|z)$, MI can be reformulated:

$$I = P(1) \int p(z|1) log \frac{p(z|1)}{p(z)} dz + P(2) \int p(z|2) log \frac{p(z|2)}{p(z)} dz \ . \tag{1.5}$$

In this context, Shannon information measures the average difference between the conditional pdfs $p(z|c)$ incorporating the information about the given stimulus and the mixture pdf $p(z)$ observing response z independent of the input stimulus. Therefore, the information describes how unique a response is determined by its input. Depending on the base of the logarithm Shannon information is given in natural units, abbreviated by **nats** (log_e), or in binary units, abbreviated by **bits** (log_2). Shannon information is positive and symmetric. In the two-class problem Shannon information is between zero and $log(2)$ (0.693 nat or 1 bit), and it can be easily normalized [51]. Like Bayes error, it is invariant to bijective transformations.

Annotation: The response set is continuous. Therefore, the inner term in Equation (1.4) ranging over all observed responses z is an integral. Nevertheless, the continuous situation is related to the discrete situation, and it can be shown that the information between two continuous random variables is the limit of the Shannon information between their quantified versions [47; 95]. Further details of information theory can be found in [47] and [221]. References about applications of information measurements in neuroscience can be found in, e.g., [24; 29; 65; 66; 181; 198].

---

[1] Information theory has been developed in the late 1940 by Hartley, Kolmogoroff, Kullback, and Shannon, among others [47].

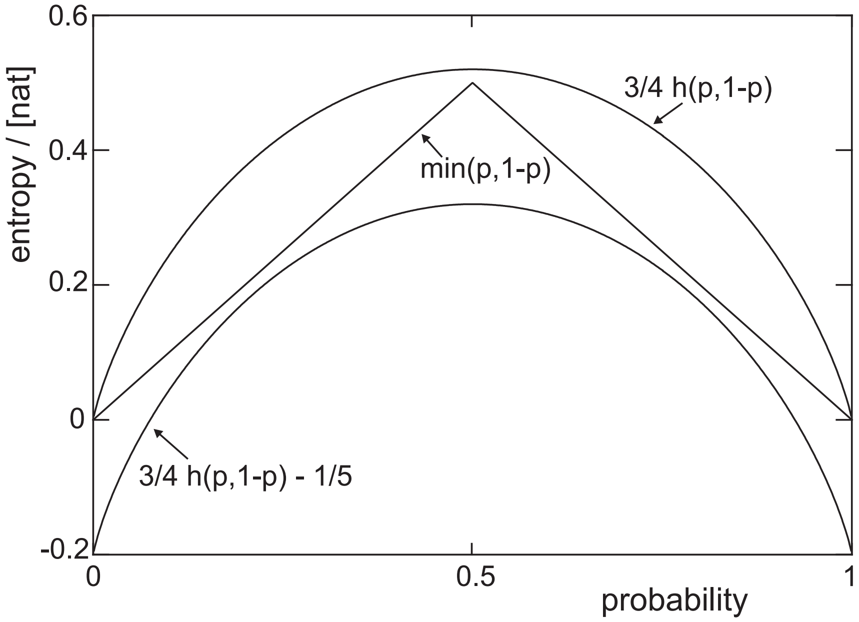

Figure 1.2: Illustration of a coarse upper and lower information boundary for the Bayes error in the two-class problem. The key element comes from the limitation of the expression $min[p, 1-p]$ by the entropy.

Shannon information and Bayes error are related to each other. An upper boundary on the Bayes error was obtained by [116] (see also [245]):

$$E \leq \frac{3}{4}h(c|z) = \frac{1}{2}(h(c) - I(C;Z)) \ , \tag{1.6}$$

with $h(c) = -\sum_{c \in C} P(c) log_e P(c)$ the entropy of the a priori class probability and $h(c|z) := -\int p(z) \sum_{c \in C} p(c|z) log_e p(c|z) dz$ the conditional differential entropy.
This can be seen by rewriting Bayes error:
$E = \int p(z) \min[p(1|z), p(2|z)] dz = \int p(z) \min[p(1|z), 1 - p(1|z)] dz$ since $p(1|z) + p(2|z) = 1 \ \forall z$, with $0 \leq p(1|z)$. From the inequality (see Fig. 1.2): $\min[p, 1-p] \leq \frac{3}{4}[-p \cdot log_e(p) - (1-p) \cdot log_e(1-p)]$, where $0 \leq p \leq 1$, Bayes error is bounded by: $E \leq -\int p(z)[p(1|z) \cdot log_e(p(1|z)) + p(2|z) \cdot log_e(p(2|z))] dz = h(c|z)$. At least, it can be shown that: $I(C;Z) = h(c) - h(c|z)$ [47]. A coarse lower bound on the Bayes error, involving the Shannon information measurement, can be derived in a similar way [246; 253]:

$$E \geq \frac{3}{4}h(c|z) - \frac{1}{5} = \frac{3}{4}(h(c) - I(C;Z)) - \frac{1}{5} \ . \tag{1.7}$$

The proof of Equation (1.7) can be accomplished using the inequality (see Fig. 1.2): $\min[p, 1-p] \geq \frac{3}{4}h(p, 1-p) - \frac{1}{5}$. Therefore, estimating the Bayes error enables one to determine boundaries for the Shannon information and vice versa. Similar boundaries using Renyi entropy can be found in [76].

1.4 Curse of Dimensionality

From a theoretical point of view Bayes error and Shannon information are prominent. Many aspects of them are well known for many decades and there are many application fields in neuroscience for Bayes statistics and information theory. Nevertheless, these approaches are not standard in the neuroscience community. The main reason why they are not in wider use currently lies in computational difficulties, at least for data of high dimensionality.

In principle, a complete description of the relation between two data sets is possible, if the conditional pdfs and their priors are known. In order to compute the Bayes or Shannon distance measurement, the main difficulty comes from the high-dimensional integral expressions, which have to be evaluated numerically, in both cases. Even if $p(z|1)$ and $p(z|2)$ are Gaussian, an analytical solution cannot be given in general (see Section 1.2). As I mentioned before, one way to avoid these difficulties is to use other distance measurements. Instead of the Bayes error, the Bhattacharyya bound or the mean square error given by: $d_{Mean} := \int (P(1) \cdot p(z|1) - P(2) \cdot (z|2))^2 dz$ are often analyzed [89]. Instead of the mutual information, some investigators have directed their attention to the quadratic information (e.g., [191]) given by: $d_{Quad} := \sum_c \int (p(c,z) - P(c) \cdot p(z))^2 dz$. These approaches reduce the computational problems to the problem of estimating $p(z|1)$ and $p(z|2)$ from the measured signals. Nevertheless, to estimate the underlying pdfs, further obstacles have to be overcome.

1.) *Small Sample Case.*

First, the number of available data is restricted. On one side, due to technological advances in micro-electrode employment and in order to understand the complex spatio-temporal processes in the brain, higher number of electrodes have been used in recent years recording neural responses simultaneously. This technological evolution has lead to data overloads. On the other side, when dealing with living organisms, the number of stimulus-response pairs available under stationary conditions from an experiment is often low [36; 129; 156; 272]. Variations due to tiredness cause instationarities in the signals which reduce the number of useful trials.

2.) *Curse of Dimensionality.*

When description of high-dimensional random variables is sought, the so called *curse of dimensionality* becomes the main factor affecting the performance [21]. The problem of investigating the *structure* of a set of responses N in a d-dimensional feature space where $N \leq d$, is not appropriately executable due to the inherent sparseness of high-dimensional spaces (see also Fig. 1.3). This means, the amount of data needed in order to get reasonably low variance estimators, becomes ridiculously high (growth generally exponential). In the small sample case without any further information, the reliable estimation of a pdf is impossible (Lugosi, personal communication (Louvain La Neuve Belgium, 2002)). With regard to the Shannon information, such approaches tend to overestimate the true distance, in the high-dimensional small sample case [173; 247].

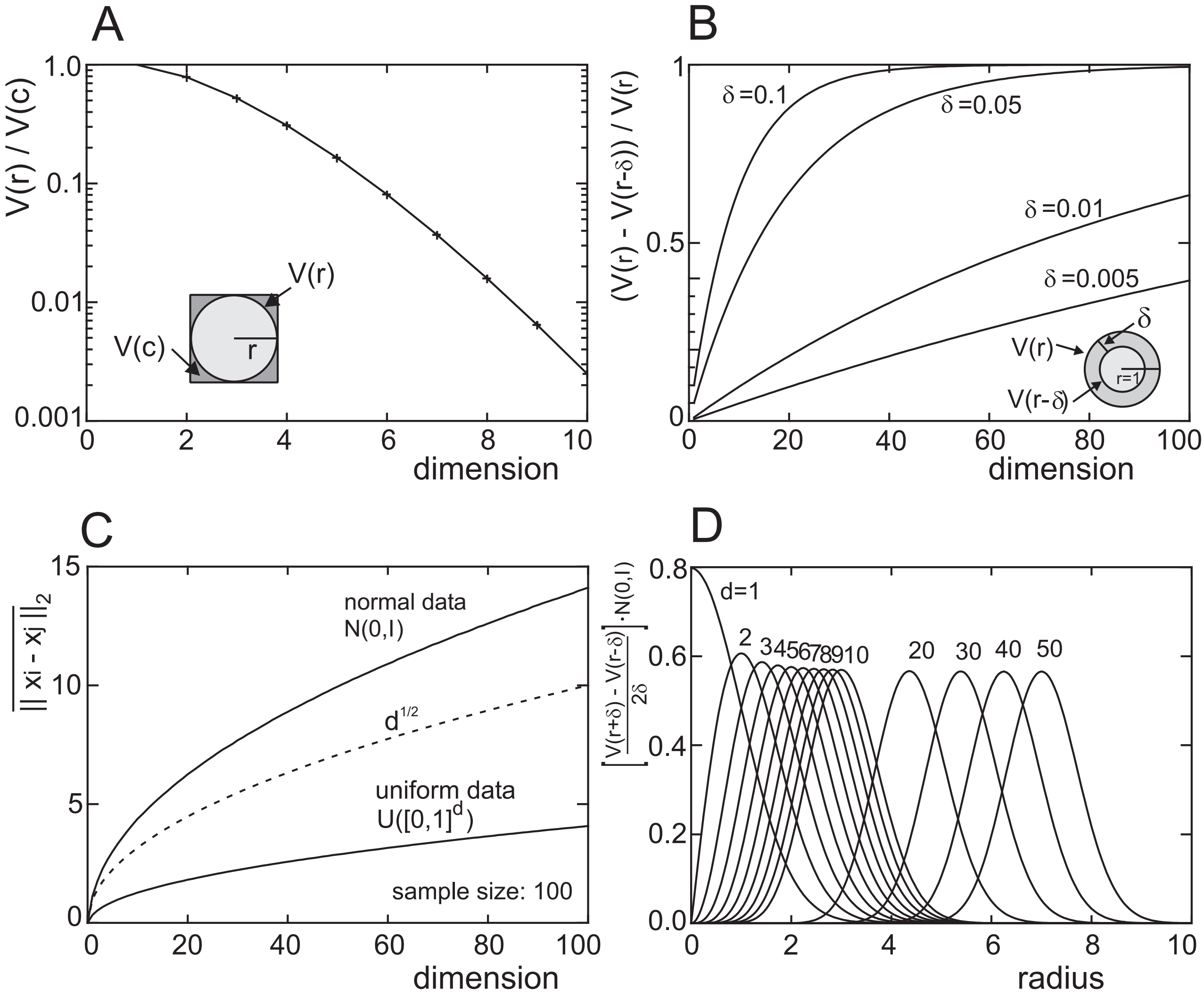

Figure 1.3: Illustrations of the curse of dimensionality and some related effects. (**A**) As dimensionality increases, the volume of a hypercube will be concentrated in its corners. The volume of the hypersphere of radius r and dimension d is known to be given by the equation: $V = \frac{2r^d}{d}\frac{\pi^{d/2}}{\Gamma(d/2)}$. (**B**) As dimensionality increases the volume of a sphere will be concentrated in its shell. The fraction of the volume in a shell defined by a sphere of radius $r-\delta$ inscribed inside a sphere of radius r is: $\frac{V(r)-V(r-\delta)}{V(r)} = 1-(1-\frac{\delta}{r})^d$. (**C**) Mean distance of 100 normal, uniform distributed points (solid). Length of the diagonal in the cube $[0,1]^d$ (dashed). (**D**) Within the high-dimensional space Gaussian distributed data tend to be collected in the tails expressed as product of the derivation of the hypersphere with a standard normal $N(0,I)$ [35, sect.1.4; 155; 256].

3.) *Complex Data Structures.*
A further difficulty comes from the underlying distributions which are often complex and contain unknown signal components not correlated with the actual task (e.g., maintained activity). Even for the same stimulus the signals indicate a large variation. In addition, in many cases the neuronal signals are similar for different stimuli. As a consequence, there is a strong overlap in their response behavior. A simple separable representation may not be given, and the signal components may not be independent [248]. Further, in some situations, the intervals between the stimulus presentations are too short, so that the response will be influenced by several previous stimuli [70]. For a reliable discrimination, all these circumstances make it necessary to handle the different recording sites simultaneously [199].

With regard to an *accurate* estimation of the Bayes error and the Shannon information, I will give reasons how to circumvent the curse of dimensionality by dimension reduction in the next Section.

1.5 Dimension Reduction

Reduction of dimensionality is widely accepted as a pre-processing and modeling tool to deal with data in high-dimensional spaces. There are several reasons to keep the dimensionality as low as possible.

1.) *In most cases, dimension reduction is necessary*.
If dimensionality isn't reduced, the computational effort of classical approaches can be daunting even for today's most powerful computers. If the data set is quite small the determination of the free parameters in a parametric approach is inappropriate, e.g., for multivariate density estimation and the variation of the estimation will be intensified by the curse of dimensionality (see Section 1.4). Highly correlated and redundant features tend to exert undue influence in classical data analysis approaches [61].

2.) *Dimension reduction enhances understanding*.
High-dimensional spaces are inherently difficult to understand. They possess properties that challenge and often contradict the intuition that have been developed from our experience with two- or three-dimensional geometry (see also Fig. 1.3). Although, dimension reduction may ignore serious aspects of the high-dimensional setting projection to one, two or three dimensions, e.g., for visualization of the data, is very helpful. A functional description can be given easier for low- than for high-dimensional data.

3.) *Dimension reduction is appropriate*.
In some cases the intrinsic dimensionality of the high-dimensional data is much lower [20]. For example, the set of vectors generated by rotation of an image describes approximately a one-dimensional continuous trajectory, embedded in a space of high dimensionality equal to the number of image pixels [260]. Besides, coherent structures lead to strong correlations between sensory

inputs, such as between neighboring pixels in an image, generating observations that lie on or close to a smooth low-dimensional manifold [203]. Even if the data complete the high-dimensional space, these might be a side effect of signal components not related with the actual task (maintained activity).

4.) *Dimension reduction is a natural process.*
Our mental representation of the world is formed by processing large numbers of sensory inputs [203]. For example, the human brain extracts a manageable small number of perceptual relevant features from about 30.000 auditory nerve fibers and two million optic nerve fibers [236]. Besides, dimension reduction plays a central role in human learning [20]. Every decision we make can be understood as a dimension reduction process.

Although there are many aspects approving dimension reduction, a critical point (bottleneck) is that the data structure will be modified. If the motivation lies in visualization, description or exploration, the challenge is to embed a set of high-dimensional observations into a low-dimensional (Euclidean) space that preserves as closely as possible the intrinsic local metric data structure [131; 236]. Methods which will be used in this context are principal component analysis (PCA), multidimensional scaling (MDS), Kohonen or self-organizing maps (SOM), and there nonlinear extensions [27; 49; 133; 142]. [2] As was mentioned in Section 1.1, I study the problem of dimensionality reduction from a discriminative point of view. In this context a projection of the high-dimensional observations to a much lower dimension has to be found: $g : \mathbb{R}^d \longrightarrow \mathbb{R}^1$, preserving the information of the most discriminative features between classes [25]. So the question comes up: What influence has the dimension reduction on the distance measures?

From a Bayesian perspective the following discriminant function: $g^*(z) = p(1|z) - p(2|z)$ together with the condition that an observation z would be assigned to class C_1 if $g^*(z) > 0$ and to C_2 if $g^*(z) < 0$ would minimize the error of misclassification. Besides this formula, the minimum-error-rate discriminant function is often expressed in the form: $g^*(z) = log\frac{p(z|1)}{p(z|2)} + log\frac{p(1)}{p(2)}$. A classifier that places a pattern in one of only two categories is called a dichotomizer [63, sect.2.4.2]. Correspondingly, the function g^* with this assignment will be called the Bayes decision function or Bayes classifier (maximum a posteriori classifier) [57; 63]. As a consequence, the high-dimensional space is divided by $g^*(z) = 0$ into separate regimes and the maximum of correct classification corresponds to a minimum of conditional risk and therefore to the Bayes error. Therefore, a randomly chosen discriminant function g increases the classification error and at the same time reduces the separability: $d_{Bayes}(C; g(Z)) \geq d_{Bayes}(C; Z)$. This can be seen since: $0 \leq E = \min_{g:\mathbb{R}^d \to \mathbb{R}} d_{Bayes}(C; g(Z))$ [57]. [3]

With regard to the Shannon information, a similar result exists, called *data processing inequality* [47, sect.2.8] and [183]. It states that the information cannot increase if the random samples were mapped by an arbitrary function g with $I(C; Z) \geq I(C; g(Z))$. In order to fulfill this inequality

---

[2] Reviews about unsupervised dimension reduction techniques can be found in [35] and [86].

[3] Antos, Devroye and Györfi have shown, theoretically, that without further conditions on the distribution $p(c, z)$ no rate-of-convergence can be obtained for the Bayes error [11].

the function $g(\cdot)$ depends only on the random variable Z and is conditionally independent of C. The proof can be given by the reformulation of the joint probability mass function based on the Bayes equation: $p(c, z, v) = p(c, z) \cdot p(v|c, z) = P(c) \cdot p(z|c) \cdot p(v|z)$, since $p(v|c, z) = p(v|z)$, with $v = g(z)$. Applying the chain rule of the differential entropy (see Section 1.8 and [47, sect.9.6]) to the Shannon information yields: $I(C; Z, V) = I(C; Z) + I(C; V|Z) = I(C; V) + I(C; Z|V)$ with $I(C; Z|V) := h(C|V) - h(C|Z, V)$. Since C and V are conditionally independent, given Z, we have $I(C; V|Z) = 0$. Since the information is always positive $I(C; Z|V) \geq 0$, we have $I(C; Z, V) = I(C; Z) + 0 \geq I(C; V)$. In this context the function $g(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^1$ can be regarded as an information channel and the data processing inequality can be interpreted so that no manipulation of the data can improve the inferences that can be made from the data. In general, the transmitted information will decrease. Therefore, every projection can be regarded as lower bound to the mutual information. Finding an optimal map means maximizing the mutual information between transformed data and their class labels. From the last equations these results may be interpreted so that maximizing the mutual information between transformed data and their class labels achieves the lowest possible bound to the error of a classifier [75]. Consequently, dimension reduction reduces the problem of underestimating the Bayes error or overestimating the Shannon information.

1.6 Supervised Learning

From the inequalities in the previous Section it is clear that an inappropriate dimension reduction is related to an information loss. An estimation of the a priori probability can be determined by the frequency with which each stimulus has been presented: $p(C_1) \approx \frac{1}{m}$ and $p(C_2) \approx \frac{1}{n}$ with $m + n = N$ the number of trials. Since, the class conditional probabilities are unknown, I propose to use a nonparametric supervised learning strategy for the adaptation of the discriminant function. From the small amount of stationary data available from neural recordings, it is essential to use as many simultaneously recorded signals as possible to adapt the free parameters. A re-substitution method, where all observations are used to design the discriminant function and used again to estimate the distance between the two classes, can lead to optimistically biased values [195].

The *sub-sampling* or *hold-out method*, dividing the observations from both classes into two disjoint subsets is a widely-used first preparation [263]. One subset, the so called learning or training set, will be used to assess a limited number of free parameters $\omega = (\omega_1, ..., \omega_k)^T$ and to adjust the discriminant function $g_\omega(z)$. The other elements of the second subset, called test set, will be projected by $g_\omega(z)$ to a low-dimensional space. [4] From the one-dimensional labeled data, the techniques in Section 1.7 and 1.8 may be used to estimate the distance between the signals from class C_1 and C_2. The sub-sampling method works reasonably well, but it often results in suboptimal performance, because, if you sub-sample enough examples to get a good test, you will not have enough examples left for training in the small-sample case. In the other case, if you hold-out enough examples to tune the discriminant function, you will not have enough samples for a reliable distance estimation (Section 1.7 and 1.8). Therefore, the hold-out method makes inefficient use of the data.

---

[4] In the two-class problem it is justified to map the high-dimensional signals to the real numbers.

There are several other, more sophisticated, alternatives. I carry out (k-fold) cross-validation, sometimes called rotation estimation [6; 231]. Cross-validation techniques are used amongst other methods in model selection [210; 229], discriminant analysis [153], regression [107], or in density estimation [219]. In k-fold cross-validation, the samples from both classes are randomly divided into k approximately equal-sized subsets. For each of the subsets, the remaining $k-1$ subsets are combined to form a training set. The discriminant function will be adapted k times, for each training set separately. The omitted subsets will be used as test sets, defining the elements for the Bayes error or information estimation. In order to avoid mismatch, the number of elements in the subsets from the two classes for the learning set should be equal or at least adapted (stratified) according to the priors, reducing the variance of the discriminant performance. Note that cross-validation is quite different from the hold-out method. The distinction between cross-validation and hold-out validation is important, because cross-validation is never worse for small data sets [102]. Cross-validation behaves very robust in statistical inference [140; 163]. There are asymptotical relations between k-fold cross-validation and some information theoretical criteria, e.g. [230]. (For an insightful discussion of the limitations of cross-validation among several learning methods, see [268].)

In general, cross-validation reduces the influence of the small data sample to the performance of the discriminant function. The whole data set can be used for training and Bayes error or information estimation. Empirical studies have shown, that cross-validation prevents from underestimating the Bayes error. Nevertheless, the properties of the discriminant function will be different for different subdivisions. Due to the finite data sets, it is also possible to overestimate the separability. Repeating cross-validation multiple times, using different splits into folds, provides a better estimate for the real error. This ensures that each element z of the data set is classified many times. Although this approach is time consuming, most of the data can be used for training while still having enough independent tests to estimate the separability. Altogether, cross-validation reduces the problem of overfitting and gives us the chance to compute some sort of confidence interval [164] (see also Chapter 6). The subdivision, necessary for cross-validation, will be carried out by sequential random sampling. For this purpose, I use an algorithm proposed by Ahrens and Dieter [4]. For the uniform random number generation, I use an algorithm from L'Ecuyer [78].
The whole procedure will be as follows:

- **Cross-validation.** Dividing the observations of each class into disjoint subsets.
- **Supervised learning.** Determine the free parameters ω in the discriminant function $g_\omega(z)$.
- **Dimension reduction.** Project the test sets applying $g_\omega(z)$.
- **Distance measure.** Estimation of the 1-dimensional distributions $p(g_\omega(z)|C_1)$ and $p(g_\omega(z)|C_2)$ associated to the classes C_1 and C_2.
- **Sequential random sampling.** Varying the order of the observations by sampling without replacement.
- **Repeating the procedure p times**, in order to estimate to what extent the classification properties depend on the subsample (centres).

Therefore, the computation of the Bayes error and the Shannon information has been replaced by two problems:

i) how to find an appropriate projection $g_\omega(z)$ and
ii) how to estimate the 1-dimensional probability density functions $p(g_\omega(z)|1)$ and $p(g_\omega(z)|2)$.

Before I introduce different methods for dimension reduction, I want to explain how to estimate the 1-dimensional conditional probability functions and how to estimate the two distance measures.

1.7 Kernel Density Estimation

The projected samples can be regarded as realizations of two univariate random variables. The probability for a special outcome z given the class affiliation is described by the two probability density functions $p(g_\omega(z)|1)$ and $p(g_\omega(z)|2)$. Since I'm interested in a reliable estimation of the Bayes error and the Shannon information, analyzing these univariate pdfs from the projected samples has to be done with care. For the estimation of univariate probability density functions different concepts exist. In our case it is not possible to deduce to which class of pdf the samples belong to. For this reason, I favor nonparametric density estimation approaches.

The histogram method is perhaps the oldest nonparametric method, partitioning in its simplest form the real line into a number of equal-sized cells of bin width h. The estimate of the density at a point x is taken to be the portion of samples in the cell n_c that straddles the point x to the total number of samples n: $p(x) = \frac{n_c}{n \cdot h}$. The quality of the histogram approximation depends heavily on an appropriate choice of the bin width. A Taylor series approximation up to second order of the mean integrated square error between the unknown pdf $p(x)$ and its histogram approximation reveals that the optimal bin width is given by: $h^* = (\frac{6}{\int f'(x)dx})^{1/3} \cdot n^{-1/3}$. Under the assumption that the unknown function can be approximated by a Gaussian distribution, the corresponding bin width is given by: $h^*_{Gaussian} = 3.490 \cdot \sigma \cdot n^{-1/3}$ [218]. The standard deviation σ can be approximated by the empirical standard deviation or another robust estimate: $\sigma = \min(\sigma, IQR/1.348)$ [219]. Although the histogram is easy to apply, it produces discontinuous estimations which are often biased [219]. In addition, histograms have bad convergence properties and show a volatile behavior (Fig. 1.4).

Another very popular method is k-nearest-neighbor density estimation [80; 159]. Given a point x, the k-nearest-neighbor approach is to fix the probability k/n and to determine the range $|x - x_k|$ which contains k samples centered on the point x. An approximation for the density is given by: $p(x) = \frac{k}{2 \cdot n \cdot |x - x_k|}$. [5] The estimator is positive and continuous everywhere, but with infinite integral, in that the tails go to zero like $1/|x|$ [100]. It can be shown that the density estimator is asymptotically unbiased and consistent if: $lim_{n \to \infty} k(n) = \infty$ and $lim_{n \to \infty} k(n)/n = 0$. The integer k is often chosen as $k \approx \sqrt{n}$. The optimal k for a Gaussian distribution can be found in [89, chap.6]. The simplicity of this approach has made it very popular among researchers for Bayes error estimation [81; 88]. For small sample sizes, the behavior will be very irregular, especially the slow decrease makes it unsuitable for compactly supported densities (Fig. 1.4).

---

[5] Some use $k - 1$ in the numerator [89, chap.6].

The density estimation based on an orthogonal expansion was first introduced by Cencov [37]. The basic approach uses the fact that a pdf can be expressed as a weighted sum of orthogonal basis functions: $p(x) = \sum_{i=1}^{\infty} a_i \cdot \phi_i(x)$. Trigonometric functions, Laguerre-, Legendre-polynomials and Hermite functions, among others, have been used in applications [55, chap.12]. The coefficients can be approximated by [68]: $a_j = \frac{1}{n}\sum_{i=1}^{n} \phi_j(x_i)$. In order to obtain a useful estimate of the density, a low-pass-filter has to be applied, since increasing the number of basis functions has the same effect as decreasing the bin width. In general, the approximation is not necessarily a density, and the values can be negative or may not integrate to 1. Therefore, a normalization procedure has to be used afterwards [100]. Further, orthogonal density approximations tend to be oversmoothed with a sensitive dependence to the number of frequency components (Fig. 1.4).

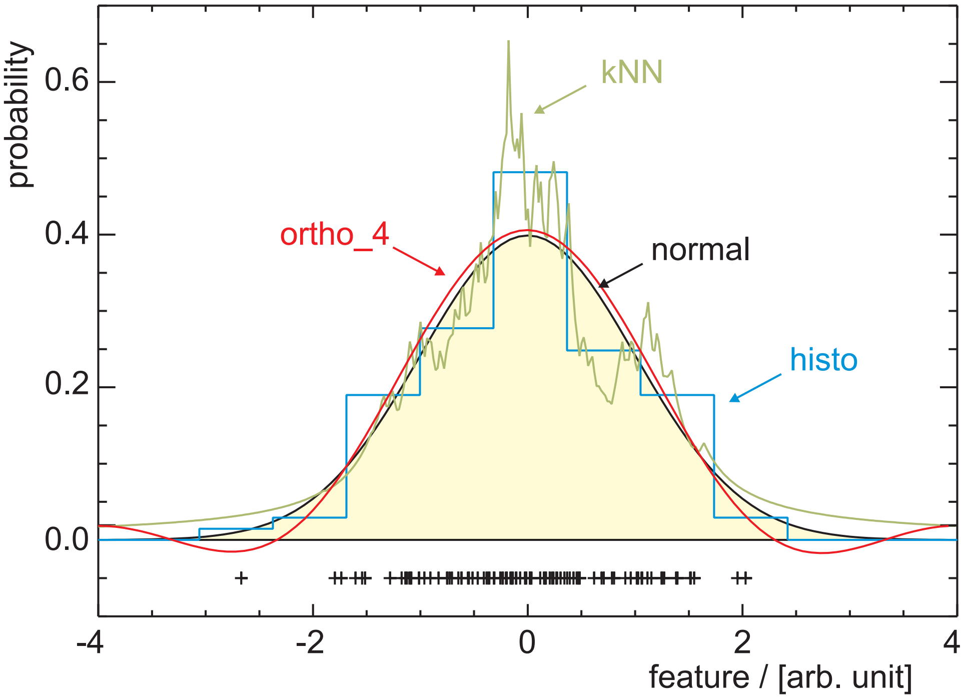

Figure 1.4: Approximation of a Gaussian distribution ($\sigma = 1$) by a histogram (histo with h = 0.68), a nearest neighbor approach (kNN with k=10) and an orthogonal series approximation (ortho with J=4) constructed from a random sample (sample size: 100). The distribution of the underlying random sample is depicted by crosses (+) below the base line. The approximations by kNN and ortho are not normalized.

For all three methods, many extensions exist. For example, more and more people use wavelets to approximate densities, and there are new data-driven bin width algorithms for the histogram construction. Nevertheless, the histogram density approximation has slow convergence properties and an irregular behavior. K-nearest neighbor shows a $1/|x|$ decrease at the boundaries, and performs inappropriate for asymmetric or compact densities, leading in general to a large bias. Orthogonal expansions tend to assess local effects, leading to under-smoothed solutions. Its behavior is strongly influenced by the number of frequencies. Due to these properties I will not investigate these methods further.

Today, most people use kernel density estimation methods (a mixture of single densities) also known as Akaike-Parzen-Rosenblatt methods [5; 185; 200]. It can be shown that most density estimators are specialized kernel density estimators (KDE). [6] Typically, the approximation has the form: $p(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i)$, with $\int K(x)dx = 1$ and the notation $K_h(x) = \frac{1}{h}K(x/h)$. Commonly used kernel functions for univariate data are the triangular kernel $K(x) = (1-|x|)1_{|x|<1}$, the Gaussian kernel $K(x) = \frac{1}{\sqrt{2\pi}}exp(-x^2/2)$ or the Epanechnikov kernel $K(x) = \frac{3}{4}(1 - x^2)1_{|x|<1}$ [74]. In general, the appropriate choice of the kernel K, which depends on the smoothness of the underlying pdf, is of minor importance for the density estimation.

More important than K, is the influence of the parameter h, which determines the width of the kernel function. If h is too small, the density is a collection of n sharp peaks, positioned at the sample points, leading, e.g., to an underestimation of the Bayes error. If h is too large, the density estimate is smoothed, and structures in the probability density function are lost. Correspondingly, the Bayes error would be overestimated, and the Shannon information would be underestimated. The different strategies for choosing the parameter h, lead in general to different values of h. The main problem to adjust the width of the kernel function, depends on the small sample size, heavily tailed densities, discontinuities, and edge effects. Most strategies focus on the problem of selecting a bandwidth to minimize some global measure of discrepancy between the estimated and the target density, like the integrated square error (L2-error) or the total variation (L1-error) [55; 219, sect.2.3]. [7] Strategies for choosing h rely on cross-validation, plug-in, maximum likelihood or information measurements. In the following, I will concentrate on so called plug-in approaches, which are very fast to evaluate and sufficiently robust [261].

1.) The simplest strategy for the bandwidth choice is the reference density method (one stage plug-in). Assuming that f is sufficiently smooth and that K is a nonnegative kernel, the asymptotic optimal bandwidth that minimizes the mean integrated square error $MISE(f_n) = E[ISE(f_n)] = E\int(f_n(x) - f(x))^2dx$ is given by: $h = (\frac{\int K^2(u)du}{n[\int u^2 \cdot K(u)\, du]^2 \int (f'')^2(x)dx})^{1/5}$. If the second order derivation $f''(x)$ is approximated by a Gaussian distribution, and if the kernel function is given by the Epanechnikov kernel, the resulting bandwidth is given by: $h_{ref,L2} = 2.345 \cdot \sigma \cdot n^{-1/5}$, with $\sigma = \frac{Q_{3n/4} - Q_{n/4}}{1.349}$ and Q the inter-quantile range [238]. Similar arguments can be used to obtain an equivalent bandwidth in the L1 setting. In this case, the corresponding bandwidth is slightly smaller: $h_{ref,L1} = 2.279 \cdot \sigma \cdot n^{-1/5}$, [22]. Terrell [237] obtains an upper bound for the optimal bandwidth in the L2 setting, seeking for a minimum of the expression $\int(f'')^2(x)dx$ under all densities with given variance $\sigma < \infty$. Towards this the optimal bandwidth in the L2 setting is bounded by: $h_{opt} \leq n^{-1/5} \cdot \sigma \cdot (\frac{243\int K^2(u)du}{35[\int u^2 \cdot K(u)\, du]^2})^{1/5}$. For the Gaussian kernel this yields to an over-smoothed bandwidth: $h_{OSG} = 1.144 \cdot \sigma \cdot n^{-1/5}$ and for a Epanechnikov kernel: $h_{OSE} = 2.532362 \cdot \sigma \cdot n^{-1/5}$ [238]. Although, in many cases the underlying density will be oversmoothed, these approaches are amazingly stable [22].

---

[6] Comprehensive reviews of state-of-the-art kernel smoothing methods are available in [219; 226; 261].

[7] L1 pays more attention to the tails [110]. For heavy-tailed distributions, it might be suitable to transform the initial data to a bounded interval [162].

2.) A very successful extension in the L2 setting, where a nonlinear equation must be solved, has been proposed by Sheather and Jones [224]. In the following, I use a simpler two-stage plug-in version reported in [40; 58]. Instead of approximating the term $\int f''(x)dx$ by a reference density, they approximate it by a kernel density. The estimation begins with a pilot bandwidth given by: $h_o = \sigma(\frac{32}{5\cdot n\cdot\sqrt{2}})^{1/7}$. The integral will be estimated by: $\rho = \frac{1}{n^2\cdot h_o^5}\sum_{i,j} L''''(\frac{x_i-x_j}{h_o})$, where L'''' is the fourth derivation of a smooth kernel, for which I will take the Gaussian density function $L = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$, as in [22]. The fourth order derivation can be easily obtained using Hermite polynomials. The resulting bandwidth is given by the minimum of two bandwidth: $h_{PI,L2} = \min[h_{ms,L2}, (\frac{15}{n\cdot\rho})^{1/5}]$ with $h_{ms,L2} = 2.532362\cdot\sigma\cdot n^{-1/5}$, $\sigma = \frac{Q_{3n/4}-Q_{n/4}}{1.349}$, the over-smoothed bandwidth for the Epanechnikov kernel. In the limit case, this kernel density estimation is universally consistent, which means that for all f: $lim_{n\rightarrow\infty}E\int|f_{h_{PI,L2}} - f| = 0$ [58].

3.) The third kernel density estimation method that I investigate has been proposed by Hengartner [118]. First a pilot kernel estimator has to be computed: $\bar{f}(x_j) = \frac{1}{n}\sum_i K_{h_o}(x_i - x_j)$ approximating the unknown density function at the sample points x_j. After that, the resulting kernel density estimation is given by: $f(x) = \frac{1}{n}\sum_j K_{h_1}(x_j - x)\bar{f}(x)/\bar{f}(x_j)$. Hengartner proves that for twice continuously differentiable densities with $h_1 = c\cdot n^{-1/5}$ and $h_o = c\cdot n^{-\alpha}$, $0 < \alpha < 1/5$, the corresponding kernel density estimation is asymptotically unbiased.

Annotation: The proposed universal bandwidth selection methods are, in general, not suited for discontinuous pdfs, e.g., uniform or exponential pdfs, but small samples from ill-behaved pdfs are all but indistinguishable from same size samples from smooth, small-tailed pdfs.

In the following, I apply these kernel density estimators to different pdfs varying the sample size between 20, 50 and 100. I compare the two-stage plug-in method with the method from Hengartner ($\alpha = 0.1$), both with Epanechnikov kernels, and the reference rule where the universal bandwidth h is twice that in the L2 setting together with a Gaussian kernel $h_{OSG} = 2\cdot 1.144\cdot\sigma\cdot n^{-1/5}$. To test the three estimators, data from four different probability density functions have been used. I investigate the behavior of the different methods varying the overlap between the pdfs (Fig. 1.5). For each constellation 5000 Monte Carlo simulations have been carried out, so that the standard error of the mean values are always lower than 0.004 .

1.) The first data set consists of two symmetric Gaussian probability density functions with equal variance ($\sigma = 1$) but different mean values (Fig. 1.5, first row). The underlying pdf is given by: $f(x) = 1/\sqrt{2\pi\sigma}exp(-\frac{(x-\Delta)^2}{2\sigma^2})$. Random samples have been generated by the Box-Muller method according to [97, pp.88]: $X = \sqrt{-2\cdot ln(U_1)}\cdot cos(2\pi U_2) - \frac{\Delta}{2}$ and $Y = \sqrt{-2\cdot ln(U_1)}\cdot sin(2\pi U_2) + \frac{\Delta}{2}$ with X and Y independently distributed as $N(-\frac{\Delta}{2}, 1)$ and $N(+\frac{\Delta}{2}, 1)$. U_1 and U_2 are independently distributed as uniform in $(0, 1)$ [78]. As you can see, the three kernel methods behave similarly. Independent of the sample size, all approximations tend to zero, as the overlap decreases. The approximation by the plug-in and the local adaptive method is better than the approximation by the reference method, which tends to overestimate the Bayes error. Interestingly, all three methods indicate a slight difference in the case where the two pdfs are identical ($\Delta = 0$).

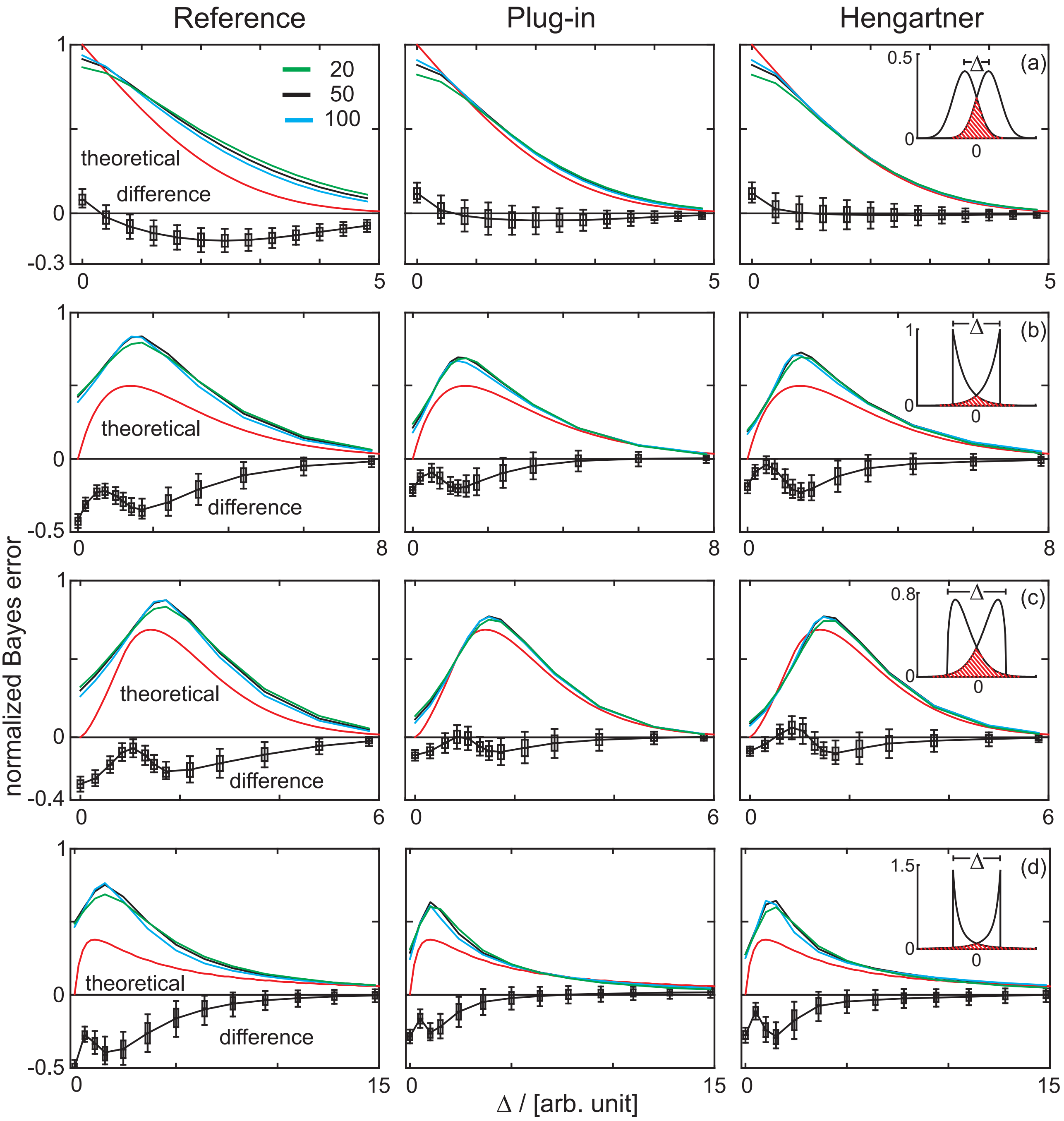

Figure 1.5: Overlap (normalized Bayes error) estimation in the small sample case by three kernel density estimation methods for four different distributions (**a**) Gaussian, (**b**) Exponential, (**c**) Weibull, (**d**) Pareto. Left column: Over-smoothed reference method, middle column: two-step plug-in method, right column: local adaptive method after Hengartner. Sample size varies between 20 (green), 50 (black) and 100 (blue) samples for each class. The boxplots (10, 25, 75, 90 quantiles) show the deviation of the 50 sample curves from the theoretical curve (red). Further descriptions can be found in the text.

This underestimation of the Bayes error becomes smaller for larger sample sizes.

2.) In the second test set, I use the three kernel methods to estimate the overlap between two exponential probability density functions where one pdf is mirrored and shifted relative to the other pdf (Fig. 1.5, second row). The underlying exponential probability density function is given by: $f(x) = \lambda exp(-\lambda(x-\Delta))$ for $0 < \lambda$ and $\Delta \leq x$. I establish the parameter λ equal to one in both classes. Samples have been generated by the following pseudo-random number generation method: $X = -log(U_1) - \frac{\Delta}{2}$ and $Y = +log(U_2) + \frac{\Delta}{2}$ [97, pp.92]. Due to the asymmetry, the normalized Bayes error increases and reaches a maximum. In the following, for larger distances ($\Delta \rightarrow \infty$) the theoretical overlap goes to zero. Qualitatively, the three KDE methods show a similar behavior. The most conservative estimation comes from the over-smoothed reference method. The performance of the plug-in and the local adaptive method are nearly equal.

3.) In the third example, I investigate two Weibull probability density functions where one pdf is mirrored and shifted relative to the other pdf (Fig. 1.5, third row). The underlying pdf is given by: $f(x) = \frac{a}{b}x^{a-1}exp(-\frac{1}{b}x^a)$ for $0 < a, b$ and $0 \leq x$. I choose $W(a = \sqrt{2}, b = 1)$. The pseudo-random number generation has been done by: $X = -log(U_1)^{1/a} - \frac{\Delta}{2}$ and $Y = +log(U_2)^{1/a} + \frac{\Delta}{2}$ [97, pp.99]. As before, the differences between the methods are mostly quantitative. Due to the discontinuity of the underlying pdf and the global construction of the bandwidth h, the normalized Bayes error approximation is far apart from the theoretical value for $\Delta = 0$. In contrast, for $\Delta \approx 1$, the local adaptive method and also the plug-in method show the tendency to underestimate the Bayes error.

4.) The last two data sets belong to two Pareto probability density functions. One pdf is mirrored and shifted relative to the other pdf (Fig. 1.5, fourth row). The underlying pdf formula is given by: $f(x) = \frac{a \cdot b^a}{x^{a+1}}$ for $0 < a$ and $0 < b \leq x$. I choose $P(a = \sqrt{2}, b = 1)$. Samples have been generated by the following pseudo-random number generation procedure: $X = U_1^{-\frac{1}{a}} - 1 - \frac{\Delta}{2}$ and $Y = -U_2^{-\frac{1}{a}} + 1 + \frac{\Delta}{2}$ [56; 97, pp.102]. The last example is important since the Pareto distribution has very long tails. If the centres are far apart from each other ($\Delta >>$) all kernel density estimation methods have numerical problems, resulting in an underestimation of the true Bayes error. In addition, due to the small sample size some methods show an inconsistent behavior.

Conclusion: Under the assumption that the generated samples represent the underlying probability density function – what is definitely the case – it is clear that the Bayes error estimation by a KDE approach has to be done with care, especially in the small sample case. With regard to the different methods, I can say that twice the L2 reference method seems to be too conservative. The plug-in method should be preferred to Hengartner's approach, since this local adaptive method shows a tendency to underestimate the real Bayes error. The underestimation can be intensified by multimodal, unusual data sets, or if the two data sets have a huge overlap. Naturally, many extensions exist for all proposed methods. For example, in the kernel density estimation setting, I have focused on a global bandwidth. More general approaches try to adapt the smoothing factor by $h = h(x_i)$ or $h = h(x)$ [1; 206; 226]. Interestingly, Terrell and Scott report good performance for small to moderate sample sizes, but performance deteriorates as sample size grows [239]. Another idea is to vary the *anchor point* of the kernel function [184]. Although it may be that these extensions are superior in special situations, first, they do not guarantee a correct estimation in every case, and second, they are very time consuming. With regard to the computational effort, I favor the L1 reference method. Furthermore, this method is more conservative than the two-stage plug-in method by Sheather and Jones but not as conservative as twice the L2 reference method (data not shown).

1.8 Differential Entropy Estimation

I have carried out (Section 1.3) that Shannon's information measure can be expressed as:

$$I(C;Z) = P(1)\int p(z|1)log\frac{p(z|1)}{p(z)}dz + P(2)\int p(z|2)log\frac{p(z|2)}{p(z)}dz\ . \tag{1.8}$$

As I previously mentioned, a kernel density estimation method might be used to approximate the conditional probability density functions $p(g(z)|1)$ and $p(g(z)|2)$ after projection and to estimate Shannon's information by numerical integration. Instead of this procedure, which has been successfully applied by many people, I pursue another strategy. In order to compute $I(C;g(Z))$, I use the fact that:

$$IC;g(Z)) = h(p(g(z))) - P(1)h(p(g(z)|1)) - P(2)h(p(g(z)|2))\ , \tag{1.9}$$

with $h(f(x)) = -\int f(x)logf(x)dx$, the differential entropy or mean uncertainty of a pdf [47, pp.225]. [8] So, the problem of information estimation has been reduced to the estimation of differential entropy. There are some similarities and differences between the entropy of a discrete random variable and the differential entropy. In the following, I will report a few of them. First, in contrast to the entropy of a discrete random variable, differential entropy can be negative. As in the discrete case a chain rule for differential entropy exist: $h(X_1,...,X_n) = \sum_{i=1}^{n} h(X_i|X_1,...,X_{i-1})$ and therefore the inequality: $h(X_1,...,X_n) \leq \sum_{i=1}^{n} h(X_i)$. Thereby, the conditional differential entropy of X given Y is defined as: $h(X|Y) := -\int\int p(x,y)log(p(x|y))dxdy = h(X,Y) - h(Y)$. In the discrete setting it can be shown that: $H(g(X)) \leq H(X)$. Interestingly, the same does not hold for the differential entropy $h(X)$ [47, pp.43]. Differential entropy has been extended for example by Renyi: $h_r(x) = \frac{1}{1-r}log[\int f^r(x)dx]$ with $lim_{r\to 1}h_r(x) = h(x) = -\int f(x)logf(x)dx$ [197]. The differential entropy is not invariant under linear transformations since: $h(aX) = h(X) + log(|a|)$.

Many methods exist for the computation of the differential entropy, and a lot of them can be found in the review article of Breilant et al. [19] which reviews powerful approaches when the underlying probability measures are known a priori to possess a given degree of smoothness. In the following, I compare three methods motivated by (1) kernel density estimation, (2) nearest neighbor search, and (3) the cumulative distribution function (cdf).

---

[8] The category of entropy was introduced in 1864 by Rudolf Clausius into physics and in 1949 by Claude Shannon into information theory. The word *entropy* stems from the greek word $\tau\rho o\pi\eta$, which means conversion or circular.

(1) The first method that I have analyzed has been proposed by Hall and Morton [111]. The estimator is given in natural units by: $h_H = arg\min_h -\frac{1}{n}\sum_i ln(\bar{f}_{i-}(x_i))$ where $\bar{f}_{i-}(x_i)$ is a KDE computed from (n-1) samples which exclude the point x_i, evaluated at the point x_i. Notice, this *leave-one-out* kernel density estimation does not involve numerical integration. The estimator is motivated by the fact that $\bar{h} = \frac{1}{n}\sum_i ln(f(x_i))$ is an unbiased and root-n consistent entropy estimation if $\int f(lnf)^2 < \infty$. Under certain regularity and smoothness conditions of the kernel function and the tails of the density the estimator $\bar{h}_H$ converges to $\bar{h}$. I use a double exponential kernel: $K_h(x) = 1/(2h)\cdot exp(-|x|/h)$ proposed in [71].

(2) The second method proposed by Kozachenko and Leonenko [144] has been successfully applied with the analysis of single-channel spike trains [259] and extended by [101; 128]. In the one dimensional space the estimator is given in natural units by: $h_L = \frac{1}{n}\sum_i ln(\rho_{i,k}) + ln(n-1) - \psi(k) + ln(2)$ with $\psi(z) = \frac{d}{dz}ln(\Gamma(z))$, the digamma function and $\rho_{i,k}$ the distance between x_i and its kth nearest neighbor. For integer values of k, $\psi(k)$ is given by: $\psi(k) = -\gamma + 1/1 + 1/2 + 1/3 + ... + 1/(k-1) = \sum_{i=1}^{k-1}\left(\frac{1}{i} - \gamma\right)$, with $\gamma = 0.57721566490$ the Euler-Mascheroni constant. It can be shown under very weak conditions of the density function that h_L is unbiased and consistent as $n \to \infty$.

(3) The third estimator relies on a reformulation of the integrand in terms of the cumulative distribution function [254]: $h(f(x)) = \int_0^1 ln\left(\frac{d}{dp}F^{-1}(p)\right)dp$. The estimator is given by rank ordered sample statistics: $h_V = V_{n,m} - ln(n) + ln(2m) - (1 - \frac{2m}{n})\cdot\psi(2m) + \psi(n+1) - \frac{2}{n}\sum_{i=1}^{m}\psi(i+m-1)$ with $V_{n,m} = \frac{1}{n}\sum_{i=1}^{n} ln(\frac{n}{2m}(x_{i+m} - x_{i-m}))$ and $x_{i-m} = x_1$ for $i - m < 1$ and $x_{i+m} = x_n$ for $n < i + m$. The positive integer m ($m < n/2$) determines the order of the spacing. Vasicek proved that the estimator h_V is consistent for densities with finite variance, as $n \to \infty$, $m \to \infty$ and $m/n \to 0$.

The computational effort of h_L and h_H is similar, whereas the computational effort of h_H is more expensive. A generalization to higher dimensions is possible for the first two methods but not for the Vasicek entropy estimator. One advantage of the entropy estimator h_H is based on the kernel density estimation approach, which can be used for visualization of the underlying density. In the following, I investigate the behavior of the three entropy estimators applying them to densities with different characteristics. The sample size varies between 20, 50 and 100. Table 1.2 summarizes the results. The theoretical values have been taken from [255] and [47], except for the last density $M(0,2)$. Altogether, the entropy of eight pdfs has been estimated: Uniform, Gaussian, Exponential, Logistic, Chauchy, Weibull, Pareto, and Bimodal Gaussian (Mixture).

Random number generation has been done according to [97] and [56]. I perform 10.000 Monte Carlo Simulations except for the Pareto distribution where I perform 50.000 iterations. This leads to a standard error of the mean values below 0.005. In the table (1.2), you can find the mean values plus minus their standard deviations. On average, the entropy estimator by Vasicek h_V shows the best performance, which is in accordance to the results of [266]. The mean values of Leonenko's entropy estimator h_L are not far behind, but in most cases their variance is larger. The variance of

PDF	Formula
Uniform	$U(0,1) = 1$ for $0 < x < 1$
Gaussian	$N(0,1) = \frac{1}{\sqrt{2\pi}} exp(-0.5x^2)$
Exponential	$E(0,1) = exp(-x)$ for $0 \leq x$
Logistic	$L(0,1) = exp(-x)(1 + exp(-x))^{-2}$
Cauchy	$C(0,1) = \frac{1}{\pi(1+x)^2)}$
Weibull	$W(\sqrt{2},1) = \frac{\sqrt{2}}{x^{\sqrt{2}-1}} exp(-x^{\sqrt{2}})$
Pareto	$P(\sqrt{2},1) = \frac{\sqrt{2}}{x^{\sqrt{2}+1}}$ for $1 \leq x$
Mixture	$M(0,2) = \frac{1}{2}(N(0,1) + N(2,1))$

Table 1.1: Probability density functions used as benchmark for the three differential entropy estimators.

Hall's entropy estimator h_H is similar to that of Vasicek's but especially for distributions with long tails (Cauchy, Weibull, Pareto) it shows a large bias to the right. The entropy estimator h_L and h_V show no problems with long tail pdfs but h_L shows a tendency to the left, underestimating the real entropy value. Interestingly, for the uniform probability density function all methods have the tendency to underestimate the theoretical entropy with a long tail to the left (not shown).

The standard deviation decreases rapidly with larger sample sizes. For example, analyzing 1000 samples generated from a Gaussian distribution (data not shown), the standard deviation of h_L reduces to 0.035 nat ($std(h_V) = 0.025$ nat, $std(h_H) = 0.024$ nat). Although, on average, the three entropy estimators behave differently, it might be possible that the mutual information could be identical. For example, let us identify the mixture density $M(0,2)$ with the unconditional pdf composed by two normal distributions with different mean values and same variances. According to the equations: $I = h(p(x)) - 0.5 \cdot h(1) - 0.5 \cdot h(2)$ and $h(X + c) = h(X)$ the underlying mutual information can be estimated from the table for the three methods: $I_H(50) = 0.34$ nat, $I_L(50) = 0.34$ nat, $I_V(50) = 0.34$ nat. At least it should be mentioned that, by this procedure, the resulting information could be negative. In such a situation, the corresponding information value will be replaced by the value zero.

	U(0,1) h=0	N(0,1) h=1.418	E(0,1) h=1	C(0,1) h=2.531	L(0,1) h=2	W($\sqrt{2},1$) h=0.822	P($\sqrt{2},1$) h=1.360	M(0,2) h = 1.755
$h_H(20)$	$.15 \pm .10$	$1.48 \pm .18$	$1.24 \pm .27$	$3.01 \pm .83$	$2.07 \pm .21$	$.95 \pm .21$	$2.05 \pm .72$	$1.83 \pm .16$
$h_V(20)$	$.00 \pm .09$	$1.35 \pm .18$	$0.99 \pm .24$	$2.64 \pm .51$	$1.93 \pm .21$	$.78 \pm .19$	$1.42 \pm .42$	$1.70 \pm .16$
$h_L(20)$	$.03 \pm .18$	$1.34 \pm .23$	$0.95 \pm .28$	$2.28 \pm .40$	$1.90 \pm .25$	$.76 \pm .24$	$1.25 \pm .40$	$1.70 \pm .21$
$h_H(50)$	$.10 \pm .05$	$1.45 \pm .10$	$1.19 \pm .17$	$3.15 \pm .77$	$2.04 \pm .13$	$.91 \pm .12$	$2.08 \pm .59$	$1.79 \pm .09$
$h_V(50)$	$.00 \pm .04$	$1.39 \pm .11$	$0.99 \pm .15$	$2.60 \pm .29$	$1.98 \pm .13$	$.81 \pm .11$	$1.38 \pm .25$	$1.73 \pm .09$
$h_L(50)$	$.01 \pm .11$	$1.38 \pm .15$	$0.98 \pm .18$	$2.41 \pm .27$	$1.95 \pm .16$	$.80 \pm .15$	$1.31 \pm .26$	$1.72 \pm .14$
$h_H(100)$	$.07 \pm .03$	$1.44 \pm .07$	$1.15 \pm .12$	$3.22 \pm .71$	$2.03 \pm .09$	$.89 \pm .08$	$2.07 \pm .52$	$1.78 \pm .06$
$h_V(100)$	$.00 \pm .03$	$1.40 \pm .07$	$0.99 \pm .10$	$2.57 \pm .19$	$1.99 \pm .09$	$.81 \pm .08$	$1.37 \pm .18$	$1.74 \pm .07$
$h_L(100)$	$.00 \pm .08$	$1.39 \pm .10$	$0.99 \pm .13$	$2.47 \pm .19$	$1.97 \pm .11$	$.81 \pm .11$	$1.33 \pm .19$	$1.73 \pm .09$

Table 1.2: Comparison of three approaches for the estimation of the differential entropy. The abbreviation $h_H(20)$ stands for the differential entropy estimator by Hall and Morton applied to a data set with 20 samples. Correspondingly, the abbreviation h_V (h_L) stands for the differential entropy estimator by Vasicek (Kozachneko and Leonenko). The entropy values are given in natural units [nat]. Eight different probability density functions have been analyzed. The sample size varies between 20, 50 and 100 samples.

1.9 Significance Analysis

In order to determine if the computed values for the Bayes error and the Shannon information exceed a predefined significance level, I use a nonparametric test based on area statistics [212]. This statistic is motivated by the receiver operating characteristic (ROC), also known as p-p-plot, a classic methodology from signal detection theory [69; 103], that is common in psychology and medical diagnosis and has recently begun to be used more generally in machine learning work [192; 257]. The hypothesis test can be seen as an L1-version of the Cramer-von Mises test under general alternatives. Let $F(x) = \int_\infty^x p(t|1)dt$ and $G(x) = \int_\infty^x p(t|2)dt$ denote the cumulative distribution functions of the projected conditional pdfs. Then, the two maps: $F(G^{-1}(p))$ and $G(F^{-1}(p))$ are restricted to the range $[0,1]^2$. Each map contains information about the relation of the two random variables. The graphs of the two maps coincide with the diagonal of the unit square, if and only if, $F = G$. Therefore, the area between the two curves given by:

$$T = \int_0^1 \left| F(G^{-1}(p)) - G(F^{-1}(p)) \right| \, dp \, , \tag{1.10}$$

can be taken as a measure of dissimilarity (Fig. 1.6, right).

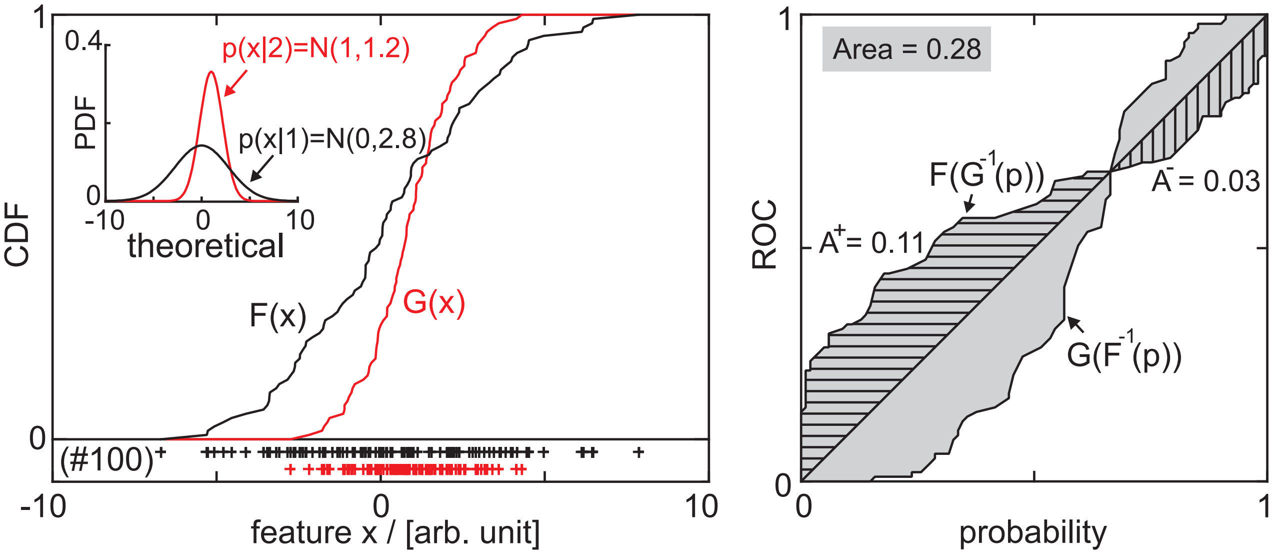

Figure 1.6: Left: Cumulative distribution functions of two random samples according to two Gaussian distributions illustrated in the upper left. Right: Receiver operating characteristic. The area from T relative to the diagonal amounts to 0.28 (see text). The two pdfs are significantly different ($p < 0.01$) but there is no dominance of one distribution against the other. The area under the curve $F(G^{-1}(p))$ is 0.58.

If the quantity T is zero the two distributions are identical (not separable). If the area between the two distributions is one ($T = 1$), both distributions are completely different (separable). An important feature of this distance measurement is that it is invariant against one-to-one transformations of the decision axis. A corresponding test statistic can be defined using the empirical distribution functions (edf) $F_n(x)$ and $G_m(x)$ determined by the random samples [212]:

$$T_{m,n} = \left(\frac{mn}{m+n}\right)^{1/2} \int |F_m(x) - G_n(x)| \frac{m \cdot dF_m(x) + n \cdot dG_n(x)}{m+n} , \tag{1.11}$$

or in terms of rank order statistic [13]:

$$T_{m,n} = \left(\frac{nm}{(n+m)^3}\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^{m}\left|R_X(i) - i\frac{m+n}{m}\right| + \frac{1}{m}\sum_{j=1}^{n}\left|R_Y(j) - j\frac{m+n}{n}\right|\right) , \tag{1.12}$$

where $R_X(i)$ and $R_Y(j)$ are the ranks of x_i resp. y_j in the ordered pooled sample. This test is consistent against any alternative $F \neq G$, not only location and scale. Quantiles of the distribution of $T_{m,n}$ under the null hypothesis H_o (F=G) for the balanced case (m=n) can be found in [212] and for the unbalanced case in [13]. In the balanced case, this test performs better than other well known two-sample tests like the Kolmogoroff-Smirnov test, which can be illustrated as the absolute maximal vertical distance between the ROC curve and the diagonal, or the Wilcoxon-Mann-Whitney test, which evaluates differences in the mean. Further, specializations of this statistic in terms of dominance in one direction of F over G are given in [145]. Within this work the following two nonlinear rank statistics are relevant:

$$A^+_{m,n} = \frac{1}{m}\sum_{i=1}^{m} \max\left(\frac{R(i)-i}{n} - \frac{i}{m}, 0\right) , \tag{1.13}$$

and

$$A^-_{m,n} = \frac{1}{m}\sum_{i=1}^{m} \max\left(-\frac{R(i)-i}{n} + \frac{i}{m}, 0\right) . \tag{1.14}$$

A^+ and A^- correspond to the area between the graph and the diagonal in the unit square, each above and under the diagonal (Fig. 1.6, right). These measurements enable us to find out at which side of the diagonal the graph $F(G^{-1}(p))$ runs or if the graph crosses the diagonal without visual inspection. This will be important to get a deeper understanding of the various dimension reduction methods proposed in the next Chapter. In addition, these quantities have some relevance in terms of misclassification.

2 Projection Methods

'There is, however, a strong sentiment in the neuroscience community that there is a bottle-neck in data analysis, a sense that there are not yet adequate tools for understanding multi neuronal recordings.' (Vaadia, E. (2000) Nature,405:523)

In this Chapter I will present six discriminant functions adapted to the small sample case. Two linear methods and four nonlinear methods have been taken into account. The behavior of the different approaches will be compared from a theoretical point of view. A numerical comparison will be given in the next Chapter.

2.1 Demands on the Discriminant Function

The choice of the discriminant function is an essential part of the dimension reduction process, as has been expressed by the mapping properties of Bayes error and Shannon information in Chapter 1. After that, the projection of the data should maximize the discriminability of the two sets, or in terms of information theory, maximize the statistical dependency between the stimuli and the corresponding responses. From the abundance of possible approaches I have chosen those which further fulfill the following criteria. The method has to be:

i) applicable in high dimensional spaces,
ii) applicable for small data samples,
iii) robust, and
iv) easy to implement.

In all cases, the relative distances between the samples play a central role. For the distance between two realizations I apply the Euclidean-norm: $\|x - y\| = (\sum_k (x_k - y_k)^2)^{1/2}$.

2.2 Linear Correlation Classifier (LCC)

A popular measure of separation is the distance of the sample means [195]. At first, the mean responses $\overline{x} = \frac{1}{m}\sum_i x_i$ and $\overline{y} = \frac{1}{n}\sum_j y_j$ of the two classes have to be computed. An unknown sample z will be projected according to:

$$g_{LCC}(z) \; = < \overline{x}, z > - < \overline{y}, z > = < \overline{x} - \overline{y}, z > \quad . \tag{2.1}$$

This approach is named differently in the literature. Some prefer the name *minimum-distance classifier* [63, sect.2.6], using the function $g(z) = \|z - \overline{x}\|^2 - \|z - \overline{y}\|^2$. I prefer the name *linear correlation classifier* according to Fukunaga [89, pp.127], since the terms $\|\overline{x}\|^2$ and $\|\overline{y}\|^2$ are only interesting in classifications adapting the threshold. For the discrimination process, these quantities are not relevant. In addition to the zero passage, the slope of the discriminant function $g(z)$ is also irrelevant for the Bayes error and Shannon information estimation, which are invariant under linear transformations. The construction of the linear projection maximizes the mean distance between the two classes so that the feature space is separated into parallel hyperplanes (affine sets) by $g_{LCC}(z) = const$, which are orthogonal to the distance vector $\overline{x} - \overline{y}$. Therefore, the discriminant function $g(z)$ gives an algebraic measure of the distance between z and the hyperplane.

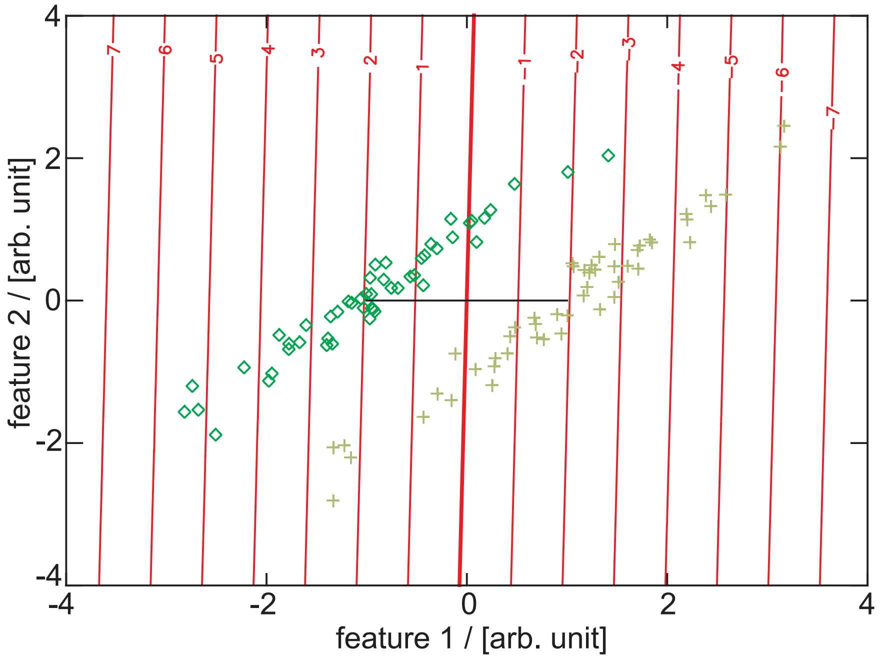

Figure 2.1: Contour plot according to the linear correlation classifier (LCC) for a 2-dimensional artificial data set. The underlying probability density functions have the same covariance matrix but different mean vectors. The distance between the centres of the two data sets is given by the straight line. According to the contour lines, projection by LCC corresponds to a projection to the first feature of the data. Similar to a linear projection to the abscissa or the ordinate separation by LCC is imperfect.

Overall, the LCC method can be easily implemented and has a unique solution independent from the size of the training set or the dimensionality. An extension to multi-class problems is straight forward. Different metrics can be used (not necessarily Euclidean) and, in general, no overfitting occurs. Although, LCC takes only first order moments into consideration, it behaves in some cases very well. There are many investigators who gave hints for the use of linear methods in the small sample case [131]. Furthermore, it is the basis for many other modern pattern recognition methods. Nevertheless, even if the data are linearly separable, the LCC hyperplane does not have to show this. The influence of outliers becomes greater for very small training sets. Instead of the mean values, more robust measures like the median can be used or another weighted combination of the training samples. The different training sets from cross-validation lead to differences in the mean values and therefore to a variation in location of the decision surface (direction of projection).

2.3 Linear Fisher Discriminant (LFD)

A very popular extension of LCC is the so called *linear Fisher discriminant*, first described in 1935 by Mildred Barnard at the suggestion of Roland A. Fisher. She applied this method to discriminate different flowers [79]. As in the previous case the linear map looks like:

$$g_{LFD}(z) \; = < \alpha, z > \; . \tag{2.2}$$

The direction of the projection vector α is determined by maximizing the square distance between the projected means of the two classes while minimizing the variance within each class [63]:

$$\alpha = argmax_w \; \frac{[w^T(\overline{x} - \overline{y})]^2}{\frac{1}{m}\sum_i [w^T(x_i - \overline{x})]^2 + \frac{1}{n}\sum_j [w^T(y_j - \overline{y})]^2} \; , \tag{2.3}$$

where $\overline{x}$, $\overline{y}$ represent the average response of each data set. Rewriting the right-hand side in matrix notation gives:

$$\alpha = argmax_w \; J(w) = argmax_w \frac{w^T S_B w}{w^T S_W w} \; , \tag{2.4}$$

with $S_B = (\overline{x} - \overline{y})(\overline{x} - \overline{y})^T$ the symmetric *between-class scatter matrix* and

$$S_W = \frac{1}{m}\sum_i (x_i - \overline{x})(x_i - \overline{x})^T + \frac{1}{n}\sum_j (y_j - \overline{y})(y_j - \overline{y})^T \; , \tag{2.5}$$

the symmetric *within-class scatter matrix*. This expression is well known in mathematical physics as the generalized Rayleigh quotient (denoted to John William Strutt, third Baron Rayleigh, 1842-1918). In signal theory, this criterion is also known as the signal-to-interference ratio. It is easy to show that a vector w that maximizes $J(.)$ must satisfy $S_B w = \lambda S_W w$, for some constant λ, which is a generalized eigenvalue problem. In our particular case, it is unnecessary to solve for the eigenvalues and eigenvectors of $S_W^{-1} S_B$ due to the fact that $S_B w$ is always in the direction of $(\overline{x} - \overline{y})$. In fact, the rank of the numerator matrix is at best one. Maximizing this criterion yields a closed form solution that involves the inverse of a covariance-like matrix. Since the scale factor of w is immaterial, we can immediately write the solution for the w that optimizes $J(.)$: $w = S_W^{-1}(\overline{x} - \overline{y})$. When the number of features (dimensionality) exceeds the number of training vectors, the sample estimate of the *within-class scatter matrix* will be a singular matrix with $rank(S_W) < N$. The smallest eigenvalues are estimated to be zero. Several methods to determine an optimal discriminant, when S_W is singular, have been described [262]. For example, the pseudo inverse will be used by [64; 196]. Chen et al. (2000) propose to use ordered eigenvectors [41]. I use a perturbation by the identity matrix to regularize the scatter matrix, which is fast and reliable and similar to [124] and [84]: $S_W(\beta) = (1-\beta) \cdot S_W + \beta \cdot I$ with $0 < \beta < 1$, in which the second term stabilizes the matrix S_W. In doing so, the influence of the small eigenvalues and the corresponding eigenvectors to the discrimination direction will be reduced at the expense of potentially increasing the bias [84]. The bias variance trade-off can be regulated by the parameter β, which controls shrinkage toward a multiple of the identity matrix. In the limit case (infinite number of samples), LFD gives the same linear separating decision surface as Bayesian maximum likelihood discrimination in the case of equal class covariance matrices. But, there is no reason to believe that the solution yields a separating vector in the small sample, linearly separable case. The computational complexity of finding the optimal w for the Fisher linear discriminant is dominated by the calculation of the within scatter matrix S_W and its inverse.

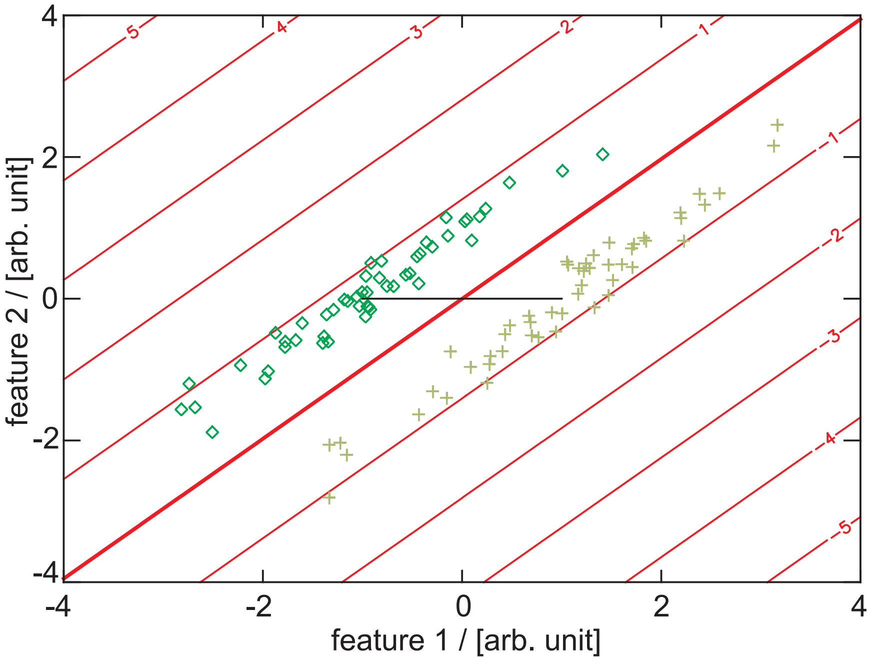

Figure 2.2: Projection according to the linear Fisher discriminant method (LFD) for a 2-dimensional sample. In contrast to LCC the separators are parallel to the first principal components of the single classes.

There are many connections to other state-of-the-art discriminant methods. For example, Shashua [222] has shown that the solution of LFD on the set of support vectors is normal to the hyperplane obtained by linear support vector machines (SVM). And Soumen [39] and Cooke [44] used this fact to build an iterative approximation to the linear SVM algorithm.

2.4 Regularized Discriminant Analysis (RDA)

Regularized Discriminant Analysis has been proposed by Friedman [84]. It can be understood as a mixture of Fisher's linear discriminant and a quadratic discriminant. The projection of a sample z in the two-class problem is given by the difference of the Mahalanobis distances:

$$\begin{aligned} g_{RDA}(z) &= \|\Sigma_1^{-1}\,(z-\overline{x})\|^2 - \|\Sigma_2^{-1}\,(z-\overline{y})\|^2 \\ &= (z-\overline{x})^T\,\Sigma_1^{-1}\,(z-\overline{x}) \;-\; (z-\overline{y})^T\,\Sigma_2^{-1}\,(z-\overline{y})\;. \end{aligned} \tag{2.6}$$

The sample covariance matrices are given by:

$$\Sigma_1 = \frac{1}{m}\sum_{i=1}^{n}(x_i-\overline{x})(x_i-\overline{x})^T, \quad \Sigma_2 = \frac{1}{n}\sum_{j=1^m}(y_j-\overline{y})(y_j-\overline{y})^T\;. \tag{2.7}$$

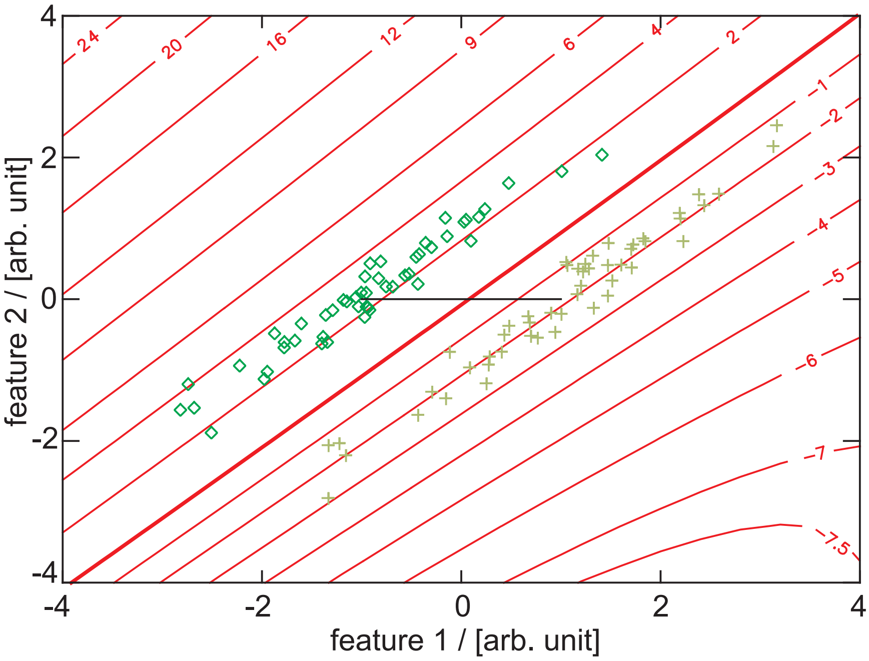

Figure 2.3: Projection according to the regularized discriminant analysis (RDA) for a 2-dimensional sample. The 'zero line' divides the two classes.

To overcome the singularity in the small sample case, Friedman proposed to use a linear combination of the class covariance matrix with the pooled covariance matrix Σ [113, chap.4.3]:

$$\Sigma_i(\lambda) = (1-\lambda)\cdot\Sigma_i + \lambda\cdot\Sigma\;, \tag{2.8}$$

and in addition shrinking the pooled covariance matrix itself toward the scalar covariance σI. A good pair of values for λ and σ can be obtained by cross-validation. Instead of the pooled covariance matrix, which is singular in the small sample case, I use the identity matrix I for regularization: $\Sigma_i(\lambda) = (1-\lambda)\Sigma_i + \lambda I$. The regularization parameter λ is established by the condition number of $\Sigma_i(\lambda)$ in order to guarantee a stable inversion. RDA is strongly related and motivated by Gaussian assumptions [63, chap.2.6]. This can be seen by the discriminant functions:

$$g_i(x) = -\frac{1}{2}(x-\mu_i)\Sigma_i^{-1}(x-\mu_i) - \frac{d}{2}ln(2\pi) - \frac{1}{2}ln(\Sigma_i) + ln(P(i)) \ , \tag{2.9}$$

with $i = 1, 2$ for two multivariate Gaussian densities, abbreviated by $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$ with prior probabilities $P(1)$ and $P(2)$. The decision surfaces are hyperquadrics which need not be simply connected. The numerical effort is twice as large as for Fisher's approach, where only one system of linear equations has to be solved. In addition, for RDA, two covariance matrices have to be stored, and their inverse has to be determined. Notice, the number of free parameters is higher in RDA than for FDA. Because the sample size is small compared to the dimensionality, even more parameters have to be shrinked by regularization.

2.5 Kernel Fisher Discriminant (KFD)

A very successful idea in pattern recognition is pre-processing. Mapping the data to another high – not necessarily finite dimensional – feature space can greatly simplify the task of separation. For example, data which are not linearly separable in the original feature space, might be linearly separable in an appropriate high-dimensional transformed feature space. Many possibilities exist for the choice of the transformation and the size of the feature space. Fortunately, for certain inner product feature spaces and corresponding mappings ϕ, there is a highly effective trick using kernel functions [168]. With kernel functions, scalar products can be implicitly computed without explicitly using or knowing the mapping [175]. Independent from each other several approaches use this idea to perform Fishers discriminant in a kernel feature space [16; 170; 201]. The *kernel Fisher discriminant* implements the Fisher linear discriminant in a feature space induced by a Mercer kernel [168], giving rise to a non-linear pattern recognition method. The projection of a test sample z can be expressed by KFD in the following form:

$$g_{KFD}(z) = \sum_{i=1}^{m} \alpha_i k(x_i, z) + \sum_{j=1}^{n} \alpha_{j+m} k(y_j, z) \ , \tag{2.10}$$

with kernel function: $k(x, y) = \phi(x) \cdot \phi(y)$. In the following, I use a Gaussian kernel with a dimension depending parameter c: $k(x, y) = exp(-(x-y)^2/c)$ according to [170]. For other possible kernel functions see, for example [214; 251]. Analogically to LFD, the weight vector α can be computed by maximizing a generalized Rayleigh quotient [172]:

$$J(w) = \frac{w^T M w}{w^T N w} \ . \tag{2.11}$$

The *between-class scatter matrix* can be computed from the samples by: $M = (M_1 - M_2)(M_1 - M_2)^T$ with $M_1 = \frac{1}{m} K \mathbf{1}_\mathbf{i}^\mathbf{1}$ and $\mathbf{1}_\mathbf{i}^\mathbf{1} = \mathbf{1}$ if i belongs to class C_1 and $M_2 = \frac{1}{n} K \mathbf{1}_\mathbf{i}^\mathbf{2}$ if i belongs to class

C_2. It can be proved that the *within-class scatter matrix* of the transformed feature space is given by: $N = K(I - v_1 v_1^T - v_2 v_2^T)K^T$, where I is the identity matrix and v_1 is a vector with elements $(v_1)_i = 1/\sqrt{n}$ if i belongs to class C_1 and zero otherwise. The corresponding formula holds for v_2. K represents the kernel matrix $(K_{ij})_{i,j=1}^{N} = k(x_i, x_j)$ between all training samples. Since the inverse of N does not exist, Mika et al. [170] proposed to add a multiple of the identity matrix I with a regularization parameter μ: $N(\mu) = N + \mu I$. Other forms of regularization are possible leading to solutions with different properties (see also Section 2.7). Simulations with different regularization parameters show that the influence of the parameter μ is, in most cases, not so critical (data not shown) [152]. In Chapter 3 and 4, the regularization parameter has been chosen to be: $\mu = 10^{-3}$.

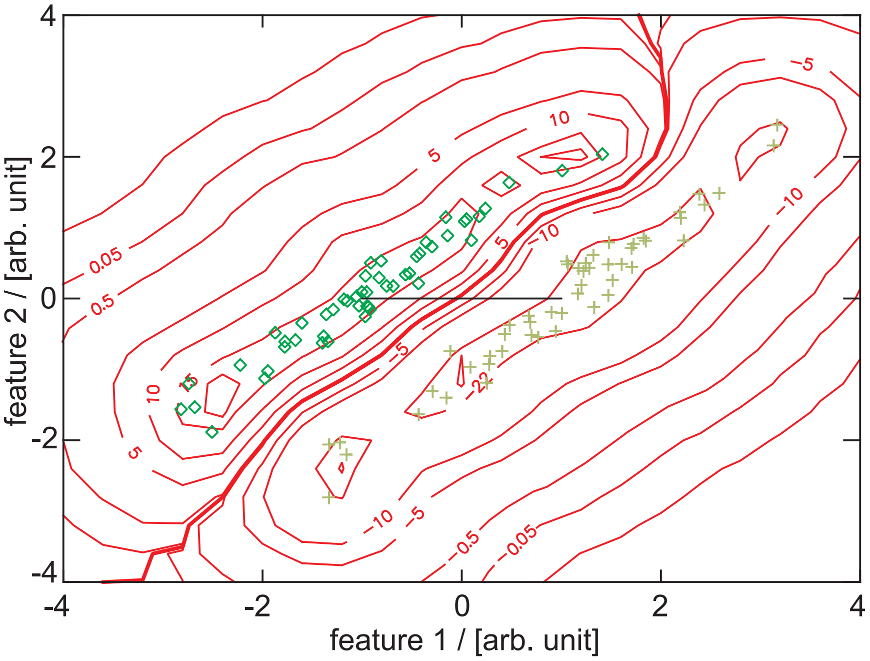

Figure 2.4: Projection according to the kernel Fisher discriminant method (KFD) for a 2-dimensional sample. The 'zero line' separates the two data sets. According to the Gaussian kernel, values far apart from the two data centres tend to zero.

Annotation: The now familiar *kernel trick* has been used to derive non-linear variants of many linear methods borrowed from classical statistics, e.g., principal component analysis [213], ridge regression [209], canonical correlation analysis [154], as well as more recent developments such as the maximal margin classifier, giving rise to the support vector machine family [45]. Kernel methods including SVM became very popular in recent years. Introductions into this field can be found in [214; 223; 252]. A disadvantage of these approaches is that a complex optimizing problem has to be solved (see also Chapter 6).

An illustration of the KFD behavior is given in Figure (2.4) where the projection to the input space is shown as contour plot. In a large collection of benchmark datasets, it has been demonstrated that KFD and its variants are capable of producing state-of-the-art results competitive to SVM or Adaboost [170] (see also Chapter 6). The similar performance of KFD and SVM is not surprising, since KFD has relations to SVM. While the SVM approach optimizes for a large minimal margin, in KFD, the average margin will be maximized. This property makes KFD very promising for the small sample case. In addition, it has been proved that KFD is equivalent to so called least squares support vector machines (LS-SVM) [234]. This can be shown via some complex mathematics, as in Gestel et al. [250], or by using the quadratic programming formulation [171]. Correspondingly to LFD, KFD is optimal (in the limit case) for data sets which are Gaussian distributed in the transformed feature space with equal covariance.

2.6 Approximate Distance Classifier (ADC)

The next method, has many similarities to nearest-neighbor classification. The *approximate distance classifier* is about a nonlinear projection that approximately preserves inter-class distances. It has been shown to be very successful in high dimensional data sets that strongly cluster [48]. The training set – called witness set by Cowen and Priebe – represents a template or prototype set, where each sample will be regarded as a characteristic realization of its class. As with the nearest neighbor function, ADC is simple to implement, and its applicability is not conditioned on the knowledge of the form of the underlying densities [130]. Irregular class boundaries can be represented with enough prototypes in the right places. In the two-class case a test sample z will be projected according to:

$$g_{ADC}(z) = \min_{i=1,..,m} (\|x_i - z\|) - \min_{j=1,..,n} (\|y_j - z\|) \ . \tag{2.12}$$

The main computational work is founded on sorting. Nevertheless, there are many fast algorithms for this problem, (e.g. [83]). In addition, in the small sample case this point has no influence. To quantify the proximity between an unclassified sample z and the templates from the two classes, the Euclidean distance is preferred. Implicit in this course of action, is the assumption that the class probability is roughly constant in the neighborhood, and no direction stands out [113, pp.427]. Generalizations have been made for example by [85] and [112] where the data structure is explicitly incorporated in the metric. In contrast to the other analysis algorithms described here, ADC stores all of the data in the relation instead of forming an abstract model of the data. Because it uses only the training point close to the query point, the bias of the ADC estimate is often low, but its variance is high [113, pp.417]. In contrast to the previous methods, ADC is a 'local' method.

Interestingly Cover and Hart found that half of the information for discrimination in an infinite collection of classified samples can be extracted from the nearest neighbor [46]. The asymptotic misclassification rate of the nearest-neighbor rule satisfies the condition: $E_{Bayes} \leq E_{1NN} \leq E_{Bayes}(2 - 2E_{Bayes})$. The inequality may be inverted to give [262, pp.95]:

$$0.5 - \sqrt{0.5}\sqrt{0.5E_{1NN}} \leq E_{Bayes} \leq E_{1NN} \ . \tag{2.13}$$

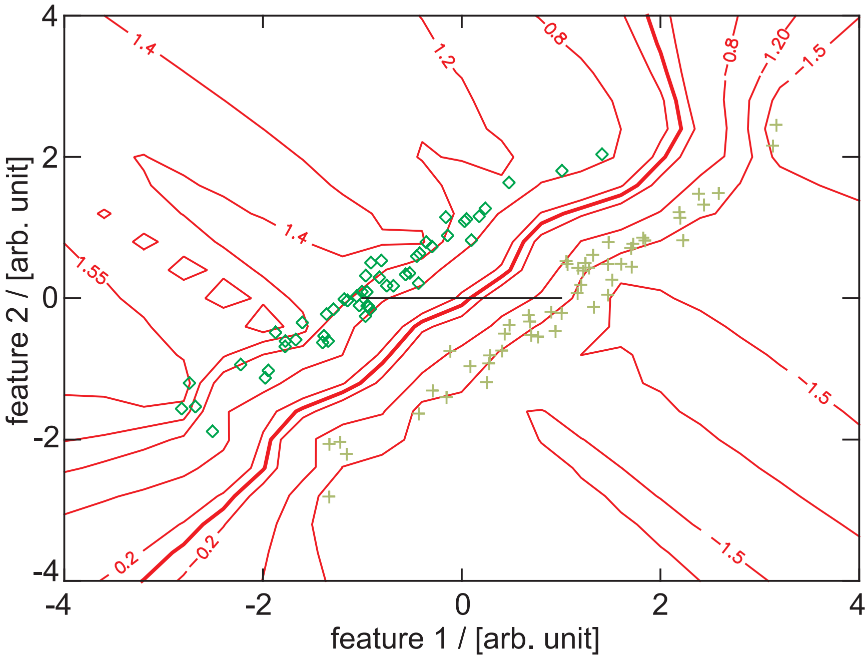

Figure 2.5: Projection according to the approximate distance classifier (ADC) for a 2-dimensional sample. The 'zero line' discriminates the two data sets. The shape of the contour plot is irregular and rugged.

The effects on the error rates of kNN rules of finite sample size data set have shown that the bias in the nearest neighbor error decreases slowly with sample size, particularly when the dimensionality of the data is high [88]. Despite this theoretical aspects nearest neighbor approaches have been successful in classification of ECG patterns and handwritten digits [113, pp.416]. As in the case of the nearest neighbor method a generalization in sense of averaging over k-samples can be easily constructed and used to prevent the method from over-fitting. The larger the value of k, the more robust is the procedure. Furthermore, the prototypes in our case are built by the training set. In general, this is not necessary.

2.7 Radial Basis Function (RBF)

The last projection method which I investigate is based on *radial basis function approximation*. RBF creates a continuous hypersurface from the training set, approximating the differences between the two classes [114, pp.256]. The shape of our RBF approach can be expressed in the same form as for KFD:

$$g_{RBF}(z) = \sum_{i=1}^{m} \alpha_i k(x_i, z) + \sum_{j=1}^{n} \alpha_{j+m} k(y_j, z) \ . \tag{2.14}$$

The kernel, called *radial basis function* $k(r), r \geq 0$, can be for example Gaussian $k(r) = exp(-r^2/\sigma)$, multiquadric $k(r) = (r^2 + c^2)^{1/2}$ or anything else (see [190]). For further descriptions of the RBF method, see [34; 125]. Possible applications in neuroscience, for example to ECG data or local field potential recordings from the visual cortex of awake monkeys, can be found in [136; 137; 146].

The weight vector $\alpha = (\alpha_1, ..., \alpha_N)^T$ is given by the following interpolation condition: $A\alpha = b$ with $A = (k(x_i, x_j))_{i,j=1}^N$ and $b = 1$ if x_i belongs to class C_1 resp. $b = -1$ if x_i belongs to class C_2. According to the choice of the weight vector, $g_{RBF}(\cdot)$ represents a continuous hyper-surface mapping the training vectors from class C_1 exactly to the value $+1$ and the vectors from class C_2 to the value -1. Therefore, points nearby a sample from class C_1 (C_2) become positive (negative).

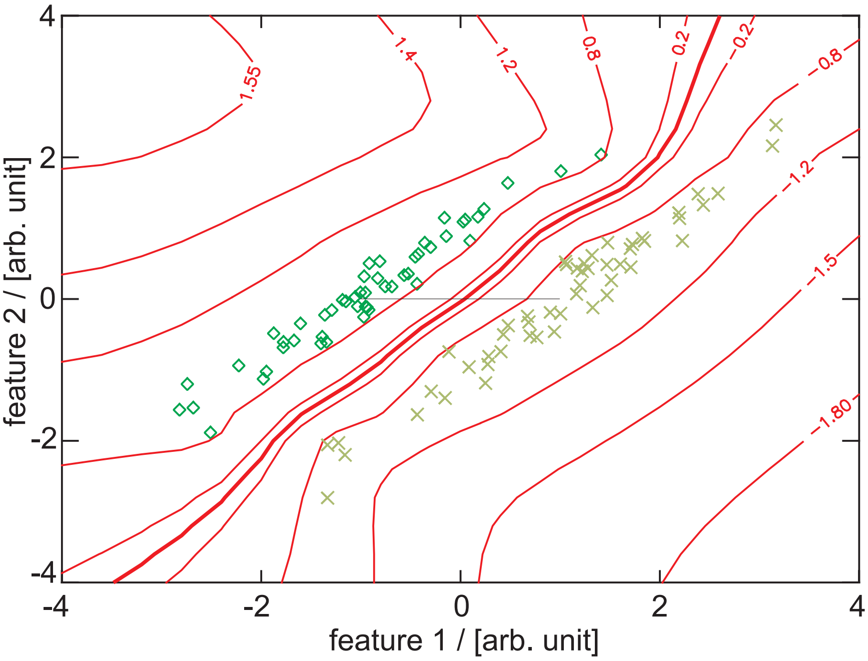

Figure 2.6: Projection according to radial basis function approximation (RBF) for a 2-dimensional sample. The 'zero line' divides the plane into two disjunct parts. According to the multiquadric kernel, RBF performs some sort of normalization. All training points of class C_1 will be mapped to the value $+1$. The training points of class C_2 will be mapped to the value -1.

The weights are uniquely determined, since it has been shown that for a large class of basis functions $k(r)$ and under very weak conditions on the geometry of the centres x_i – provided they are distinct – the symmetric distance matrix A is positive definite, and therefore non-singular [169; 190]. The performance will be influenced by the smoothness (complexity) of the underlying densities. In high-dimensional spaces, this assumption is justified, for the most part (see Section 1.4). In addition, no regularization is necessary for the matrix inversion. For solving the linear system of equations, I use a preconditioned conjugate gradient method [15]. Other methods may be as well possible ([18; 217]. Because of numerical as well as segregation aspects, the choice of the appropriate basis function, the shape parameter c, and the dependence on the number of the centres x_i, is very important. An unsuitable choice can lead to a badly conditioned, fully occupied matrix A, with poor discriminative behavior. With regard to the amount of training data, which is restricted in this work to the small sample case, the size of the matrix A plays a minor role. Many research groups use Gaussian kernel functions, but I have found by extensive simulations, that the multiquadric $k(r) = \sqrt{r^2 + c}$ and even the Euclidean distance $k(r) = r$ provide mostly better results (data not shown) leading as well to moderate condition numbers. As in the previous case, the training vectors

$\{x_k \in \mathbb{R}^d |\ k = 1, .., N\}$ take over the role of a prototype set, but there are other interpretations possible. RBF can be regarded as a global method with a distributed representation (Equ. 2.6). Although, the multiquadric kernel function is not local, this approach shows a local behavior (see Fig. 2.6). Therefore, the negative influence of outliers or a bad choice of training samples are strongly diminished. RBF has many connections to the previously described methods.

1) At first, let me explain why the radial basis function approximation can be regarded as an extension of the linear correlation classifier. For this the linear correlation classifier is written in the from: $\varphi(z) = \sum_{i=1}^{m} \frac{1}{m} \quad < z, x_i > - \sum_{j=1}^{n} \frac{1}{n} \quad < z, y_j >$, assuming that $z \in \mathbb{R}^d$ is a new response that has to be classified with respect to C_1 or C_2. Restriction to the case of normalized responses $\|v\| = \|x_i\| = \|y_j\| = 1$ and Euclidean distances makes it obvious that a radial basis function Φ can be regarded as a nonlinear function f of the correlations between z and x_i: $\Phi(\|v - x_i\|) = \Phi(\sqrt{2(1 - < v, x_i >)}) = f(< v, x_i >)$. Therefore, the corresponding radial basis function, given by: $s(v) = \sum_{i=1}^{n} \lambda_i \ f(< v, x_i >) + \sum_{j=1}^{m} \mu_j \ f(< v, y_j >)$, is a nonlinear extension of the linear correlation classifier (Equ. 2.1). Comparing the Equations (2.1) and (2.6) shows that the linear correlation classifier uses only first order properties. Correspondingly, the coefficients in (Equ. 2.1) are constant and depend only on the number of samples. In contrast, the coefficients α_i and α_{j+m} in the RBF expansion (Equ. 2.6) depend both on the data and the mapping.

2) Trained with only one sample x_1 and y_1 from each class the function $g_{ADC}(\cdot)$ is identical to $g_{RBF}(\cdot)$, provided the basis function is given by: $k(x, y) = ||x - y||$. In general, the solutions show a unequal behavior as can be seen from Figure (2.5) and (2.6).

3) RBF has a direct relation to KFD. Depending on the kernel function and the regularization, it can be shown that the weight vectors of RBF and KFD are the same (up to a scale factor). For this the kernel function in Equation (2.4) has to be the same as in Equation (2.6) and instead of regularization with the identity matrix $I = KK^{-1}$ the invertible matrix KK^T has to be used. If we order the samples like: $\{x_1, ..., x_m, x_{m+1}, ..., x_N\}$ defining $x_{m+j} := y_j$, which can be done without loss of generality, then the between-class scatter matrix M can be written in the form: $M = K \cdot (\frac{1}{m} I_0^1 - \frac{1}{n} I_1^0)$ with $I_0^1 \in \mathbb{R}^N$ equal to one in the first m elements and $I_1^0 \in \mathbb{R}^N$ equal to one in the last n elements. With this convention, the RBF regularized KFD vector w can be computed according to: $[K(I - \frac{1}{m} I_0^1 I_0^{1T} - \frac{1}{n} I_1^0 I_1^{0T} + \sigma I) K^T] w = K(\frac{1}{m} I_0^1 - \frac{1}{n} I_1^0)$. The solution vector v of the linear system of equations: $[(1 + \sigma) I - \frac{1}{m} I_0^1 I_0^{1T} - \frac{1}{n} I_1^0 I_1^{0T}] v = (\frac{1}{m} I_0^1 - \frac{1}{n} I_1^0)$ is given by: $v = 1/\sigma \cdot (1/m, ..., 1/m, -1/n, ..., -1/n)^T$. While doing so, the entry $1/m$ belongs to the first m components and the entry $1/n$ belongs to the remaining n components. For equally sized samples $n = m$ the vector v is identical to the vector b up to the scale factor σ/m. As I have mentioned previously, the scale factor is irrelevant for the estimation of the Bayes error and the Shannon information. The only point that is important in this context is that our kernel density estimation approach might be influenced by an inappropriate scale factor (see Chapter 3).

3 Numerical Comparison

'The traditional machinery of statistical processes is wholly unsuited to the needs of practical research. ...The elaborate mechanism built on the theory of infinitely large samples is not accurate enough for simple laboratory data. Only by systematically tackling small sample problems on their merits does it seem possible to apply accurate tests to practical data.' (Fisher, R.A. Statisitical Methods for Research Workers. Oxford, 1970)

In the following, I compare the six dimension reduction methods introduced in the previous Chapter. For this purpose, various artificial data sets have been generated. In doing so I took different properties into consideration which are relevant for the analysis of real multi-channel signals. Great emphasis lies on the performance in high-dimensional spaces, and different training sizes. Altogether, the numerical comparison reveals no unique picture. However, with respect to numerical qualities, RBF behaves in a very robust and predictable manner.

3.1 Uniform Data

In all cases I use uniform distributed samples. The uniform distributions have the advantage that Bayes error and Shannon information can be computed analytically. With regard to the peculiarities of uniform pdfs I refer to [58] and to Section 1.4 of the first Chapter. For the uniform random number generation I apply the same pseudo-random algorithm as before [78]. It has been shown, that this algorithm is reliable also in the multivariate case. In general, I have found that the type of the data (uniform, exponential, Gaussian, etc.) has a minor effect on the properties of our methods, in contrast to other parameters, e.g., the training size or the dimensionality of the data (see also Section 3.6). Since neurophysiological multi-channel recordings generate high-dimensional data sets, I vary the dimensionality between 2, 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100. Furthermore, the training set has been restricted to 10, 20 or 50 samples per class and the prior probabilities have been chosen equally: $P(1) = P(2) = 0.5$. The training data are randomly sub-sampled out of a set of 1000 samples per class (see also Section 1.6). This procedure ensures that resources were concentrated in regions of higher data density [33]. The samples that are not used for the model construction will be used to test the performance and to approximate the Bayes error or the Shannon information. The huge number of observations has the advantage of reducing bias effects for the density and the entropy estimation. According to the results in Section 1.7, I use a kernel density estimation approach for the Bayes error approximation. The bin width will be estimated by the L1 reference method. Since the projected data are one-dimensional, I employ the spacing method of Vasicek for entropy estimation, as described in Section 1.8. In order to reduce the influence of the training set, the subsampling is repeated twenty times. In addition, 20 Monte

Carlo simulations have been performed, leading to 400 distance values per method and dimension. Therefore, the standard error of the arithmetic mean does not exceed a value of 0.009. In most cases its variation is about 0.005. The mean values, corresponding to the largest training sets (50 samples per class), are twice as accurate as for the smallest training set (10 samples per class). The regularization parameter of LFD is 10^{-5}, for KFD $\lambda = 0.001$, and for RDA, the regularization factor is 0.15, on average. The first nearest neighbor $k = 1$ has been used for ADC and for the RBF map, the multiquadric basis has been reduced to the Euclidean distance.

Overall, four different artificial data sets have been investigated. Figure (3.1) gives you an impression of the different data constellations in two-dimensions. From a neurophysiological perspective, each feature space (dimension) may be identified with a separate recording channel. Correspondingly, a d-dimensional random sample (signal) $x = (x_1, x_2, ..., x_d)$ may be identified with the mean activity recorded with d electrodes in parallel evoked by on of the two stimuli C_1 or C_2.

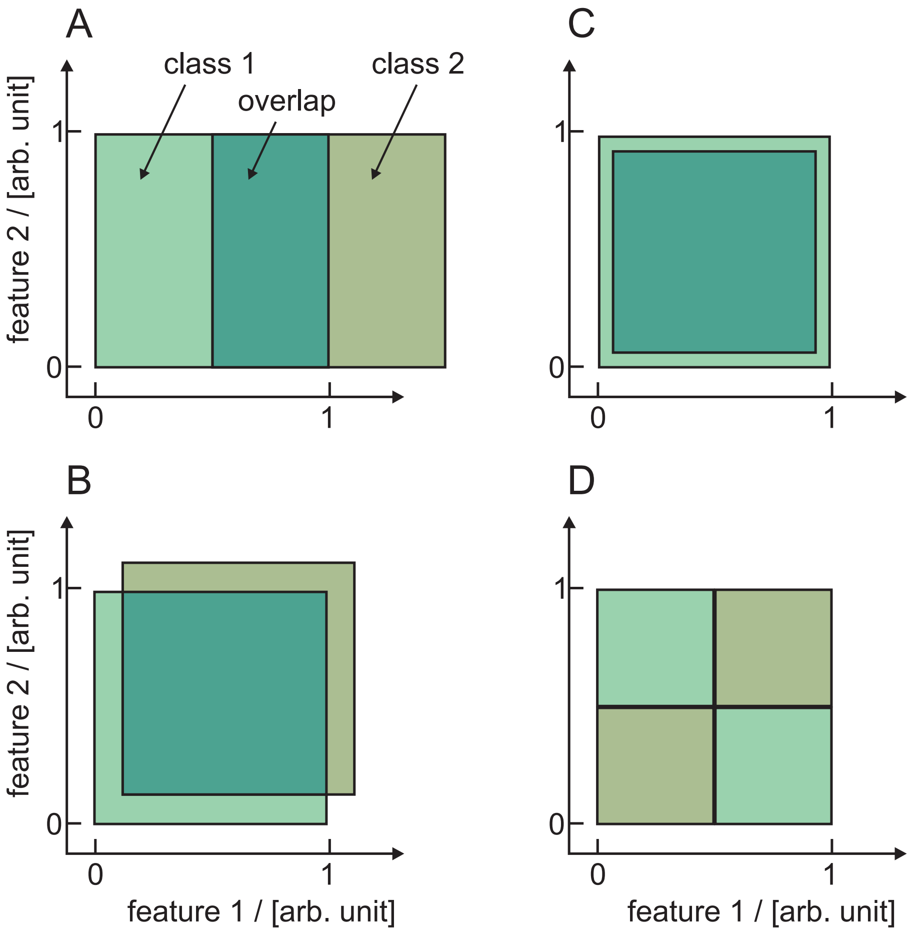

Figure 3.1: Two-dimensional illustration of the different data constellations used for the numerical comparison of the six projection methods from Chapter 2. (**A**) The difference appears only in the first dimension. (**B**) A slight difference exists in all feature spaces with a mean shift of 0.1 for class C_2. (**C**) The two data sets differ in the variance. (**D**) The two data sets have different correlation structures. For a detailed description see the text below.

3.2 Difference in One Component

The first artificial data set consists of two uniform distributions with one of them shifted in one dimension to the right (Fig. 3.1A): $C_1 = U(0,1)^d$, $C_2 = U(0.5,1.5) \times U(0,1)^{d-1}$. The difference in the first dimension could be interpreted as a stimulus-specific effect increasing the mean activity if stimulus C_2 is presented (Section 3.1). The signals of all other $d-1$ electrodes are either equally effected or unaffected by the two stimuli. Correspondingly, increasing the number of recording channels (signals) has no effect on the discrimination. At the opposite, investigating the signals from more recording channel reduces the ratio between relevant and irrelevant signal components, in this situation. Theoretically, Shannon information will be: $I = h(X) - 0.5 \cdot h(X|1) - 0.5 \cdot h(X|2) = -0.5 \cdot log_e(0.5) \approx 0.346$ nat, independent of the investigated dimension. The same is true for the normalized Bayes error which will be 0.5 for equal priors.

Looking at the simulation results (Fig. 3.2), you can see that for all methods, independent of dimensionality, the approximated Bayes error is always greater than 0.5 which is in accordance with the theory (Section 1.5). At the same time, the theoretical mutual information builds an upper boundary for all methods. In two dimensions all projection methods behave similar with slight deviations from the theoretical value. In general, the deviations increase with dimensionality. Thereby, the behavior of the linear correlation classifier and the radial basis function approach is comparable. For ADC and KFD, the error increases faster than for RBF or LCC, with increasing dimensionality. Only LCC, KFD, and RBF indicate a significant difference ($p < 0.01$) for all dimensions and all training sets. On the other hand LFD, RDA, and ADC indicate, in high dimensions, a worse segregation and a lower information content, which is not significant. ADC, LCC, KFD, and RBF show a consistent performance. For larger training sets they become better but only LCC, RBF, and KFD become significantly better if the training size increases from 10 to 50 samples per class. For the linear discriminant function of Fisher peaks appear. These are related to the ratio between the size of the scatter matrices and the dimensionality. In the literature, this phenomenon is called *peaking* or *Hughes effect* [64; 126]. RDA shows a similar behavior, but the peaks are not so distinctive. An interesting property is the fact that in the high-dimension setting ($d > 80$) the segregation of RDA and LFD, using a larger training set (50 samples per class), is worse compared to the small training set (10 samples per class). On the basis of the different boxplots it can be seen that the variability over the 400 sub-samples is relatively similar for all methods. The standard deviation shows no dependence to the dimensionality.

For this data constellation, the linear correlation classifier shows the highest segregation, followed by the radial basis functions and the kernel Fisher discriminant function. This is not astonishing, since the optimal segregation hyperplane can be obtained from the difference between the mean values of the two classes. The linear correlation classifier should be preferred as well from a numerical point of view. The condition number, which is a sign for problems with the inversion of a linear system of equations, varies for RBF from 10 to 10.000. The condition number of the regularized KFD scatter matrix is much higher – up to 10^6, especially for small dimensions. For LFD and RDA, the condition number increases with increasing dimensionality from 10 to 10^4 (10^6).

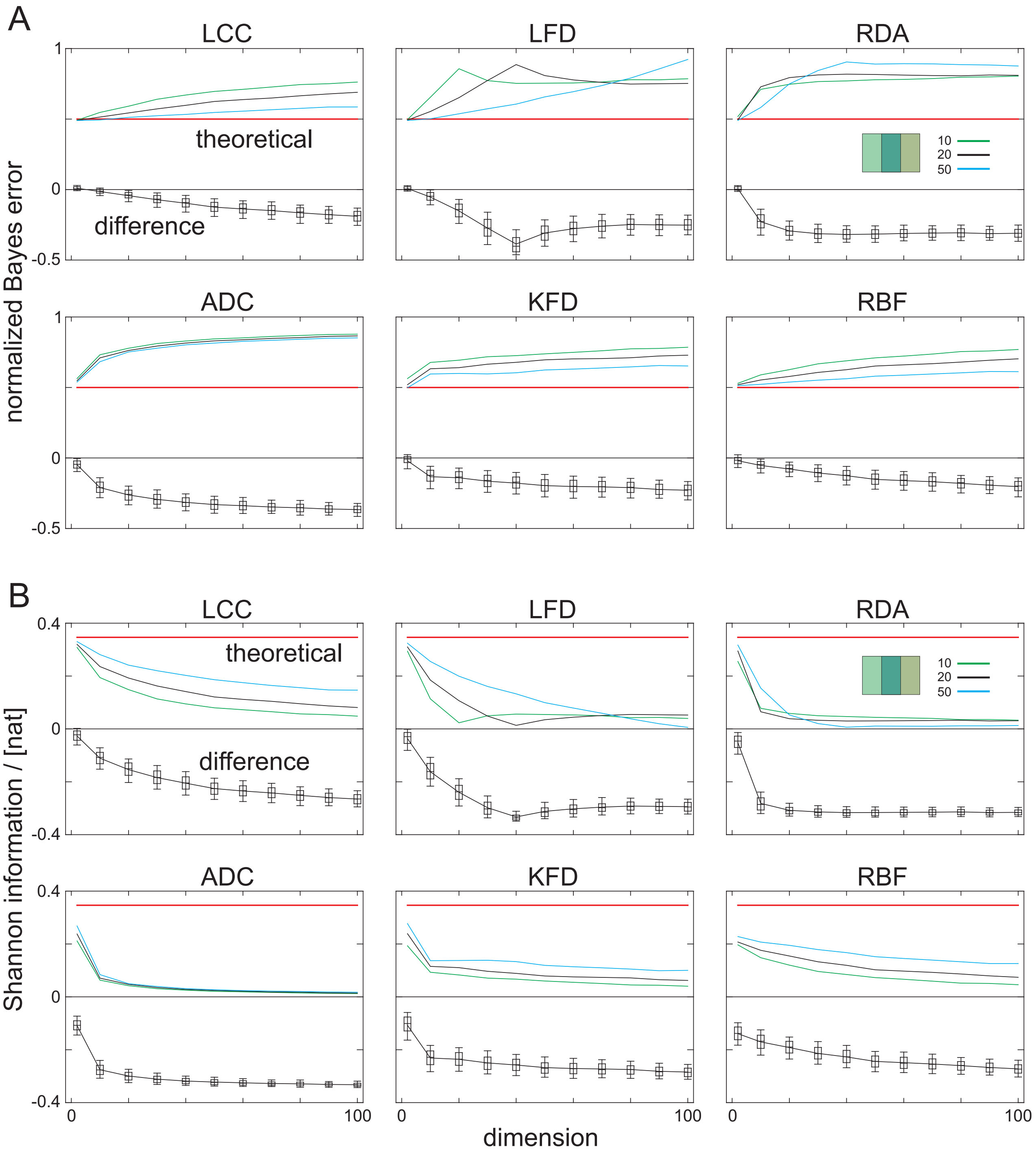

Figure 3.2: Performance of the six projection methods in the small sample case for data with a difference in one component. Sample size for projection parameter estimation varies between 10 (green), 20 (black) and 50 (blue) samples per class. (**A**) The upper six plots belong to the normalized Bayes error (overlap area). (**B**) The lower six plots belong to the Shannon information given in natural units. Red curve: theoretical Bayes error (Shannon information). The boxplots with 10, 25, 50, 75, and 90 quantiles display the deviation of the 20 sample estimation curves from the theoretical curve (red).

Since the Bayes error increases with dimension for most of the methods, this development indicates that most of the recording channels carry no information about the two stimuli. So, the dimension reduction methods can be used to discover the relevant signal components, working as a pre-processing tool in this special case (see also Chapter 6).

3.3 Difference in All Components

The second artificial data set consists of two uniform distributions which can be partially distinguished, since the samples from the second class have a small shift in all dimensions to the right (Fig. 3.1B): $C_1 = U(0,1)^d$, $C_2 = U(0.1, 1.1)^d$. Applying the same neurophysiological interpretation as before (see Section 3.1), then the shift in the second distribution may indicate, a spatially more expanded stimulus influence, in which all recording channels show a stimulus specific difference. Although each signal component is conditionally assigned, the signals of the different recording channels are statistically independent. Similar to the previous situation (Section 3.2), the variation of the signals for different trials is the same for both classes and is therefore not stimulus specific. In contrast to the previous data, the number of signal components has a pronounced effect onto the separability. Therefore, using the signals from more recording channels in parallel, reduces the overlap between the two distributions (see also Section 1.4). This can be seen from the evaluation of the theoretical Bayes error. Theoretically, the normalized Bayes error decreases to zero with growing dimension, according to: $d_{Bayes} = 0.9^d$. Correspondingly, Shannon's information measure increases with dimension: $I = ln(2) \cdot (1 - 0.9^d)$, reaching the maximal value of $ln(2) \approx 0.693$ nat for infinitely high dimension. The question is, what dimension reduction method will reveal this information increase and preserves it?

Figure (3.3) indicates that most of the methods, analyzed here, indicate an increase of the transmitted information with higher dimensions, although the deviation from the theoretical value is smallest in the 2-dimensional setting. LCC, ADC, KFD, and RBF show a consistent behavior. The Bayes error goes down with larger training samples. As in the previous setup, the LFD and RDA performance is not monotone. For example, the Bayes error obtained from Fisher's linear discriminant method using 20 samples per class, becomes largest if the dimension of the scatter matrix coincides with the dimension of the feature space. At its maximum the deviation from the theoretical value is about 80 %. In addition, the variability for the 400 different sub-samples per training size and dimension is higher for LFD than for the others. Except for the LFD approximation, with 50 training samples per class, all methods indicate a significant distance ($p < 0.01$) between the two data sets, analyzing signals in a 50 dimensional space. As in the previous test situation (Section 3.2) LCC, KFD, and RBF provide the best results, and the condition number of RBF, LFD, RDA, and KFD is in the same range. The smallest spread was found for the linear correlation classifier method, followed by the radial basis function approach.

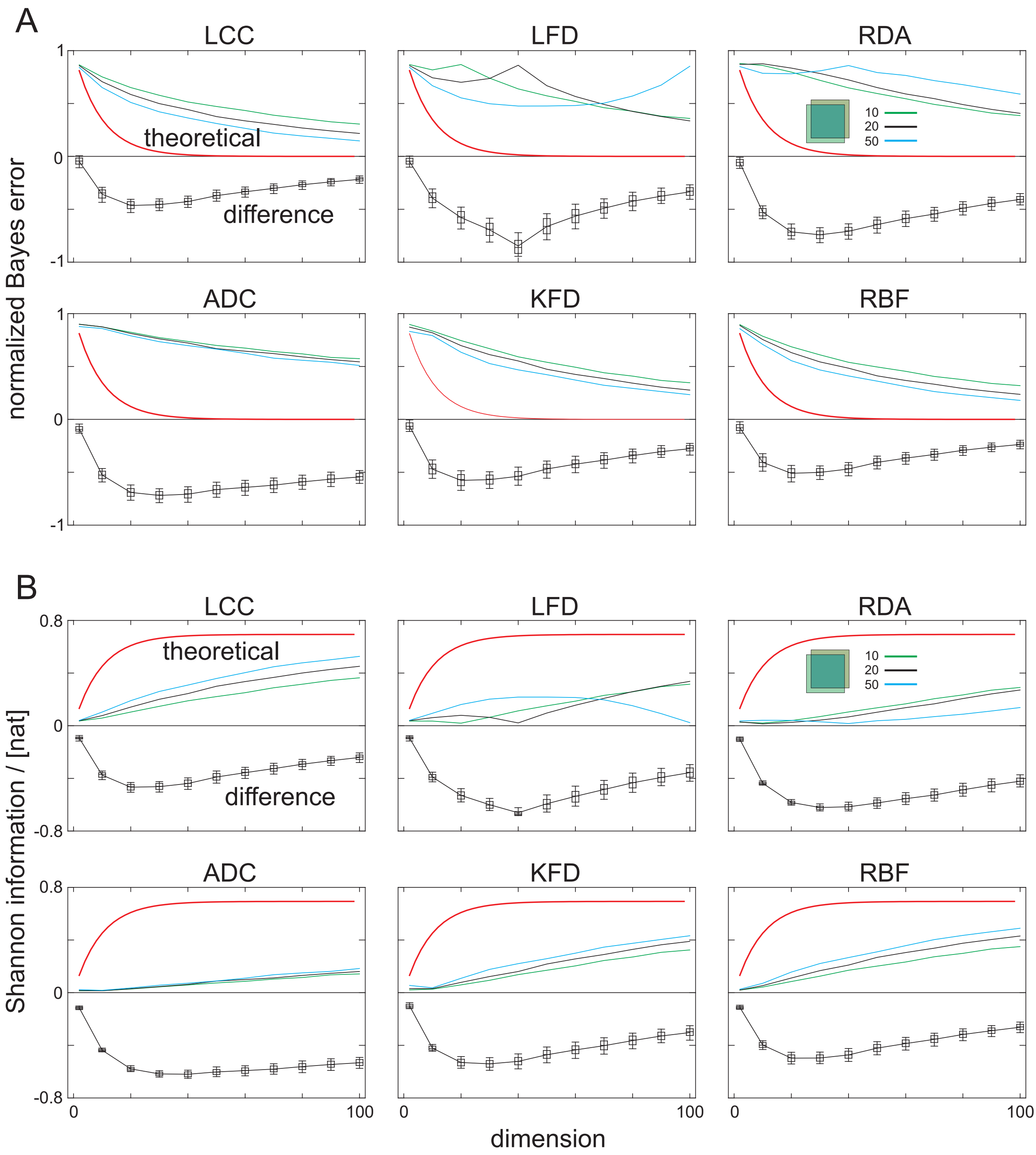

Figure 3.3: Behavior of the six projection methods (see Chapter 2) applied to two data sets with a small difference in all components. (**A**)The upper six plots show the approximation of the normalized Bayes error. (**B**) The lower six plots illustrate the performance of the methods used to approximate the Shannon information. The theoretical curves are red. The boxplots show the difference between the theoretical curve and the approximation according to a training size of 20 samples per class.

3.4 Differences in the Noise Level

The distinction in the data of Section 3.2 and 3.3 was based on a difference in the centres of the two classes. In the following artificial data sets the mean signals are equal, but the variances are different (Fig. 3.1, C): $C_1 = U(0,1)^d$, $C_2 = U(0.1, 0.9)^d$. The noise level of the signals in the second class is smaller. As in the previous tests, analyzing each feature space separately, reveals a poor discrimination. However, using the high-dimensional information at once should lead to better results. As before, the normalized Bayes error draws nearer to zero: $d_{Bayes} = 0.8^d$ as the dimension increases. At the same time, Shannon information reaches a maximum with increasing dimension: $I = 0.5 * ln(2)(1 - 0.8^d) - 0.8^d pln(p) - 0.5dln(0.8)$ with $p = 0.5(1 + \frac{1}{0.8^d})$. Due to the correlation between the two distance measures (Bayes error and Shannon information) and the dimensionality, one might expect that the six methods show a similar behavior (Section 3.3).

Interestingly, this is not the case (see Fig. 3.4). Not only the linear Fisher discriminant method, but also the linear correlation classifier shows a low performance. The estimated Bayes error is inappropriate, except for low-dimensional data sets. An explanation for the poor performance of the two linear approaches can be given: First, the two data sets are not linearly separable, and second, the mean values are equal, which means that an unknown sample will be projected to zero by both methods. So, what we get by these methods is approximately what we can reach using only one-dimensional information. In contrast, the kernel methods are better adapted to this kind of data combination. The performance of the kernel Fisher method is especially high. For high-dimensional distributions ($d > 80$), KFD comes very close to the theoretical value. The information increase achieved by radial basis functions is not so steep, but it shows a continuous profit with increasing training size. For KFD this increase is not so pronounced. On the contrary, for the training sets with 50 samples per class, KFD becomes slightly worse. Overall, the discrimination indicated by RBF and KFD is always significant ($p < 0.01$), whereas the values of LCC and LFD are not significant. RDA and ADC, which perform slightly better than the linear projection methods, indicate a significant difference only for the large training sets. For such constellations, KFD should be preferred, followed by RBF.

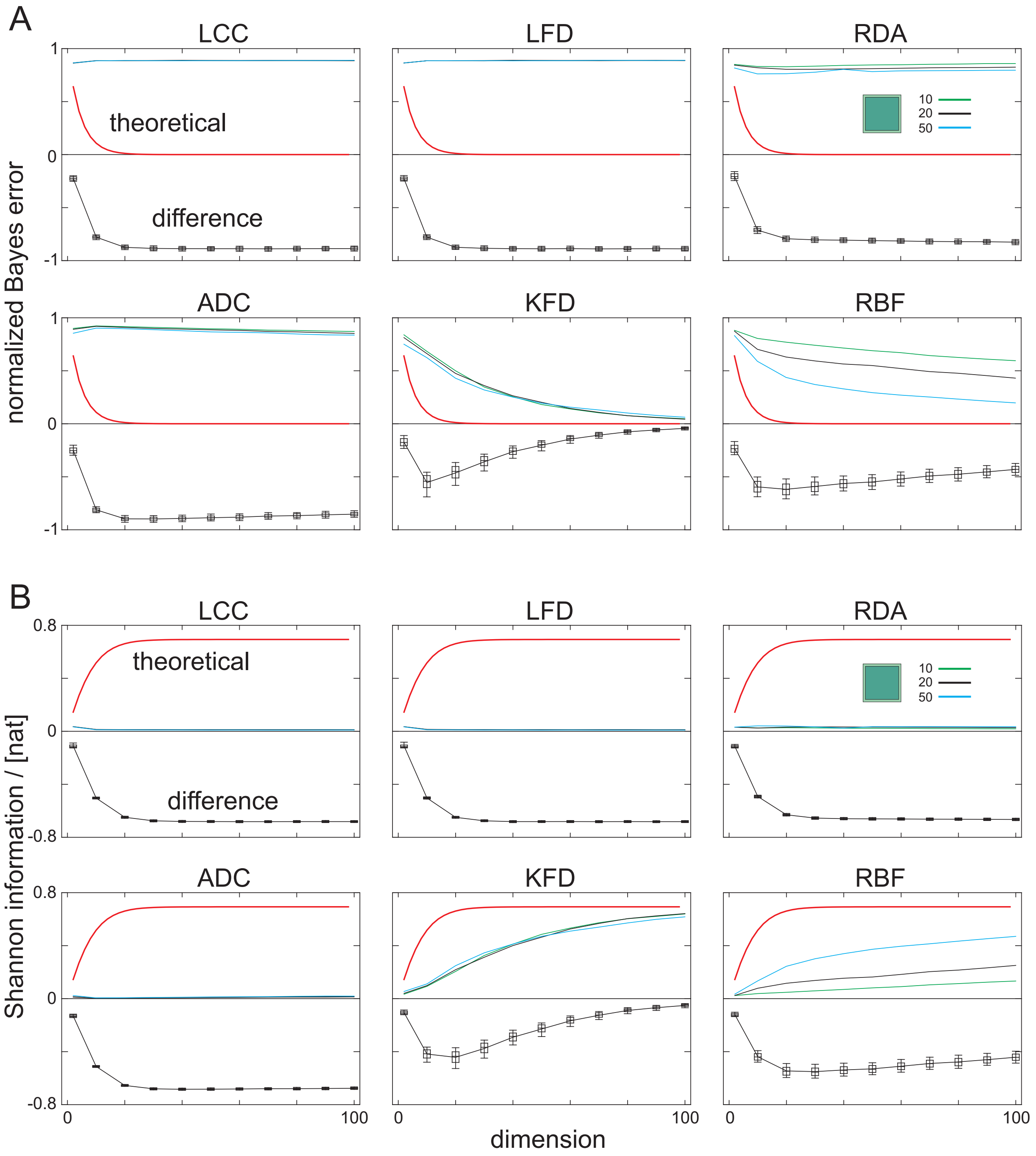

Figure 3.4: Differences in signal variability. (**A**) The upper six plots show the normalized Bayes error (theoretical curve in red) and the approximations of the different methods, depending on the dimensionality of the data and the training size. (**B**) The lower six plots illustrate the same curves for the Shannon information.

3.5 Nonlinear Separable

The last artificial data set consists of two nonlinear separable distributions with no overlap. A chessboard arrangement has been constructed (Fig. 3.1D), which is a popular benchmark in statistical learning and pattern recognition, [113, pp.255; 223, pp.102]:

$C_1 = 2 \cdot [U(0, 0.5), U(0.5, 1)] + 2 \cdot [U(0.5, 1), U(0, 0.5)] \times U(0, 1)^{d-2}$ and

$C_2 = [2 \cdot U(0, 0.5)^2 + 2 \cdot U(0.5, 1)^2] \times U(0, 1)^{d-2}$ with $d \geq 2$.

Applying the same interpretation as in Section 3.1, the signals from two recording channels are correlated. Within one class the mean activity at two recording channels is synchronized. Within the other class the mean activity of the corresponding recording channels is anti-correlated. If the activity is low within one electrode it is high in the other. The signals obtained from all other recording channels are uncorrelated and stimulus independent. Since there is no overlap, the theoretical Bayes error is always zero: $d_{Bayes} = 0$. Correspondingly, Shannon information is maximal: $ln(2)$ nat. How do the different methods behave?

The results of the discriminant analyses are illustrated in Figure (3.5). Both linear approaches indicate a significant difference for low dimensions but with a huge variability (see the corresponding boxplots). This is interestingly, because the mean values of the two distributions remain the same, independent of the dimensionality. The reason for the behavior of the linear methods lies in the fact that the overlap differs with the projection direction. ADC and RDA provide the smallest Bayes error and the greatest Shannon information in the high-dimensional space. This can be explained as follows: ADC is appropriate with clustered data. Due to the *curse of dimensionality*, most of the data are concentrated to the edges (see also Section 1.4). RDA performs well, because the class covariance matrices differ in this situation (see Section 2.4). RBF and KFD are not as good as ADC and RDA. Whereas, ADC and RDA indicate a significant segregation up to 100 dimensions, RBF and KFD indicate no significant discrimination for samples with more than 40 components. Both approaches show some kind of over-smoothed behavior. On the other side, all methods become worse at higher dimensions, which is a destructive effect of the increase in uncorrelated stimulus independent signal components.

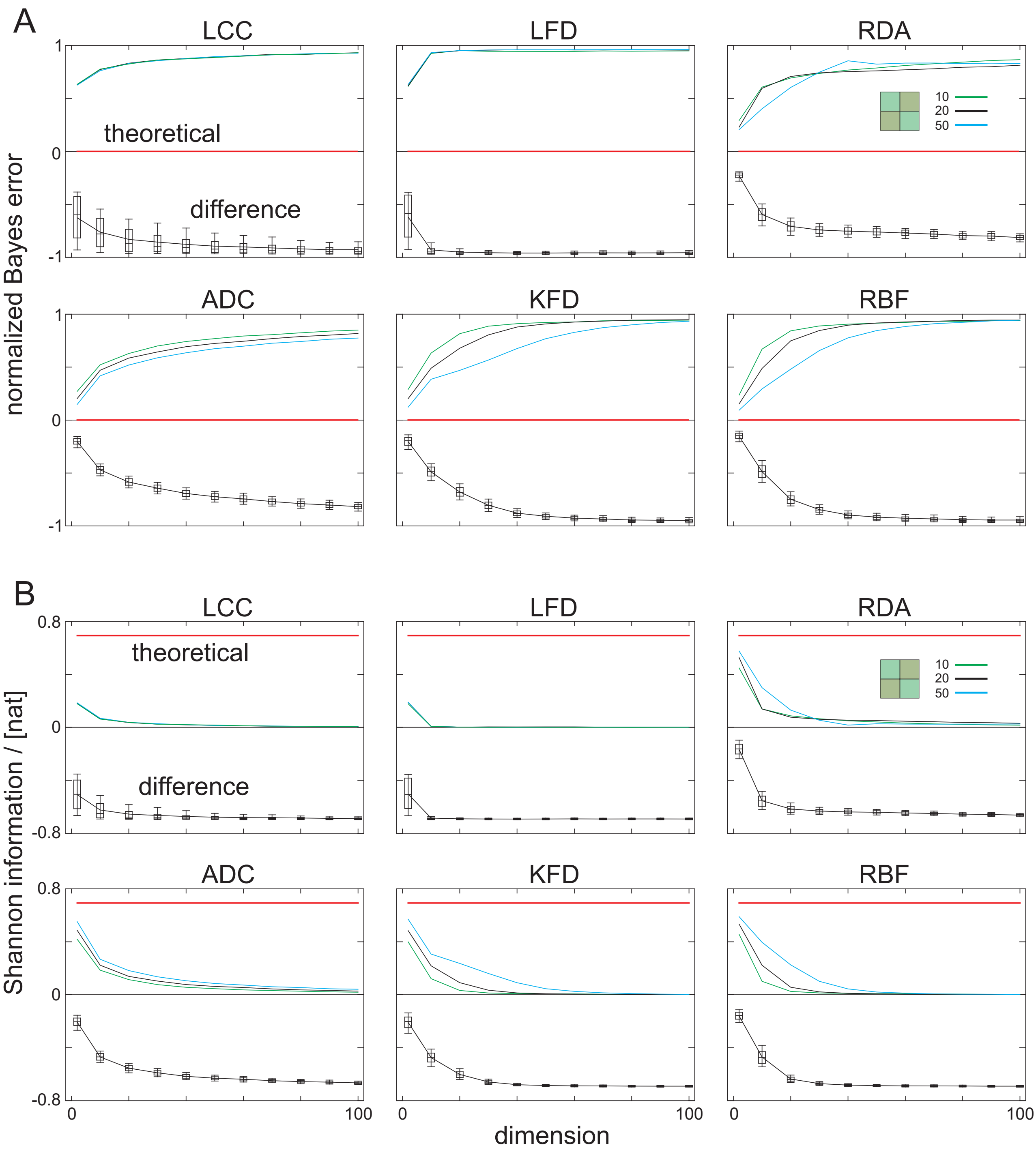

Figure 3.5: Approximation of the normalized Bayes error and the Shannon information for two nonlinear separable data sets. (**A**) The upper six plots show the results for the Bayes approximation. (**B**) The lower six plots illustrate the results for the Shannon information approximation. See also the explanations in the previous figures.

3.6 Discussion

Referring to the four test sets, no method is generally superior to the others. Although the nonlinear methods have a higher potential they are not better in every situation. Linear approaches are appropriate for data sets characterized by a difference in the mean, even if the discrimination boundary is not a linear hyperplane. Whereas in two or three dimensions the differences between the six methods are minor and independent of the investigated artificial test sets, they differ significantly at higher dimensions.

In most situations, the difference between the theoretical information and the information estimation after the projection (information loss) increases with dimensionality. This phenomenon is mainly attributed to the *curse of dimensionality*. Correspondingly, the number of learning samples per class plays an important role. While the performance of LFD, RDA and KFD becomes worse for larger training sets in certain circumstances, this has not been found for LCC, ADC and RBF which show a consistent efficiency. An explanation for this phenomenon has to be attributed to the regularization. In addition, LFD and RDA show a non-monotone behavior, depending on the ratio between the number of training samples and the dimensionality. Therefore, LFD and RDA should be handled with care if the same analysis has to be done for data sets of different dimensionality. In particular, in the case where the Bayes error is very high and the Shannon information is very low, contradictions can appear (see also Chapter 6). Although, in the present analyses no contradiction has been found, I cannot exclude that in the small sample case the low-dimensional approximation underestimates the true error.

Altogether, the variability of the six projection method becomes smaller with more training samples. A duplication of the training set from 10 samples to 20 samples per class reduces the variability (10-90 % Quantile) to half. The comparison of the six projection approaches reveals that the radial basis function approximation makes the most profit from larger training sets. From this, it is clear that an appropriate use of the data, e.g., by k-fold cross-validation may improve the segregation significantly, and makes it possible to determine confidence intervals. According to the shape of the function $g(\cdot)$ in Chapter 2, the six mappings have different projection properties. The discriminant functions of LCC and LFD have a simple structure, followed by RDA, RBF, and KFD. The discriminant surface of the nearest neighbor approach is rough, since ADC is a local method. With regard to the six mappings, a slight advantage of RBF over the other projection approaches comes from the normalization (Section 2.7).

The six investigated methods are similar from a numerical point of view. ADC and LCC are easy to implement. For the remaining methods, a linear system of equations has to be solved. RBF has a unique solution if the training samples are different. The solutions of the other methods are affected by the regularization. A trade-off between stability and performance has to be chosen for the regularization parameter. The poor outcome of LFD can be ascribed to the symmetry in the data constellations. Applying the Rayleigh quotient to 'asymmetric' distributions would lead to better results (see Section 2.3). This is also true for RDA.

I have done some studies with Gaussian distributed data, too. These results (not shown) are comparable to the above presented (Section 3.2 to 3.5). From this, I conclude that the shape of the two probability density functions has a minor effect on the error and information estimation, in contrast to other variables, e.g., the training size, the overlap or the dimensionality of the data. The performance of the different methods becomes worse with increasing overlap of the two distributions [163, pp.10]. The six projection methods are different in their parametrization. Especially, the kernel methods belong to a huge class of projection functions. Because of that, I have done tests with various kernel functions in the KFD and RBF setting. From this study, I found that KFD with polynomial kernels is not as good as with the parameterized Gaussian kernel. Within the RBF context, the separability can be enhanced by the multiquadric factor. Beside this parametric generalization, some researchers favour the combination of different methods [227]. In contrast to the six projection methods, described in Chapter 2, these strategies are much more time consuming. I will come back to this point in Chapter 6.

In summary, the results provide a highly sophisticated picture. With regard to multi-channel cortical recordings I cannot say which methods should be preferred. There are evidences in terms of robustness and efficiency to favour the RBF approach, but, at the same time, other methods, like LCC or KFD partly reveal better results. In order to obtain more realistic data sets, I propose to apply the different projection methods to artificial neuronal signals. For this, a neuron model has been developed in Chapter 4.

4 Neural Network Modeling

'Scientists have broken down many kinds of systems. They think they know most of the elements and forces. The next task is to reassemble them, at least in mathematical models that capture the key properties of the entire ensembles.' (Wilson, E.O. Consilience: The unity of knowledge. London, 1998)

After a short overview about cortical modelling concepts, I introduce a simple neuron-like model, which incorporates important functional aspects of real nerve cells. Afterwards, this simple neuron model is expanded to build small neural networks with different types of connection. The corresponding cortical-like signals serve as test set for the different projection approaches. Emphasis will be placed on the signal-to-noise ratio, the temporal dynamics, and differences in the coupling structure.

4.1 Historical Overview

The brain is under investigation for more than two thousand years, leading to many speculations about its function and its working strategy. For example, Aristoteles had the view of a brain acting as a heat exchanger in order to cool down the overheated heart and in Descartes' opinion the brain could be explained mechanically, with the pineal gland being the centre of our consciousness [180, pp.30]. Today we know that these assumptions are wrong and that the brain is a much more complex organ.

The complexity arises in three different ways. 1.) It can be found in the differentiated branching structure of individual nerve cells and the elaborated connection patterns between neurons. [1] 2.) Aside the *anatomical complexity*, *functional complexity* is associated with the reaction of the neural system. Physiological observations indicate that neurons are incorporated in functional circuits or modular units, e.g. [235]. Small circuits consist of two or three neurons and form the building blocks used in constructing larger modular units. The cue point of this interacting network is a complex and intertwining cooperation including *feedback* and *feedforward* control loops. 3.) *Dynamic complexity* can be identified with the ability to react to the same stimulus in different ways. Our brain does not behave in a static manner. Neither its functional behavior nor the underlying neural connections are fixed and unchangeable [73]. [2] For the generation of artificial neural data, I concentrate on functional aspects of small neural networks. Therefore, I develop a neuron model placed on a mesoscopic level, [138]. This model has no explicit spatial extent, and

---

[1] The human nervous system is composed of $10^{10} - 10^{12}$ neurons [179]. Typically each neuron receives inputs from $10^3 - 10^4$ neurons and in turn sends information to large numbers of neurons [265].

[2] A detailed introduction into the organization and function of the brain and all the different models lies outside the scope of this work. Therefore, I refer to the vast literature and the references therein [12; 98; 135; 179].

reflects a level of abstraction similar to that of the integrate-and-fire model [228] with relations to mean field modeling [8; 158; 267]. Although the model is highly simplified, it captures important dynamic features of neural information processing.

4.2 Mathematical Description

A neuron can be divided into an input area (synapse, dendrite, soma) and an output area (axon-hillock, axon) (Fig. 4.1). In the neuron model the signal transmission between these two areas will be fully described by four functions V, U, θ and I. The function V describes the output signal of the neuron transmitted to other neurons. The U measurement can be identified as the voltage across the cell membrane at the axon-hillock. The θ measurement characterizes the temporal dynamics of the threshold and the I measurement represents random fluctuations therein referred to as noise (see Section 4.2.3).

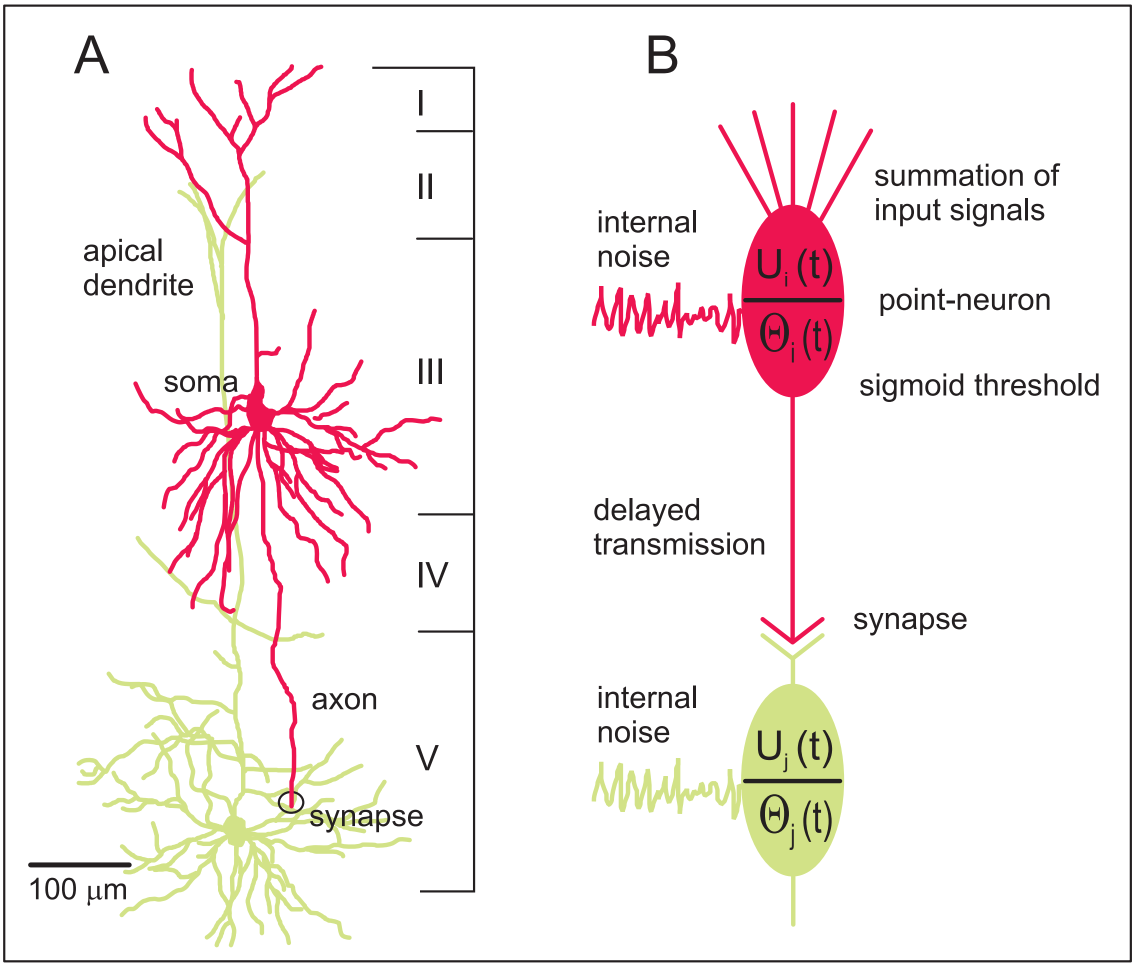

Figure 4.1: (**A**) Depiction of two pyramidal cells with synaptic connections extracted from the primary visual cortex (Layer III to Layer V) after [242, pp. 674]. (**B**) Schematic representation of the reduced neuron model including internal noise, a sigmoidal threshold mechanism and a delayed signal transmission (for further description see text below).

4.2.1 Neural Output Signal

The output signal is calculated as the difference between U and θ:

$$V_i(t) = S(U_i(t) - \theta_i(t)) \ , \tag{4.1}$$

with $S(\cdot)$ a monotone ascending function. In the following, $S(\cdot)$ has a sigmoid shape: $S(x) = \frac{1}{1+e^{-\alpha x}}$, with $S : \mathbb{R} \rightarrow (0,1)$. The positive parameter α, adjusts the gradient of the sigmoid. In the limit ($\alpha \gg$), $S(\cdot)$ has the form of a step function. For $\alpha \ll$ the sigmoid $S(\cdot)$ converges to a straight line. The derivation of the sigmoid is given by: $\frac{dS(x)}{dx} = \frac{\alpha \cdot e^{-\alpha \cdot x}}{(1+e^{-\alpha \cdot x})^2} = \alpha \cdot S(x) \cdot (1 - S(x))$. In contrast to real neurons, in which information is transmitted by patterns of impulses, the output $S(U - \theta)$, describes the probability of a neuron to transmit an impulse pattern [98]. The reason for selecting this model is motivated by the nature of the experimental data analyzed in Chapter 5, and by the assumption that sensory areas of the cortex process information primarily by changes in neural population densities, e.g. [3; 166; 186].

4.2.2 Membrane Voltage

The membrane voltage $U_i(t)$ is made up of the input signals $V_j(t)$ from other neurons and internal fluctuations $\xi_i(t)$, which I will describe in more detail in the next Section. Mathematically, the signals from other neurons will be low-passed filtered and weighted by a constant factor w_{ij}:

$$U_i(t) = \xi_i(t) + \sum_j \int_{-\infty}^{t-\delta_{ij}} V_j(\tau) w_{ij} h_{ij}(t - \tau - \delta_{ij}) d\tau \ . \tag{4.2}$$

By using this low-pass filter and factoring in the delay parameter δ_{ij}, various properties of the synapses and the dendrites are incorporated and consolidated into simpler forms. The impulse response of a single synapse has the form: $h(t) = \frac{a \cdot b}{b-a} \left[e^{-a \cdot t} - e^{-b \cdot t} \right]$ with $0 \leq t$ and 0 otherwise. The rise time and the decline time are given by the two parameters: $a \in \mathbb{R}^+$ and $b \in \mathbb{R}^+$. The area under the impulse response equals one: $\int_0^\infty h(t) dt = 1$. The synaptic transfer function is given by: $H(f) = \frac{a \cdot b}{b-a} \left[\frac{1}{a+2\pi i f} - \frac{1}{b+2\pi i f} \right]$. By adjusting the time constants to physiological values, results generated by this model can be made to approximate the behavior of real neurons needed in the present test for the comparison of the projection methods (Fig. 4.2).

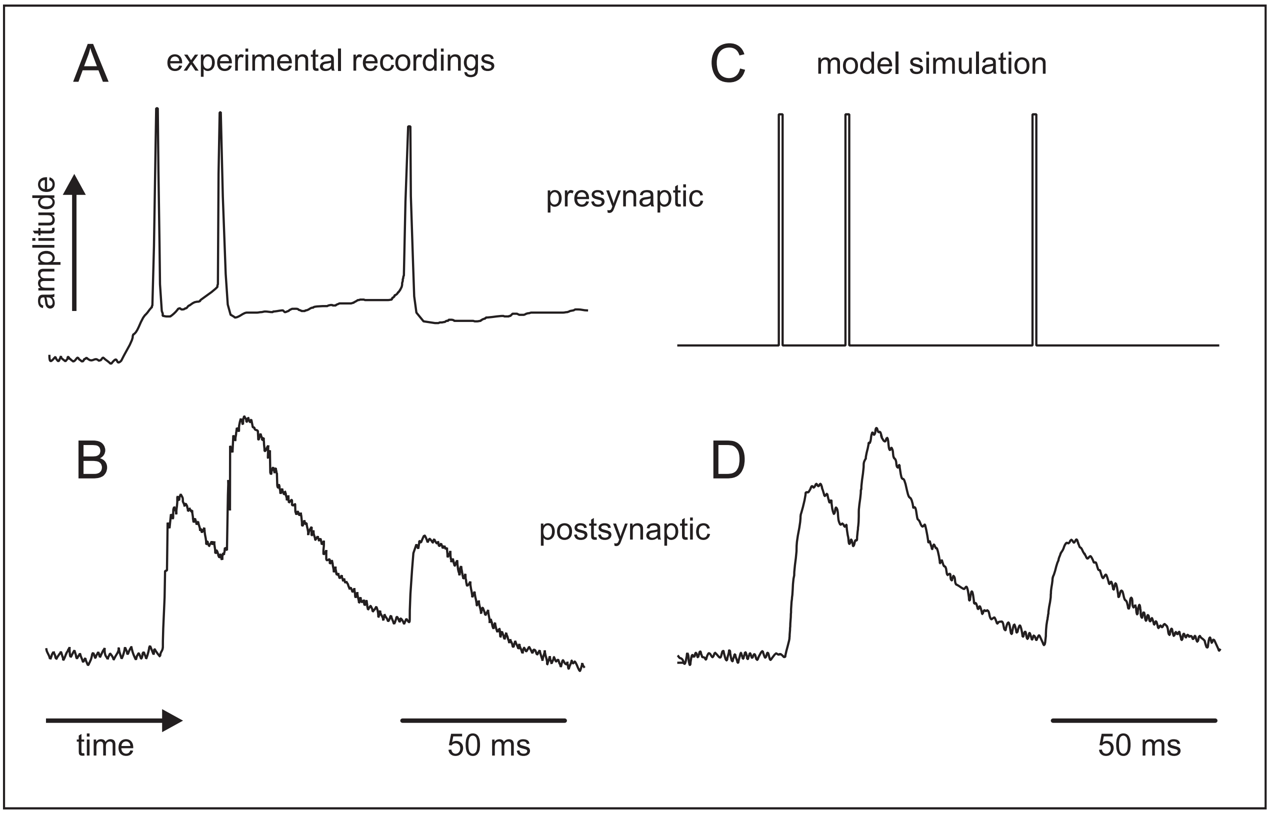

Figure 4.2: (**B**) Sequence of three excitatory postsynaptic potentials (EPSP) generated by a brief sequence of three presynaptic spikes (**A**) broadcasted by a layer III pyramidal cell to a layer V pyramidal cell. Redraft from [242, pp.677b]. (**D**) Response of the model neuron to three rectangular pulses of 1 ms duration (**C**) with the same inter-spike intervals as in (**A**). The stepsize has been fixed to h = 0.02 ms (see also Appendix A.1).

4.2.3 Noise

The temporal dynamic of cortical nerve cells is highly corrupted by random fluctuations therein referred to as noise. Roughly, one can distinguish between intrinsic noise sources evoked, e.g., by the random kinetic of ion channels and extrinsic noise sources created by network effects [98; 161]. Especially, the uncorrelated continuous strong barrage of synaptic inputs creates a highly fluctuating intracellular membrane voltage, with a pronounced effect on the behavior of individual neurons [28; 53; 204]. Due to the very dense synaptic connectivity in the cortical network, this ongoing spontaneous neuronal discharge activity will increase the trial-to-trial variability of the neural response to repeated presentations of a stimulus. Furthermore, recent studies on the effects of noise in nonlinear dynamical systems have shown that novel behaviors can arise, concerning for example the filter properties of individual neurons [10], or the information transmission [38; 160]. Noise is therefore an essential part of the cortical dynamic and is therefore incorporated in my neuronal network simulations as well. In the model, I implement a bandwidth-limited colored noise process $\xi_i(t)$ with a Lorentzian power spectrum up to 500 Hz added to the membrane voltage in an additive statistical independent manner: Other noise sources, e.g., random changes of the parameter setting or thermal fluctuations are ignored. It should be mentioned as well that the random process is independent of the stimulus.

4.2.4 Threshold

In real neurons, an action potential will be released if the membrane voltage U exceeds a certain threshold level. At the molecular level a fast inflow of sodium-ions and a temporal shifted slower outflow of potassium-ions proceeds [123]. Subsequent to this concentration change, the membrane voltage will increase for a short time (≈ 1 ms) and then fall below the resting potential. After this rapid change, the activation of another action potential will be inhibited or reduced for some time in most neurons. Consequently, the probability to emit an impulse is reduced. In my model, I simplify this highly nonlinear process and instead of changing the membrane voltage, the threshold value is manipulated. Mathematically, the temporal dynamic of the threshold is given by:

$$\theta_i(t) = \theta_{io} + \int_{-\infty}^{t} S(U_i(t) - \theta_i(t))\epsilon(t-\tau)d\tau \; . \tag{4.3}$$

The actual value of the threshold function is composed of a constant threshold offset and the filtered output signal. Thus, the impulse response is given by: $\epsilon(t) = \frac{c}{\gamma}e^{-t/\gamma}$ for $t \geq 0$ and 0 otherwise. The two parameters γ and c are used to control the inactivation and therefore how fast the threshold changes if the membrane voltage U changes. The parameter γ basically determines the decay of the exponential function. The parameter c regulates the amplitude of the threshold giving: $\int_0^\infty = \epsilon(t)dt = c$. The transfer function for the descending exponential function has the form: $E(f) = \frac{c}{1+2\pi i f\gamma}$.

To summarize, this system of equations constitutes an analog mathematical abstraction of a physical neuron consisting of a nonlinear system with a low-pass filter, a dynamical threshold, internal fluctuations and a delay for the axonal and synaptic transmission.

4.3 Numerical Computation of the Neuron Dynamic

4.3.1 Threshold Computation

For the numerical evaluation of the threshold function Equation (4.3) will be transfered into a first order ordinary differential equation:

$$\theta'(t) = \frac{1}{\gamma}(\theta_o - \theta(t)) + \frac{c}{\gamma} S(U(t) - \theta(t)) \ . \tag{4.4}$$

This nonlinear equation can be successfully evaluated by an implicit or at least linear implicit approach. Four different strategies have been tested. In all cases, the partial derivation $J = df/d\theta$ has to be computed, where the right hand side has been summarized in the form:

$$f(t, \theta) = \frac{1}{\gamma}(\theta_o - \theta(t)) + \frac{c}{\gamma} S(U(t) - \theta(t)) \ . \tag{4.5}$$

The partial derivation of f with respect to θ at the point (t_n, θ_n) is given by:

$$D_\theta f(t_n, \theta_n) = -\frac{1}{\gamma} I - \frac{c \cdot \alpha}{\gamma} S(U(t_n) - \theta_n) \cdot (1 - S(U(t_n) - \theta_n)) \ . \tag{4.6}$$

A pleasant property of the Jacobi matrix is based on its diagonal character. Thereby, within complex network simulations the multidimensional nonlinear system may be transformed in a system of scalar problems which have to be solved numerically. The first method that I have tested is the well known linear implicit Euler method [108, pp.138]. Numerical integration gives:

$$(I - h \cdot J)(\theta_{n+1} - \theta_n) = f(t_n, \theta_n) \ . \tag{4.7}$$

The second method which has been applied successfully with such models is the exponential Euler method, given by [122]:

$$(\theta_{n+1} - \theta_n) = h \cdot \phi(h \cdot J) \cdot f(t_n, \theta_n) \ , \tag{4.8}$$

with $\phi(z) = \frac{e^z - 1}{z}$. The third method that was tested is the second-order Rosenbrock method, given by the following system of equations [258]:

$$\begin{aligned} (I - h\sigma J)\, k_1 &= f(t_n, \theta_n) \\ (I - h\sigma_{21} J)\, k_2 &= f(t_n + c_2 h, \theta_n + h\alpha_{21} k_1) + h\sigma J k_2 \\ \theta_{n+1} &= \theta_n + h b_1 k_1 + h b_2 k_2 \ , \end{aligned} \tag{4.9}$$

with $\sigma = 1 + \frac{\sqrt{2}}{2}$ and $J = D_q f(t_n, q_n)$, $1 - b_1 = b_2 = 1/2$, $\sigma_{21} = -\sigma/b_2$, $c_2 = \alpha_{21} = 1/(2b_2)$. For the matrix $(1 - h\sigma J)$ we get: $(1 - h \cdot \sigma J) = (1 + \frac{\sigma \cdot h}{\gamma}) \cdot I + \frac{h \cdot \sigma \cdot \alpha \cdot c}{\gamma} \cdot S \cdot (1 - S)$. The last method that was tested, is the second-order two-step Adams-Moulton method [108, pp.143]:

$$\theta_{n+1} = \theta_n + \frac{h}{12}\left(5 \cdot f(t_{n+1}, \theta_{n+1}) + 8 \cdot f(t_n, \theta_n) - f(t_{n-1}, \theta_{n-1})\right) \; . \tag{4.10}$$

For the approximation of the value θ_{n+1} I take the solution of a first-order method (see Equ. 4.7 and 4.8). Different tests have been made to compare the four methods. The performance of the four discretization approaches, induced by a periodic variation of the membrane voltage, is illustrated in Figure (4.3).

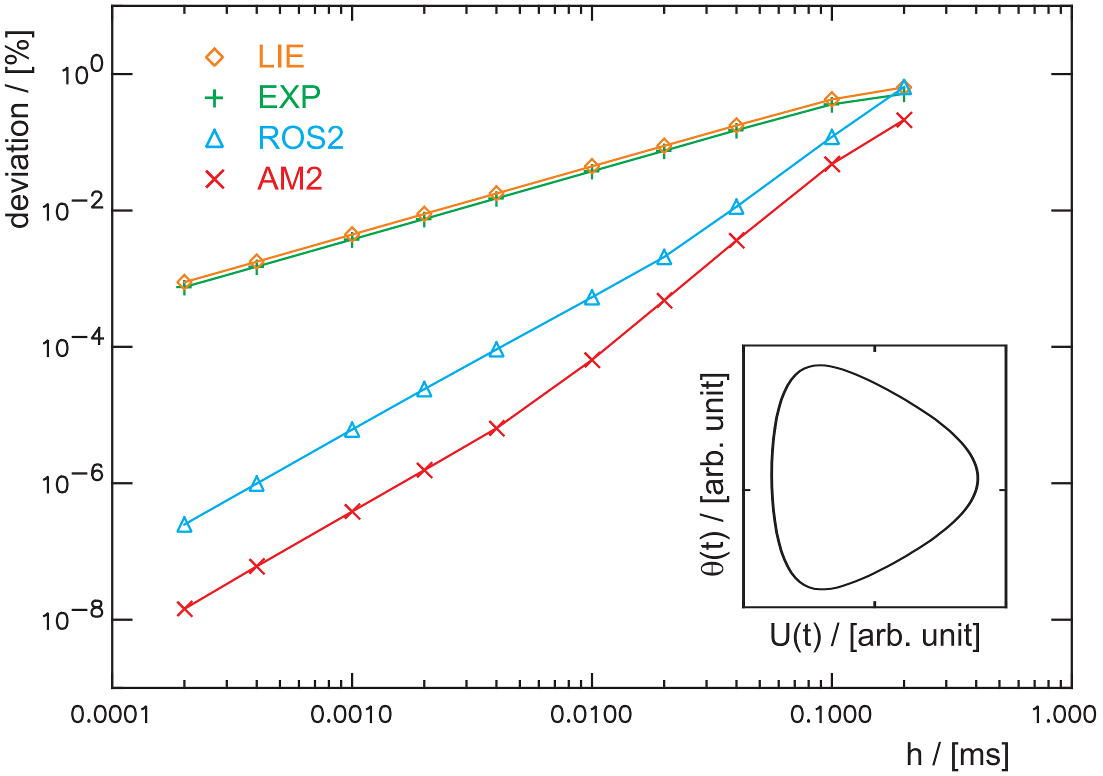

Figure 4.3: Maximal deviation between the computed threshold dynamic and the true variation ($\%$) for different step sizes h. LIE = linear implicit Euler, Exp = exponential fitted Euler, ROS2 = second-order Roenbrock, AM2 = two-step Adams-Moulton. Parameter setting: $\theta_o = 0.1\ mV$, $\alpha = 1\ mV^{-1}$, $\gamma = 50\ ms$, $c = 180\ ms$, $\theta(0) = 5.15881\ mV$. Inset: Periodic variation of the threshold originated by a sinusoidal membrane voltage $U(t) = \pi \cdot sin(2 \cdot \pi \cdot t/T)$ with $T = \ 1\ ms$.

As expected, the gradient of the two first-order methods is similar. The same is true for the second-order methods. However, the gradient of the second-order methods is steeper. The two-step Adams-Moulton method performs best, indicated by the steepest decrement with decreasing step size h.

4.3.2 Membrane Voltage Computation

At first, the integral equation of the membrane voltage (Equ. 4.2) is differentiated two times. Applying a mathematical transformation to this equation results in the following second-order delayed differential equation:

$$\begin{aligned} U''(t) + (a+b)\cdot U'(t) + a\cdot b\cdot U(t) = \\ \xi''(t) + (a+b)\cdot \xi'(t) + a\cdot b\cdot \xi(t) + a\cdot b\sum_{j=1}^{n} w(t)\cdot V_j(t-\delta)\ . \end{aligned} \tag{4.11}$$

Using appropriately selected coefficients, the left hand side of the equation becomes a conventional equation for a damped harmonic oscillator. The first part on the right hand side describes the deviation of internal fluctuations linearly superimposed onto the membrane voltage (see Section 4.2.3 and 4.3.3). The last term on the right hand side describes the filtered input from other neurons transmitted via synapses. Various computational methods are employed for numerical computation. The first method is the two-step difference method, in which the first and second derivation will be approximated by:

$$U'(t_n)) = \frac{U(t_{n+1}) - U(t_{n-1})}{2\cdot h}, \quad U''(t_n)) = \frac{U(t_{n+1}) - 2U(t_n) + U(t_{n-1})}{h^2}\ . \tag{4.12}$$

Applying this rule to the delay differential equation (Equ. 4.11) results in:

$$\frac{U(t_{n+1}) - 2\cdot U(t_n) + U(t_{n-1})}{h^2} + (a+b)\frac{U(t_{n+1}) - U(t_{n-1})}{2\cdot h} + a\cdot b\cdot U(t_n) = f(t_n)\ , \tag{4.13}$$

where the right hand side has been summarized into $f(t_n)$. Resolving the equation for $U(t_{n+1})$ is straightforward. The method is easy to implement and has a second-order error decay. A disadvantage of this scheme is the restriction to an equidistant array. Beside this discretization method I have tested two second-order Runge Kutta methods. These are the blended Lobatto method and the Radau IIA method [109]. The corresponding Butcher tables are given in Table 4.1 and 4.2.

$$\begin{array}{c|cc} 1/3 & 5/12 & -1/12 \\ 1 & 3/4 & 1/4 \\ \hline & 3/4 & 1/4 \end{array}$$

Table 4.1: Butcher table for Radau IIA (s=2).

The performance of these integration schemes has been investigated for different input signals. The example below shows the results of a single neuron's model with no synaptic input but with periodic dendritic fluctuations $\xi(t) = sin(\omega t)$ (Table 4.3).

0	$(1-\theta)/2$	$-(1-\theta)/2$
1	$1/2$	$1/2$
	$1/2$	$1/2$

Table 4.2: Butcher table for blended Lobatto (s=2).

Frequency	LIE	RadauIIA	Lobatto($\theta = 0.3$)	Diff2
10 kHz	0.03	0.05	0.01	0.03
5 kHz	0.006	0.025	0.002	0.009
1 kHz	0.002	0.005	$5 \cdot 10^{-5}$	$3 \cdot 10^{-4}$
500 Hz	0.002	0.0026	$2 \cdot 10^{-5}$	$9 \cdot 10^{-5}$
200 Hz	0.002	0.001	$6 \cdot 10^{-6}$	$1 \cdot 10^{-5}$
100 Hz	0.002	$5 \cdot 10^{-4}$	$3 \cdot 10^{-6}$	$5 \cdot 10^{-6}$

Table 4.3: Maximal distance between the approximation and the true solution for the linear implicit Euler method (LIE), the second order Radau IIA method, the blended Lobatto method with optimized parameter and the two-step difference method (Diff2).

The corresponding differential equation has the form: $U''(t) + (a+b) \cdot U'(t) + a \cdot b \cdot U(t) = -\omega^2 \cdot sin(\omega t) + (a+b) \cdot \omega \cdot cos(\omega t) + a \cdot b \cdot sin(\omega t)$. The step size has been fixed to $h = 0.01$ ms. The frequency of the periodic input has been changed from 100 Hz to 10 kHz. As you can see in Table (4.3), the error is smallest for the parameter optimized blended Lobatto method, followed by the difference method. The accuracy of Radau IIA lies between these and the linear implicit Euler method. For the data generation, I favor Diff2 since it is very fast and easy to implement and has also some advantage with regard to the incorporation of the noise signal (see below).

4.3.3 Noise Generation

As I said before, the additive noise signal corresponds to a colored random process. For the implementation, this random process has been approximated by a periodic function (sum-of-sinusoids), with uniformly distributed random phases between $[0, 2\pi]$, and amplitudes adapted to the Lorentzian function: $\xi(t) = \sum_{i=1}^{n} a_i \cdot sin(\omega_i \cdot t + \phi_i)$.

For the computation of the temporal dynamics of the membrane voltage, it should be noted that the noise process cannot simply be added to the Equation (4.13). The random process has to be adapted to the discretization method. The reason for this adaptation arises from the differential equation (Equ. 4.11) that can be also written in matrix-vector form:

$$\begin{pmatrix} u \\ v \end{pmatrix}' = \begin{bmatrix} 0 & 1 \\ -a \cdot b & -a-b \end{bmatrix} \cdot \begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} 0 \\ f(t) \end{pmatrix} . \tag{4.14}$$

The eigenvalues of the matrix will be $(-a, -b)$ and the eigenvectors will be $[1, -a]^T$ and $[1, -b]^T$. For similar constants $a \approx b$ the phase space will be distorted in one direction, which has a negative effect on the discretization. This incongruity can be eliminated by an adaptation of the noise process [202; 233]. The following example demonstrates how this can be done. Starting from the second-order differential equation:

$$\begin{aligned} U''(t) + (a+b) \cdot U'(t) + a \cdot b \cdot U(t) = \\ -\omega^2 sin(\omega \cdot t) + \omega \cdot (a+b) \cdot cos(\omega \cdot t) + a \cdot b \cdot sin(\omega \cdot t) \ , \end{aligned} \qquad (4.15)$$

with $U(0) = 0$ and $U'(0) = \omega$, the unique solution is given by $U(t) = sin(\omega \cdot t)$. Replacing the two terms $U''(t_n)$ and $U'(t_n)$ by their difference quotients: $U''(t_n) = \frac{U(t_{n+1}) - 2 \cdot U(t_n) + U(t_{n-1})}{h^2}$ and $U'(t_n) = \frac{U(t_{n+1}) - U(t_{n-1})}{2 \cdot h}$, leads to:

$$\begin{aligned} \frac{U(t_{n+1}) - 2 \cdot U(t_n) + U(t_{n-1})}{h^2} + (a+b)\frac{U(t_{n+1}) - U(t_{n-1})}{2 \cdot h} + a \cdot b \cdot U(t_n) = \\ (a \cdot b - \omega^2) sin(\omega \cdot t_n) + \omega \cdot (a+b) \cdot cos(\omega \cdot t_n) \ . \end{aligned} \qquad (4.16)$$

Choosing $U(t_n) = sin(\omega \cdot t_n)$, $U(t_{n-1}) = sin(\omega \cdot (t_n - h)) \ = \ sin(\omega \cdot t_n) \cdot cos(\omega \cdot h) + cos(\omega \cdot t_n) \cdot sin(\omega \cdot h)$ and resolving the equation for $U(t_{n+1})$ results in: $U(t_{n+1}) = \alpha \cdot sin(\omega \cdot t_n) + \beta \cdot cos(\omega \cdot t_n)$, with some constant α and β. The generalization of this procedure to a harmonic noise process built by sum-of-sinusoids is straight forward.

4.4 Small Neural Networks

In the following, small neural networks are used to generate different cortical-like signals for the comparison of the six projection methods. In the first model, I estimate the information transmission in a two-neuron excitatory-inhibitory network. Main emphasis is put on the behavior of the six projection methods according to an increase in the background activity. After that, I compare the behavior of the six projection methods in view of temporal signal changes. The last model consists of two different networks driven by the same input signal. Therein, I study the performance of the six projection methods after removing the linear components in the two data sets.

4.4.1 Wilson-Cowan Oscillator

In the first example, I compare the signals of a two-neuron network driven by two different input signals (Fig. 4.4). The first neuron has an excitatory synaptic connection to the second neuron. Alternately, the second neuron has an inhibitory synaptic connection to the first neuron. Such excitatory-inhibitory pairs of neurons are elementary circuits in the mammalian brain. They are found in all cortical areas, as well as on other brain structures, e.g. [2; 134; 243]. In the neural network literature such networks are often termed Wilson-Cowan Oscillator [267]. The first neuron is excited by a rectangle pulse of 100 ms duration, according to the time constants in the neuron model. There is a small difference in the amplitude (16.0 vs. 16.8 arb. unit), in order to generate two different data sets for the six projection approaches. The Bayes error and the Shannon information will be estimated for the discretly time sampled output signal of the first neuron. Therefore, the simulated signals are sampled at 1 kHz, producing 100 dimensional vectors. Each configuration has been repeated 800 times, overall 1600 Monte Carlo simulations are performed. The actual parameter settings can be found in the Appendix A.2, as well as a simplified stability analysis (see Appendix A.4). The behavior of the above described projection methods is examined depending on the size of the learning set (10, 20, 40 trails per class) and the power of the noise level.

The mean values of the discriminant analysis for the different projection methods and the different parameter settings are illustrated in Figure (4.5). The training and testing procedure has been repeated 400 times so that the error of the mean values is lower than 0.004. The effect of the noise level and the training size for the Bayes error can be seen on the left (Fig. 4.5 A). On the right hand side (Fig. 4.5 B) you can see the results for the Shannon information. As expected, the overlap, and therefore the Bayes error, increases with noise level. [3] From an information theoretical perspective, the statistical dependence between the stimulus pulse and the neural response becomes smaller with increasing noise. Correspondingly, the Shannon information converges to zero with decreasing signal-to-noise ratio (SNR). In the limit cases, all projection methods behave similarly. They indicate a unique discrimination at a signal-to-noise ratio above 25 dB. For a signal-to-noise ratio below -1 dB no method shows a significant difference in the small training case. Looking at the projection methods separately, RBF performs best followed by LCC and RDA. For small training sets LFD and KFD behave similarly. For larger training sets LFD behaves irregularly. Overall ADC performs worse. Interestingly, the estimated information by ADC is not significant ($p < 0.01$) for the large training set although the mean value is higher than the corresponding RBF values for a signal-to-noise ratio of -1 dB.

Generally, there is a gain in the differences between the six projection methods for larger training sets. The additional information from a larger training set increases the performance of the radial basis function approach most, so that the difference to the other method becomes larger. This effect becomes plainest comparing RBF with LFD. For example, at a signal-to-noise ratio of about 12 dB, the RBF information estimation is twice as large as that from LFD (0.8 bit versus 0.4 bit). For the information value 0.625 bit, the corresponding SNR of RBF is 10 dB, whereas it is 15 dB for LFD.

Recapitulating, for this artificial data set RBF and LCC show the best discrimination. Therefore, I conclude that RBF and LCC are less influenced in the tested model by uncorrelated random fluctuations.

---

[3] The SNR has been computed by: $20. * log_{10}(\alpha)$ with $\alpha = ((m_1 - m_2)\Sigma^{-1}(m_1 - m_2)^T)^{1/2}$, $\Sigma = 0.5 * (\Sigma_1 + \Sigma_2)$ and $\Sigma_i = E[(x - m_i)(x - m_i)^T]$ [167, pp.204] m_1 and m_2 are the mean values of the two data sets.

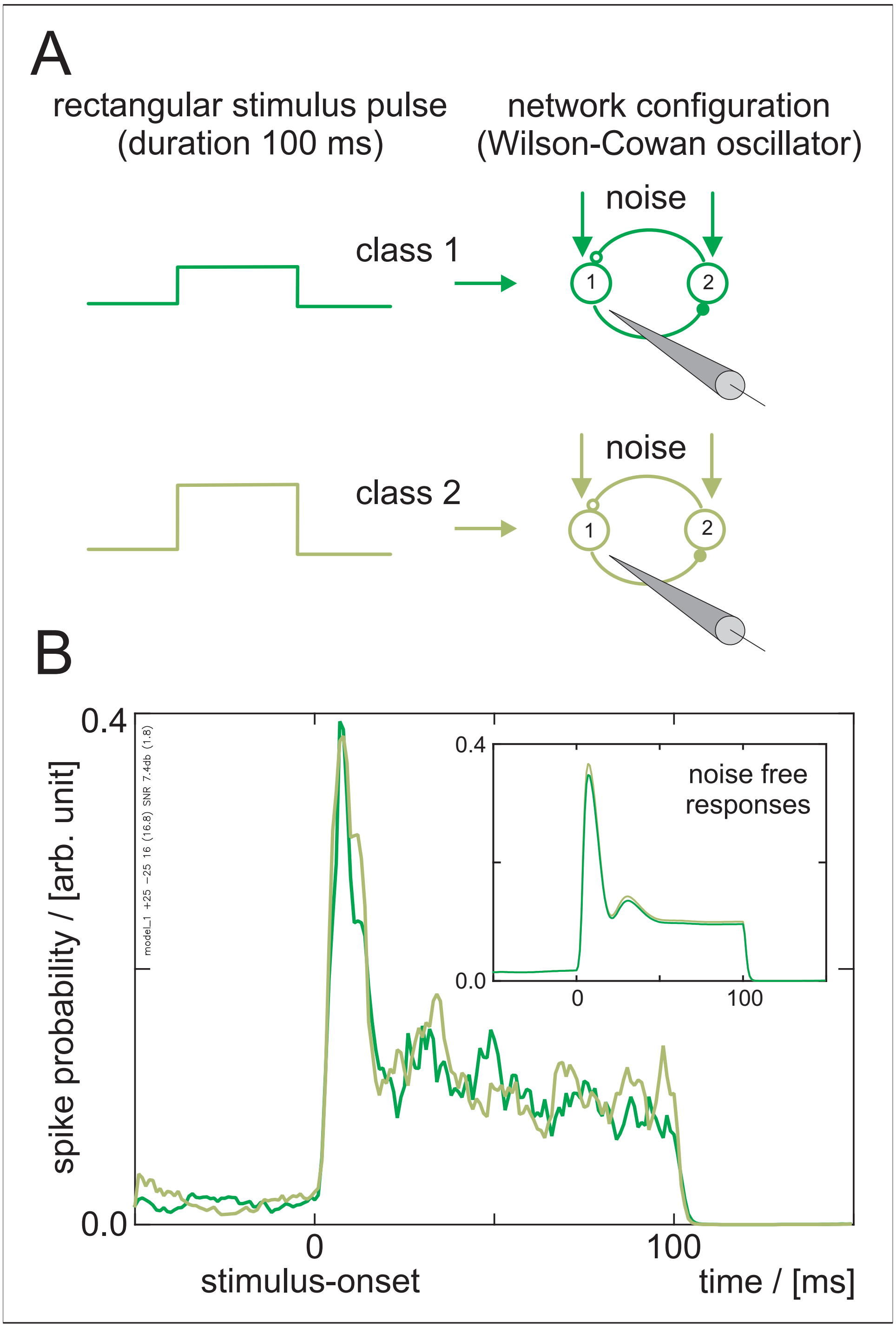

Figure 4.4: (**A**) Depiction of the neural network configuration and the shape of the square wave pulses corresponding to the two stimuli (see Appendix A.2 for parameter setting). The stepsize has been fixed to 0.1 ms. (**B**) Typical signals recorded from one neuron of the Wilson-Cowan oscillator evoked by the two stimuli at a signal-to-noise ratio of 7.4 dB. Small picture inside: noise free responses.

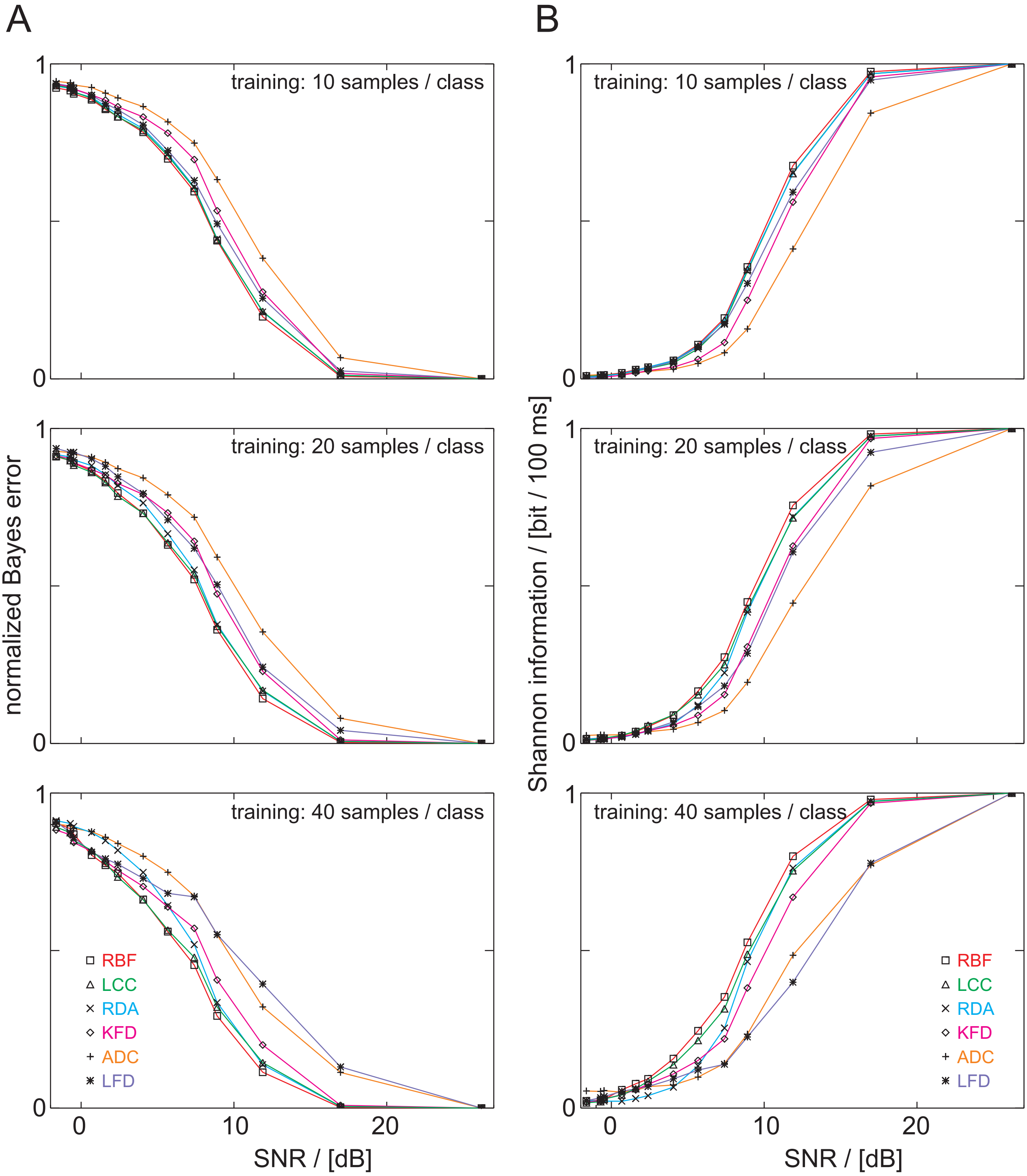

Figure 4.5: (**A**) Estimation of the normalized Bayes error for different signal-to-noise ratios (SNR) and different sample sizes computed from the signals of the two-neuron model according to Figure (4.4). The training sample size varies between 10, 20 and 40 samples per class. (**B**) Corresponding estimation of the Shannon information.

4.4.2 Time-Resolved Discriminant Analysis

In the following, I expand the previous analysis comparing the temporal dynamic of the two-neuron Wilson-Cowan oscillator signals. A sliding window technique (5 ms shifts) with varying window duration is used. The window duration varies between 4,10,20,...,90 ms. For the training of the projection methods 20 samples per class have been randomly chosen. Overall, repeated training and testing has been carried out 120 times. According to the symmetry between the Bayes error and the Shannon information only the temporal variation in the information transmission at a noise level of 11.88 dB is illustrated. The mean values of the time-resolved results for the different methods, centered around the sliding window, can be found in Figure (4.6).

As you can see, all methods indicate two local maxima 10 ms and 30 ms after stimulus-onset shifted to the right with increasing window duration. Using longer signal segments for the time-resolved analysis, what is similar to a cumulative sum techniques, results in a shift to the right because there are no differences in the two data sets before and after the presentation of the rectangular stimulus pulses. The information multiplication with longer window duration stems from the difference in the serial correlations between the two data sets. Therein, the difference is more pronounced within the first 50 ms, than in the last 50 ms. This can be explained by the mean response signals of the two classes Figure (4.7), where the overlap is smallest 10 ms and 30 ms after stimulus-onset. At the beginning and at the end of the analysis window, the signals are similar.

With respect to the performance of the six projection approaches, attention is invited to the dependence on the window duration. Figure (4.6) shows that the six methods behave different for different window durations. For example, RDA shows a gap between the 10 ms and the 20 ms window duration, and LFD performs worse for window durations between 30 to 70 ms. Furthermore, LFD contradicts the theoretical guideline, that the transmitted information increases with the duration of the analyzed signal segment. At a window duration of 4 ms, the maximal information value, indicated by KFD, exceeds the values of the other methods. For longer window durations (more than 70 ms) RBF, LCC and RDA reach the largest information values. For signal segments of more than 90 ms the maximal information value is obtained by the radial basis function approximation, what is in accordance to the previous findings (Section 4.4.1).

The comparison of the projection approach and classical techniques reveals some clear distinctions. In contrast to classical discrimination approaches where the main focus lies on differences in the mean activity (post-stimulus-time histogram), the significant differences become more clear by most of the projection methods taking the temporal correlations into account. The information profit from the serial correlation becomes most clear if we compare the result of a one dimensional analysis and the RBF result with a sliding window duration of 90 ms. The maximal value for the one dimensional analysis is about 0.22 bit (data not shown) in contrast to nearly 0.8 bit obtained by the RBF method.

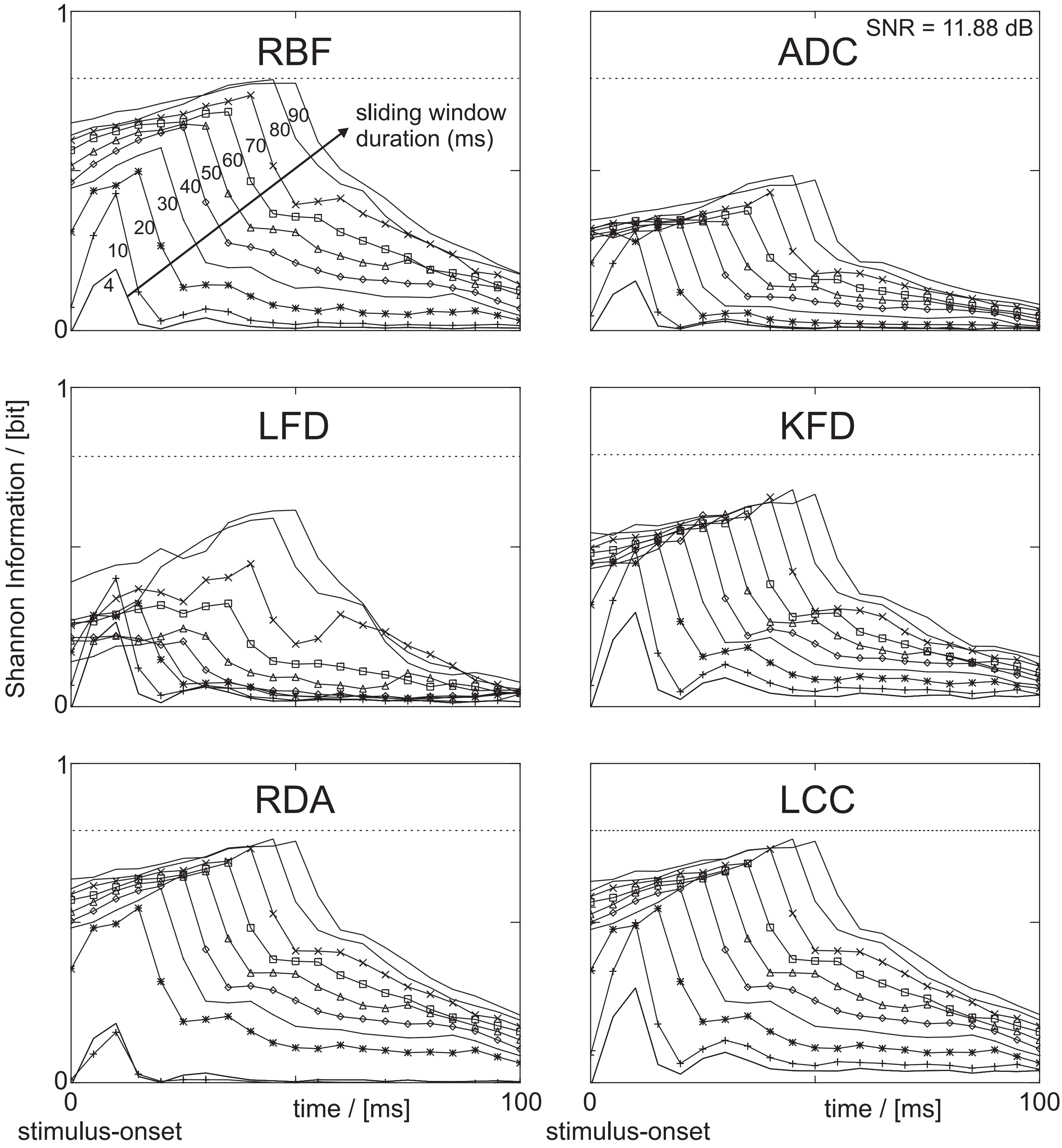

Figure 4.6: Comparison of the six projection methods (Chapter 2) by a time-resolved analysis of the data from Section 4.4.1 at a fixed signal-to-noise ratio ($\approx 12\ dB$). The discrimination results are centered around the sliding analysis-windows. Time periods extend from 4 ms, over 10 ms up to 90 ms. Learning set: 20 samples per class.

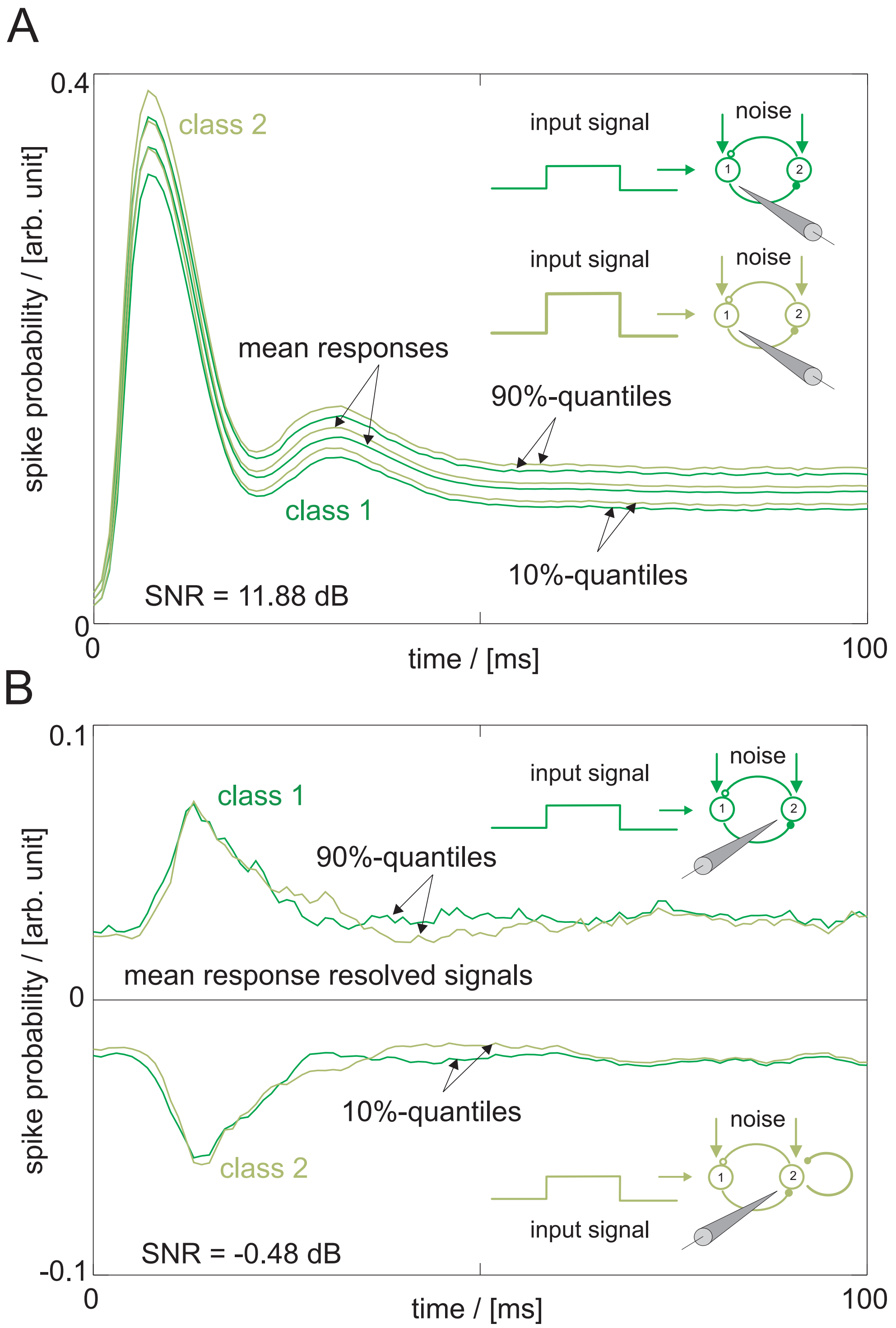

Figure 4.7: (**A**) Temporal variation of the responses from the Wilson-Cowan oscillator for the two input signals (Fig. 4.4). There is a slight difference between the mean responses evoked by the different amplitudes of the input signals. Correspondingly, responses from class 2 reach more distinctive spike probability values (on average). The inter-quantile range (10-90 %) is nearly identical for both signal classes. (**B**) Depiction of the mean response removed signals of two different neural networks at a signal-to-noise ratio of 0.48 dB (see Appendix A.3 for further parameter setting). A detailed description as well as an depiction of the discriminant analysis can be found in the next Section. Between 20 to 50 ms after stimulus-onset the temporal variation is slightly different between the two data sets.

4.4.3 Nonlinear Component Discrimination

In the following, I generate two signal classes by stimulating two different neural networks with the same rectangular stimulus pulse of 100 ms (Fig. 4.7 B). One network is identical to the previously described Wilson-Cowan oscillator [267]. The other network consists of two neurons with a positive feedback loop. In addition to this, the second neuron has a delayed self-inhibiting loop, i.e. feeding back to itself. The square wave pulse (amplitude: 10 arb. unit) has a direct influence on the first neuron in both networks (Fig. 4.4). Each network has been stimulated 800 times (overall 1600 trials), while the statistical independent random fluctuations added to each neuron have been kept at a fix noise level of -0.48 dB (see Appendix A.3 for specific parameter settings). The discriminant analysis has been carried out on the output signals of the second neuron. Before the signals are analyzed, the data are sampled at 1 kHz. The main idea of this artificial experiment is to look if a significant discrimination is possible from the nonlinear components and how the six projection methods perform in this situation. Therefore, I estimate the mean responses for each network. Afterwards, the mean responses have been subtracted from each trial. As in the previous model, a time-resolved discriminant analysis (2 ms shifts) is carried out, concentrating on the discrimination of the nonlinear components. The learning sets consist of 20 randomly selected samples per class. Each signal segment (window duration: 8 ms) has been trained and tested 120 times.

The result of this analysis can be found in Figure (4.8). In contrast to the previous depictions, I use the symmetrized area relative to the diagonal of the receiver operating characteristic (ROC) as distance measure (see also Section 1.9). The dashed lines correspond to a significance level of $p < 0.01$. The colored curves, which are mostly near the abscissa, represent the proportion of the misclassification rate, expressed by twice the area below the ROC.

As you can see from Figure (4.8), the outcome of the six projection methods is similar. All methods indicate two local peaks about 30 ms and 45 ms after stimulus-onset. These peaks are below an area value of 10 % for the linear approaches (LFD, LCC), between 10 and 20 % for the regularized and the cluster method (RDA, ADC), and above 20 % for the kernel methods (RBF, KFD). With respect to the significance level ($p < 0.01$), it comes out that the linear approaches are near the significance level. In contrast to the linear methods, the nonlinear methods exceed the significance level indisputable. Furthermore, comparing the linear and the nonlinear approaches it comes out that the area below the diagonal in the area statistic is larger for the linear methods than for the nonlinear methods. Therefore, the one dimensional distributions obtained by LCC and LFD pocess a pronounced misclassification rate. An explanation for the restricted performance of the two linear methods, can be easily given, because the data contain no first order information. Comparing all method with each other it can be seen that the RBF method reaches the most significant distance (AROC $=$ 32 %). In accordance to the receiver operating characteristic, RBF attains also the highest information content (0.1 bit), and the lowest classification error of about 38 %. The good performance of RBF can be explained by the nonlinear adaptation to this nonlinear situation and the robust behavior of this method with respect to outliers. From a numerical perspective, I have to notice that the univariate test sets for RDA are skewed due to the presence of many outliers, and that the condition numbers for KFD are very high. This particular aspect is also responsible for the huge variation of the boxplots in the KFD graphic.

An explanation for the temporal dynamic can be given from the mean responses of the two data sets, and the signal variation after removing the linear components (Fig. 4.7 B). The first increase in the neural activity of the two networks, induced by the rectangular stimulus pulse, is nearly identical and similar to the situation in Figure (4.4). Nevertheless, the following decrease is slightly different between the two networks (data not shown). Therefore, the mean resolved response distributions are different in the range 30 to 50 ms. The network with the self-inhibiting loop shows a stronger signal damping than the Wilson-Cowan oscillator. Therefore, the second upcome of the neural activity about 45 ms after stimulus-onset is not so pronounced in the self-inhibiting network (see also Fig. 4.4) As a consequence, the variation of the mean resolved signals is smaller about 30 ms after stimulus-onset for the Wilson-Cowan oscillator than for the neural network with the self-inhibition loop.

Annotation: I have chosen signal segment with a window duration of 8 ms. Interestingly, choosing longer signal segments (> 64 ms) destroys or at least substantially reduces the two significant peaks in all approaches (see Fig. 4.8).

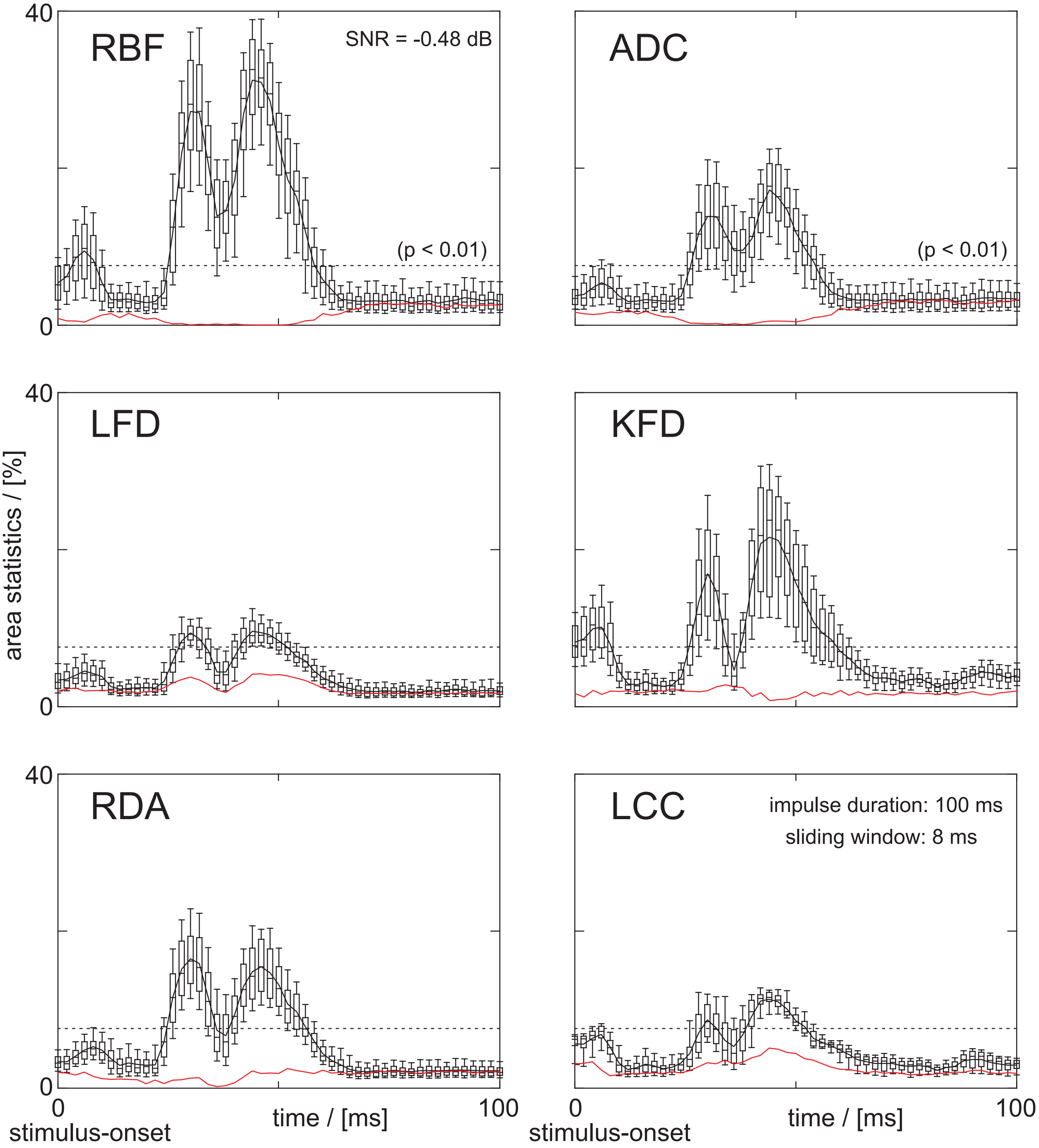

Figure 4.8: Comparison of the six projection methods from Chapter 2. The 8-dimensional data sets have been generated by two artificial neural networks where the linear signal components have been removed before the discriminant analysis was performed (see the text below and Appendix A.3 for parameter setting). The segregation is quantified by the symmetrized area relative to the diagonal of the receiver operating characteristic $A = 2 \cdot (A^+ + A^-)$ (see Section 1.9). Boxplot: 10, 25, 50, 75, 90%-quantiles of A according to 120 cross-validations. Red curve: twice the mean one-side area statistic A^- indicating one-side dominance (Section 1.9).

4.5 Summary

Surely, I omit various details of real neural networks in my network simulations. Nevertheless, the neuron model incorporates pronounced aspects of real neurons leading to cortical like responses. Therefore, the results of the preceding discriminant analyses of simulated neural signals are very important for the interpretation of the six methods and their application to high-dimensional cortical recordings. According to this outcome, let me first summarize some conclusions with respect to the six projection methods.

- LCC can be easily implemented in different programming languages. The method performs well in situations where the main discrimination comes from the mean responses. The discrimination of nonlinear data constellations is restricted. LCC behaves robust against noise and outliers, but shows a minor profit from larger training sets.

- LFD shows no profit against LCC within the analyzed data. The projection vector is motivated by a Gaussian distribution, which seems to be unattained. The method has an irregular behavior with respect to the dimensionality. Furthermore, the performance depends on the size of the training set. Good results will be obtain in low dimensions.

- ADC overs a fast implementation. The method becomes better with increasing training sets. With respect to the other methods and the simulated neural signals the estimated information is lowest in some situations.

- RDA is motivated by a Gaussian assumption about the underlying distributions, but the underlying model is more general than for LFD. The discrimination performance is superior to LFD and ADC. Within the small sample case RDA shows some numerical instabilities.

- KFD, regarded as a nonlinear generalization of LFD, can be successfully applied to nonlinear discrimination problems. The performance is worse to RBF for the neural network data, and the behavior is more influenced by noise and outliers. The outcome depends slightly on the size of the training sets and it is possible that the performance becomes worse with increasing training size.

- RBF performs best within the analyzed data sets, especially within high dimensions. The method behaves robust against noise and outliers and becomes better with increasing training sets. Furthermore, the one dimensional projection is normalized so that subsequent techniques, e.g., for the information or Bayes error estimation can be easily applied.

Although some of the projection methods seem to be inappropriate for the discriminant analysis of the simulated neural data, these results illustrate the usefulness of the dimension reduction approach in neuroscience, and enable us to get additional functional insight in the temporal dynamic between neurons. Furthermore, the projection approach can be regarded as a natural generalization of classical discrimination approaches. This becomes clear, for example from the results in Section 4.4.3. The one dimensional receiver operating characteristic obtained from the investigation of each sample point reveals a similar temporal dynamic like from the high-dimensional projection approaches (data not shown). However, the information maximum obtained from the one dimensional analysis is near the significance level and therefore much lower than, e.g. by the kernel methods. Furthermore, in contrast to classical techniques, the projection approach can be adapted to some extent to the temporal dynamic of the underlying distributions. This can be seen from the results in Section 4.4.2, in which I investigate the dependence of the segregation from the size of the sliding analysis-window. In this experiment, the projection approach makes it possible to visualize the information gain from the temporal correlation structure, what cannot be done in a one dimensional (classical) setting.

Surely, this is a restricted investigation of possible neural networks arrangements, but in addition to the presented results, I have compared the performance and the application of the six projection methods to other artificial neural structures. For example, I have generated signals from a neural network with 5 neurons put together by a random connection matrix, in which two neurons have been stimulated by different rectangular stimulus pulses. The discrimination of the signals from all neurons shows a clear distinction after the stimulus-onset (data not shown). In accordance to the results in this Chapter, RBF achieves the most significant discrimination. Furthermore, a study has been done on signals from an integrate-and-fire neural network, in which the focus lies on the discrimination properties of different continuous signal types extracted from the spiking network [147]. Lastly, I have performed various analyses with bandwidth-limited Gaussian white noise (1-500 Hz) [148].

Putting all these facts together, I can say that the statistical dependence within a time series or between different recording sites leads to an increase in the information estimation by most of the projection techniques, and that this approach has some advantage over classical techniques. Therefore, I claim that the projection approach and especially the RBF method is well suited for the spatio-temporal discriminant analysis of cortical multi-channel recordings, which will be demonstrated in the next Chapter.

5 Binocular Rivalry

'Obviously to understand brain function, we need to confront with its complexity. Various strategies have been proposed to address this issue, but one view is that the way forward is to obtain vast amounts of data characterizing the brain. Gathering the data is only the first step on this path, however; effectively mining this information and interpreting it are as difficult, and as crucial.' (Narasimhan, K. (2004) Scaling up Neuroscience. Nat. Neurosci.,7:425)

In the following, I will use the RBF method to analyze multi-channel recordings from the monkey's primary visual cortex under binocular visual stimulation. After a short description of the experimental setup, special emphasis is put on the temporal dynamic of the cortical signals. Beyond these studies, I occupy myself with aspects of neural decoding and prediction. In doing so, I find that the primary visual cortex contains signal components correlated with visual perception and motor action. In contrast to single channel approaches, it is possible to predict a monkey's reactions with high probability.

5.1 Experimental Setup

The data have been obtained from two male rhesus monkeys (macaca mulatta), abbreviated by S and H [93]. Two electrophysiological signal types have been recorded: multi-unit activity (MUA) and local field potentials (LFP). MUA has been retained from the raw broad-band signals (1 Hz - 10 kHz) by band-passing (1-10 kHz, 18 dB/oct), full-wave rectification, and subsequent low-pass filtering (140 Hz, 18 dB/oct). LFP has been retained from the raw broad-band signals by low-pass filtering (140 Hz, 18 dB/oct). From an electrophysiological perspective, MUA is composed of local (radius: 70 μm) spike density weighted by soma size and distance from electrode tip. LFP reflects a weighted average of dendro-somatic components of the postsynaptic signals of a neuronal population (radius: 400 μm), Therefore, the catchment area of the MUA signal is more local than the LFP signal. Up to 16 quartz-isolated, platinum-tungsten micro-electrodes, arranged in a regular 4×4 array with 750 μm pitch, have been used [67]. The cortical activity has been recorded from supragranular neurons (layer I-III) in the primary visual cortex (V1, parafoveal). For more details about the animal preparation and the recording techniques see [93] and the references therein.

Each eye has been stimulated separately through a dichoptical setup. The visual stimuli consisted of horizontal or vertical gratings of sinusoidal luminance. In the congruent (non-rivalrous) condition, both eyes have been stimulated by the same object. In the incongruent (rivalrous) condition, each eye has been stimulated by a grating of different orientation (Fig. 5.1 B). While in the congruent condition the visual perception is unique, it changes randomly in the incongruent condition, perceiving for some seconds only the image from one eye. This phenomenon is called

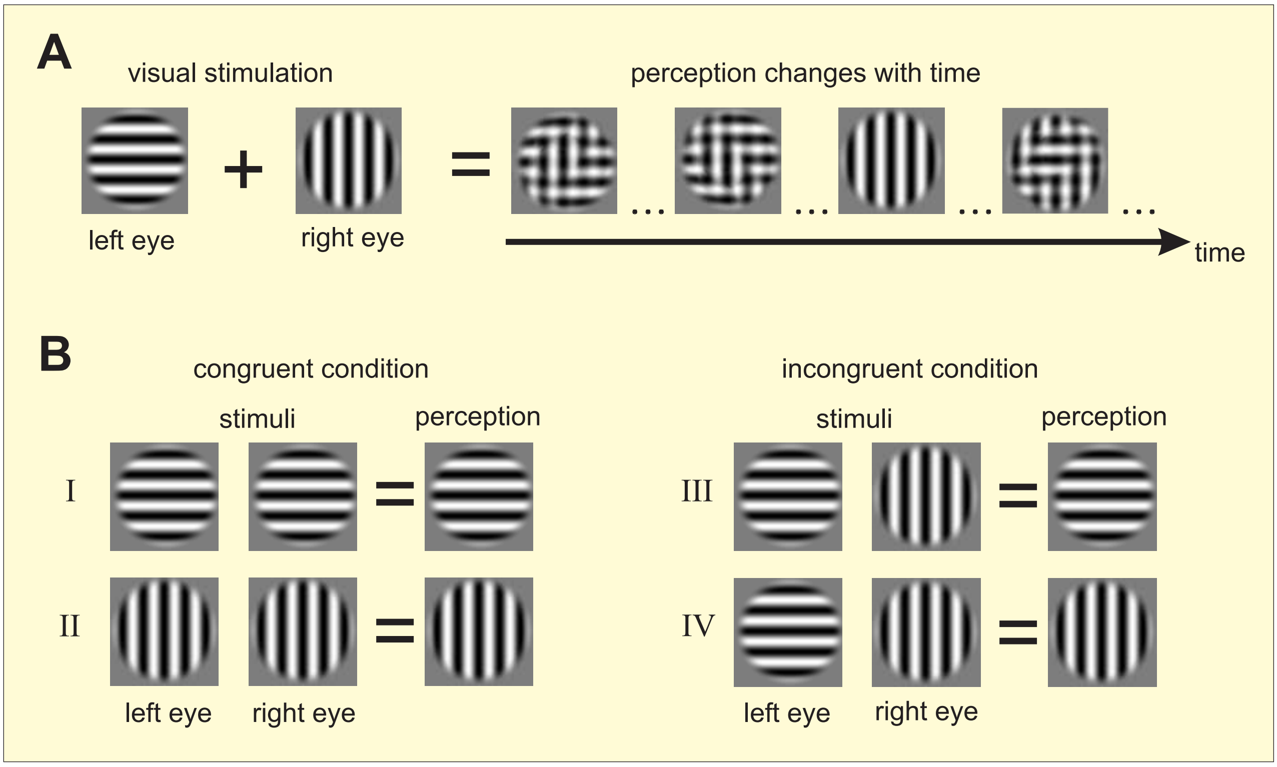

Figure 5.1: Binocular visual stimulation. (**A**) Depiction of binocular rivalry. The separate visual stimulation of the left and right eye by gratings of different orientation causes a change in the visual perception over time. (**B**) Four visual stimuli will be taken into account in the following. **I** stimulation by a horizontal grating, **II** stimulation by a vertical grating, **III** stimulation by an incongruent stimulus with horizontal perception, **IV** stimulation by an incongruent stimulus with vertical perception. The situations **III** and **IV** lead to binocular rivalry in which one eye's view dominates the visual perception for some time and is then replaced by that of the other eye.

binocular rivalry. During a single trial, the monkey has to fixate within $\pm 0.45^\circ$ at a small spot. The presentation of a special stimulus stays constant during a single trial. In the congruent condition, moreover, a stimulus looking like a locally conflicting percept called *piecemeal* (Fig. 5.1 A 3rd and 4th stimuli) was presented for a variable time (50 to 250 ms from the beginning) [94]. The monkey had to report in a time interval of 1200 ms, which orientation he perceives by pushing a lever upwards indicating a horizontal grating or downwards indicating a vertical grating. In order to prevent the monkey from behaving randomly in the rivalrous condition, *catch trials* and psychophysical experiments have been done [94]. Cortical activity has been recorded during 9 (S) and 11 (H) days.

5.2 Decoding of Neural Signals by Dimension Reduction

In the first study, I compare the signals related to the congruent condition. For that purpose I looked for differences in the data from the incongruent condition. In a pre-processing step each multi-electrode recording was aligned to the stimulus-onset and then to the time of the monkeys' response. Further, the signals of each electrode had been aligned to the mean value in the forerunnings and normalized to the standard deviation in this signal segment. For the time-resolved discriminant analysis, signal segments of 48 ms have been taken, and shifted in time by 12 ms. I examined the simultaneous recordings from 11(13) electrodes from monkey S(H). In order to apply the RBF method, the recordings from the single electrodes have been combined. The data build a 264(312) dimensional vector set for monkey S(H), corresponding to the number of electrodes, the sampling rate (500 Hz) and the sliding window duration. In addition, I studied other arrangements, e.g., varying the number of electrodes or the window duration. Their influence on the discrimination will be discussed later (Section 5.5.3).

Learning and testing was done for each session separately. During a single session signals from about 100 to 160 non-rivalrous trials were recorded. In the rivalrous condition, the trial number varied between 60 to 120 trials per session. According to the number of trials per session, the learning set size varies between 20 to 60 trials per class. Instead of quantifying the statistical dependence between the different signal groups for each session separately, I combined the various test sets. This is possible, since the RBF projection incorporates some sort of normalization (see Chapter 2). Remember, this is not the case for the other tested projection methods. This procedure is nearly as fast as if I had performed the whole discriminant analysis for each session separately and had afterwards averaged over the different quantities. One consequence of the multiple combination is that the test set increases which leads to a better statistical analysis. Instead of pooling the data from the single sessions prior to the dimension reduction, only data under physiologically identical conditions were compared. In order to reduce the influence of the randomly chosen learning sets, the cross-validation procedure has been repeated 120 (200) times in the congruent (incongruent) condition for each signal segment.

5.2.1 Congruent - Non-Rivalrous Perception

Results of the multi-channel discriminant analysis, for the non-rivalrous data, can be seen in Figure (5.2). Statistical significance is quantified by twice the area between the receiver operating characteristic (ROC) and the diagonal of the unit square (Section 1.9), in which the value 1 means perfect discrimination [212]. The significance level has been choosen to be one percent ($p < 0.01$). In addition, I have computed and plotted twice the area under the ROC (solid colored), contradicting the one-side dominance [145]. Remember, the violation of the one-side dominance means that samples from class C_1 will be classified as samples from class C_2 and reversed.

Looking at the stimulus-onset triggered curves (Fig. 5.2 A), it can be seen that in all cases no significant distinction is indicated before the presentation of a binocular grating. The first significant difference appears more than 100 ms after stimulus-onset, because of the piecemeal stimulus (see Section 5.1) used in this arrangement. After the piecemeal stimulus period the segregation becomes larger. Thereby, the gain is steeper for monkey H, who reported his perception in most cases earlier than monkey S. The time-course of the transmitted information is equivalent to that illustrated by the area statistic. About 240 ms after stimulus-onset, Shannon information reaches a local maximum for monkey H. The maximal information value is approximately 0.75 bit for both signal types (LFP and MUA) in the two-class problem. Afterwards, the Shannon information decreases in the LFP signals to a value of 0.4 bit, whereas it stays constant for the MUA signals. In terms of classification, nearly perfect prediction is possible for monkey H from the MUA recordings. The normalized Bayes error goes down to 0.14. Therefore, on average, 7 percent on average will be misclassified. For the other monkey (S) at least twice as many classification errors were made. This means best prediction is, on average, 85(75) percent for MUA(LFP).

In the response-triggered data, significant differences from chance can be found up to 300 ms pre-response. All curves (Fig. (5.2) B) increase during 300 to 200 ms pre-response. The separability of the LFP data shows a soft incursion about 80 ms pre-response. In both monkeys, the ROC-values for the MUA data stay nearly constant. In addition, the statistical dependence between the stimulus set and the multi-channel recordings is higher for the MUA data 0.63 bit (S), 0.83 bit (H) than for the LFP data 0.51 bit (S), 0.54 bit (H). Therefore, I conclude that MUA contains more information about the stimuli than LFP. In terms of prediction, a slightly better assignment is possible from the MUA than from the LFP data. Restricting to the time 200 ms before the monkeys push the lever, and using the RBF map for classification with an appropriate threshold, the Bayes error is as low as 15 $\%$ (S), 8 $\%$ (H) percent. This means, on average 85 $\%$ (S), 92 $\%$ (H) of all trials were correctly classified. Notice, in this case the area below the diagonal in the receiver operating characteristic is below the significance level.

I also tested these and the following results against surrogate data sets with randomly permutated labels. For this, I generated 100 surrogate data sets from each signal-segment. What I found is that the significant differences in the original data were significantly higher ($p < 0.01$) than the differences obtained from the surrogate data sets (data not shown). Surrogate differences were mostly below the significance level for the discriminant analysis (see also Section 6.2.4). Furthermore, I checked these and the following data for stationarity conditions with regard to the discrimination. In doing so, I compared the discrimination at the beginning of a session with the discrimination at the end of a session. In all cases the discriminant analysis was not mainly affected by the experimental progress (see also Section 5.2.2).

5.2.2 Incongruent Rivalrous Perception

The clear distinction in the non-rivalrous condition can be described by the use of two different visual stimuli. But can we discriminate between the signals under binocular rivalry, where the visual input is the same in both situations regarding the different reactions? To answer this question, as before, a time-resolved multi-channel discriminant analysis was done. The results of this study can be seen in Figure (5.3). As expected, a significant discrimination is not possible before stimulus-onset.

Compared to the congruent condition, minor differences appear in the stimulus-onset triggered data (Fig. (5.3) A). The separability is mostly near or under the one percent level for both monkeys and both signal types. A slight increase in the ROC-curves (LFP) was found for both monkeys, 300 ms after stimulus-onset. In addition, a significant episode appears for monkey H, about 140 ms after stimulus-onset. For the LFP data, correct classification is, on average, 62 % (S) and 61 % (H), using the signal segment 350 ms after stimulus-onset. For the MUA data, the ROC-curves evolve irregularly. A significant discrimination (on average) appears between 70 to 120 ms for both monkeys. Significant values are absent after that time for monkey S. Prediction from this data is near chance and the Shannon information does not exceed a value of 0.1 bit. While, the discrimination in the stimulus-onset triggered data is low, a highly significant segregation is indicated in the response-triggered data (Fig. (5.3) B). The inequality in the two data sets appear 70 ms pre-response time. Furthermore, a significant difference was found in monkey (H) about 250 to 150 ms pre-response time. A similar modulation, about 200 ms pre-response, can be seen in the data of monkey (S). It is not significant with the one percent level, but at the five percent level. In contrast to the congruent condition, the temporal dynamic of the MUA data is not consistent for the two monkeys (Fig. (5.3) B). The segregation for monkey S increases shortly before he pushed the lever. For the other monkey, a significant difference can be found 250 ms pre-response, but near the response time this effect disappears. Notice, in all cases, the one-sided dominance remained below the significance level (see also Section 1.9).

Summarizing, there are two important facts. First, the LFP discrimination is more significant than the MUA discrimination in the incongruent condition. Second, the LFP data seem to contain two successively appearing perception-related components, one at 250 to 150 ms pre-response and the other much nearer the response time. Based on the multi-electrode LFP data, correct single-trial prediction reaches to 76 % (S) and 75 % (H).

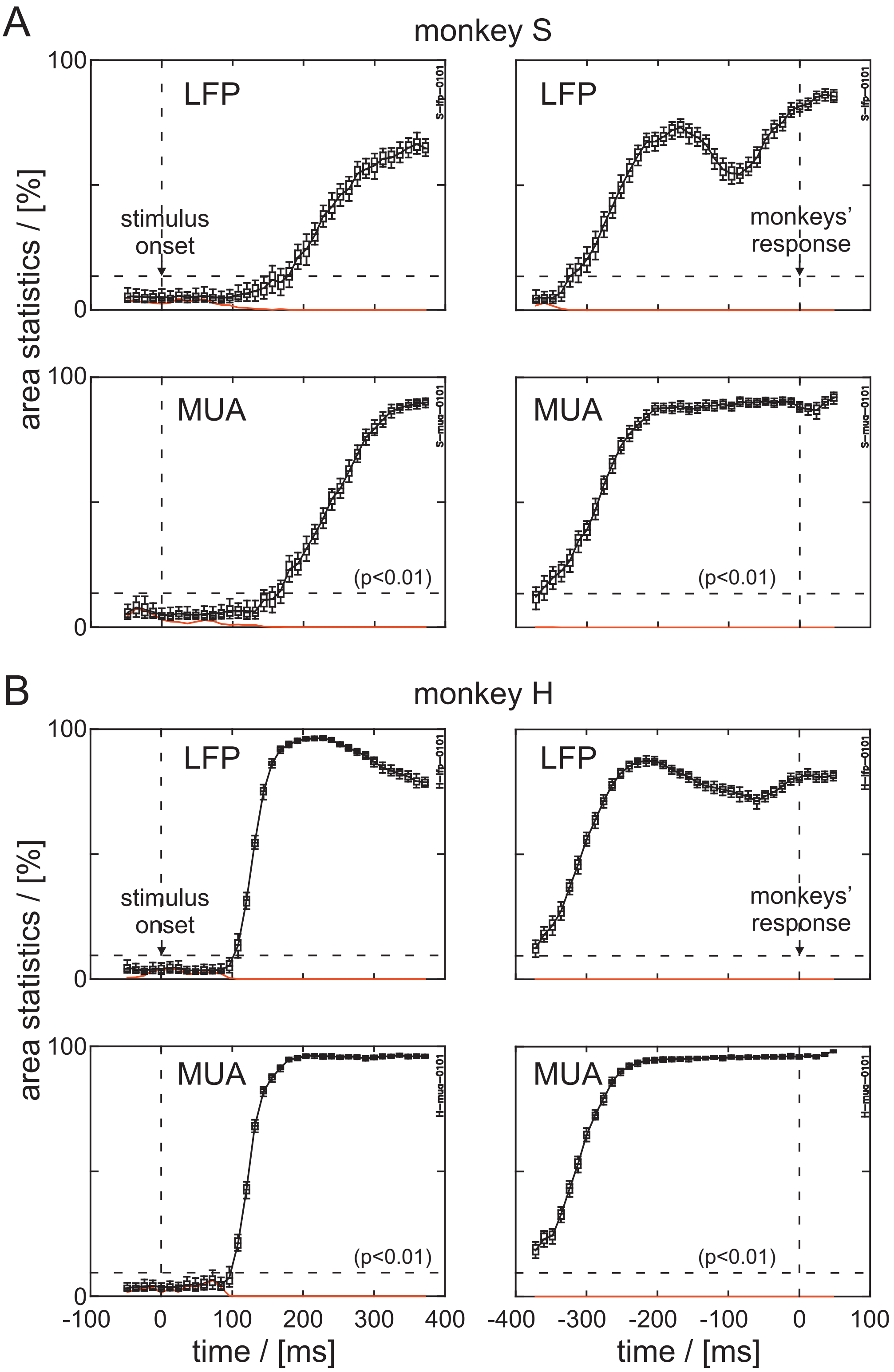

Figure 5.2: Non-Rivalrous Stimulation. Time-resolved discriminant analysis of the multi-channel recordings from monkey S (**A**) and monkey H (**B**) by the RBF method. Left column: Distinction after alignment to the stimulus-onset. Right column: Distinction after alignment to the response-time. Separability is quantified in $\%$ by twice the area between the ROC and the diagonal of the unit square. The horizontal dashed line corresponds to a significance level of $p < 0.01$. Red curve: twice the mean one-side area statistic indicating one-side dominance.

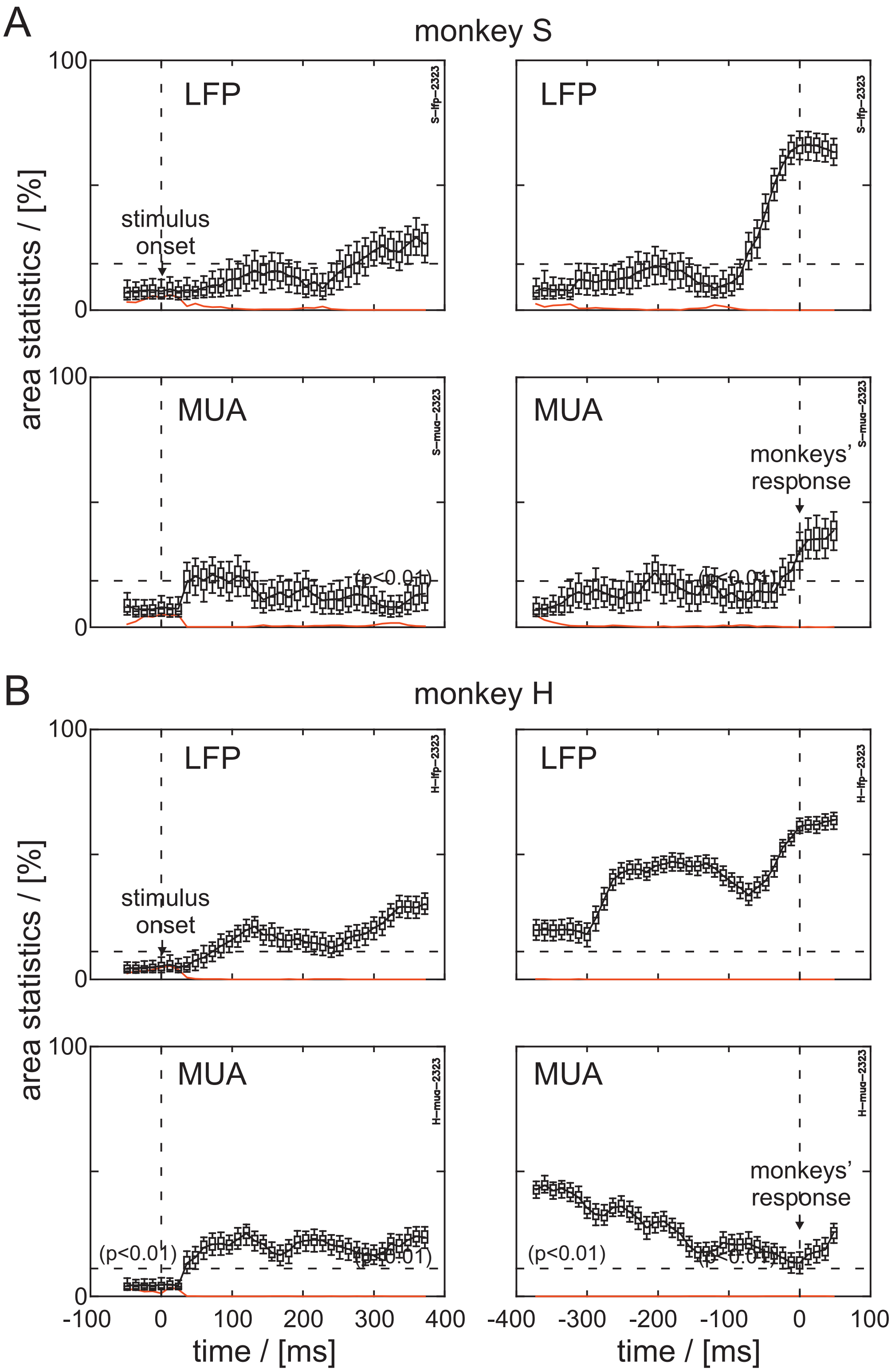

Figure 5.3: Binocular rivalry. Time-resolved discriminant analysis of the multi-electrode data from monkey S (**A**) and monkey H (**B**). Separability is illustrated by twice the area between the ROC and the diagonal of the unit square. Average one-side dominance of the distributions is indicated by the red curve. Data aligned to the stimulus-onset are shown in the left column. Data aligned to the monkeys' responses are shown in the right column. The significance level $p < 0.01$ is indicated by the horizontal dashed line.

5.3 Feature Extraction

Besides these results, many questions came up. For example, which are the main reasons affecting the observed differences? Are these effects visible in all recording sessions and signals from all electrodes? To answer these questions, I manipulated the data by various pre-processing or filtering techniques. In addition, I generated surrogate data according to basic model assumptions. Since in the stimulus-onset triggered data only minor or neglible differences occurred, I did not investigate these data further. Therefore, I will concentrate this research on the data triggered by the monkey response.

5.3.1 Centre and Power Analysis

For the following Section, I assume that the multi-channel recordings are samples drawn from unimodal (normal) distributions with different centres m_i (i=1,..,4) but the same uncorrelated variance matrix. In this situation a very simple idea is to look at the relations of the mean amplitude values with respect to the mean uncorrelated covariance matrix $(m_1 - m_2)^T \Sigma^{-1} (m_1 - m_2)$, $\Sigma = \frac{1}{2} \cdot diag(\sigma_{11}^2 + \sigma_{21}^2, ..., \sigma_{1n}^2 + \sigma_{2n}^2)$, also known as signal-to-noise ratio of the power (Section 4.4). Figure (5.4) shows two quantities which are interesting in this context.

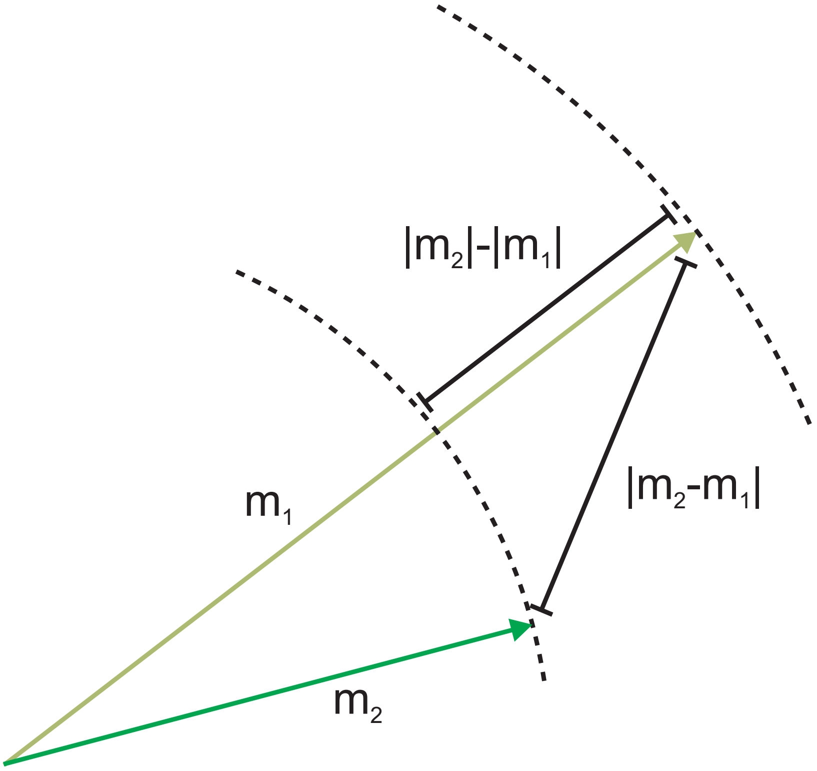

Figure 5.4: Illustration of a simple model for the comparison of two multivariate data sets. The underlying distributions have been reduced to the correlation among their centres. Two possible quantities are illustrated: the distance relative to the origin $|m_1| - |m_2|$ and the distance between the mean values $|m_1 - m_2|$.

Figure (5.5) shows the time-resolved distances between the mean values for the single recording sessions. Generally, there are differences between data of different sessions but overall the temporal dynamic is similar. Looking at the mean signal-to-noise ratio (red curve), the same temporal structure appears across data of nearly all sessions. Comparing these results with the results from Section 5.2 you can see similarities in the temporal dynamic. Therefore, I conclude that some of the differences in Figure (5.2) and (5.3) are attributed to a shift in the centres between the two multivariate data sets. Furthermore, it seems to be reasonable that the differences in the multi-channel discriminant analysis (Fig. 5.2 and 5.3) are not dominated by the recordings from a single session.

The importance of the difference between the mean responses for the discrimination is confirmed by the fact that the separability vanishes for the most part after removing the mean activity. Therefore, I estimated the mean response of the different classes for each session and signal segment and afterwards subtracted them from the raw data. Interestingly, the separability is preserved, for the LFP data shortly before the monkeys reported their decision (70 ms pre-response time). This means that the second perception-related differences, near the response time, include higher order statistics. Therefore, the serial correlations in the local field activity (LFP) transmit information about the perceptual decision. According to Figure (5.4) I compared the mean distance from each class to the origin. After it, the distance $|m_1 - m_2|$ of the centres is due to rotation and translation. In the LFP data, during rivalrous and non-rivalrous stimulation (shortly before the response time), the distance from the origin is greater during stimulation with a horizontal grating. This implies that these stimulations lead to response amplitudes (on average).

This is confirmed by another study in which I have analyzed the mean power in the multi-channel signals $p = 1/2T \ \sum_i \int_{-T}^{T} x_i(t)^2 dt$ ($i =$ electrode index). Notice, putting all electrode recordings together corresponds to a reduction to one dimension. This kind of feature extraction had a pronounced, destructive influence on the segregation. For the most part, the separability goes below the significance level (data not shown), which means that the effect cannot be explained by a difference in the mean activity. In general, the reduction is more pronounced to the MUA than to the LFP data.

In addition, I compared the multi-channel signals after the extraction of the mean power for each electrode separately $1/2T \ \int_{-T}^{T} s(t)^2 dt$. In this case, the dimensionality is reduced to the number of electrodes, used in the multi-channel arrangement. Ignoring the temporal structure in the single electrode signal segments by using the mean power for the discriminant analysis has an influence on the segregation as well. But this effect is not as pronounced as in the previous situation in which the multi-channel activity has been pooled together. With respect to the temporal dynamic (200 ms pre-response time) only a minor information loss occurred (data not shown). Generally, the separability is reduced between 5 to 30 $\%$. The minimal separability appears in the LFP data in both conditions about 120 to 80 ms pre-response time. The corresponding reduction in the congruent condition for the MUA data is low. From this, I conclude that the separability is dominated by the mean power of the amplitude signals, which is in accordance to the findings in [94]. The importance of different frequency ranges will be discussed in Section 5.3.2.

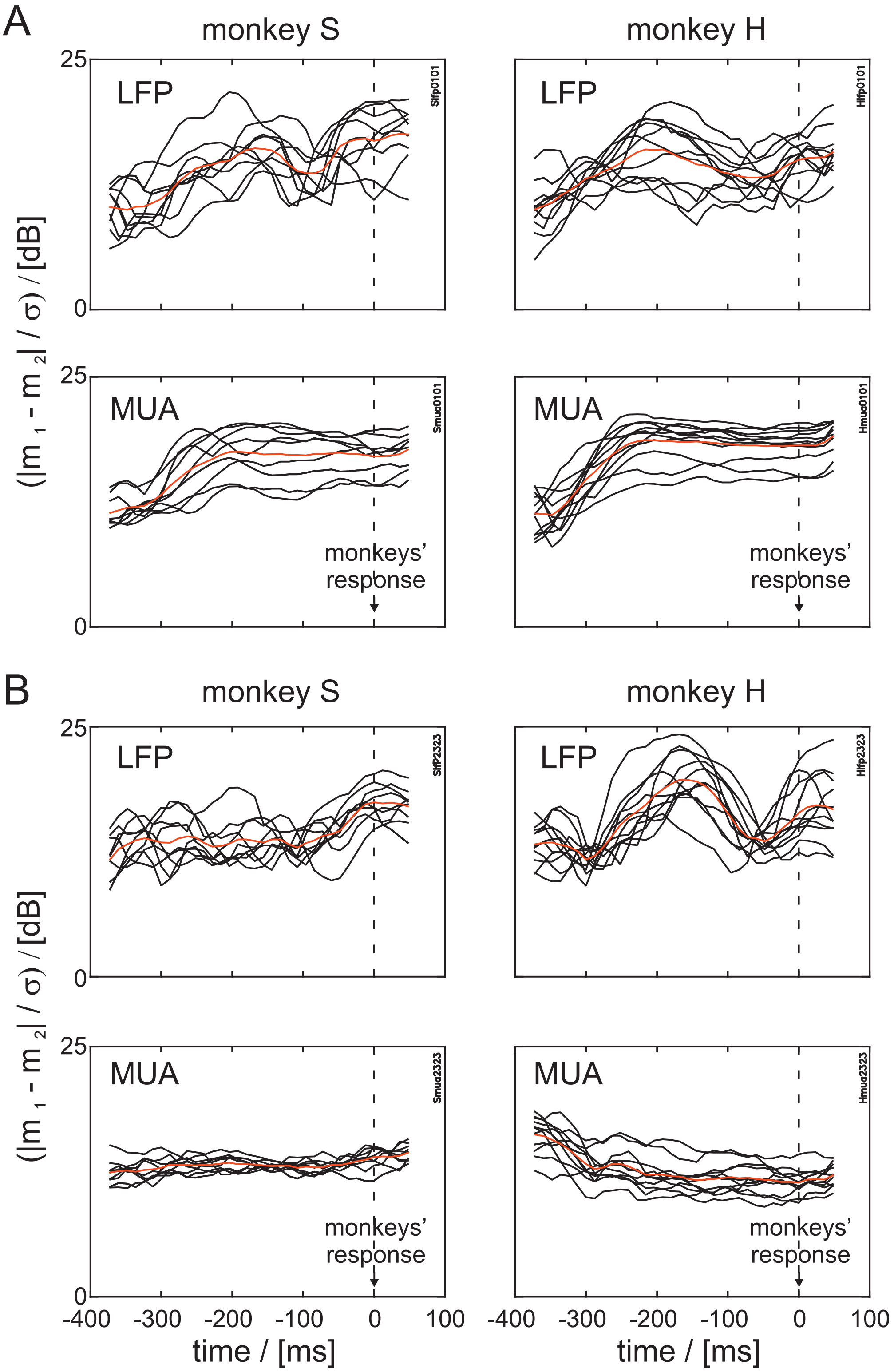

Figure 5.5: Relative distances of the mean values (SNR) between the multivariate data sets. Temporal variation of the relation between the data in the non-rivalrous (**A**) and the rivalrous condition (**B**) expressed by the distance of the mean values referred to the mean uncorrelated variance. The comparison was performed for each session, separately. Left column: results from monkey S. Right column: results from monkey H. Red curve: Mean over the 9(11) sessions from monkey S(H).

5.3.2 Time-Frequency Discriminant Analysis

The question, whether specific frequency ranges are important for the recognition and perception processes is currently part of an active investigation. In the following Section, I perform a time-frequency multi-channel discriminant analysis. By this, I wanted to find out which frequency ranges are involved in the segregation discovered in Figure (5.2) and (5.3). In accordance to previous studies sliding windows of 128 ms length-epoch with 16 ms temporal shifts have been used for the short-time Fourier analysis. At first, the mean values were estimated and subtracted from each signal segment. After that the modified signal segments have been multiplied with a Hamming window and transformed into the spectral range via Fast Fourier transformation. Discriminant analysis has been applied to the real and complex components for each frequency bin (frequency resolution 7.8 Hz). According to the number of electrodes, the overlap of the distributions in a 22(26) dimensionality feature space for monkey S(H) have been approximated by the RBF method.

The results of this study can be seen in Figure (5.6). As I expected, in the non-rivalrous condition significant differences occur more than 250 ms before the monkeys pushed the lever (Fig. (5.6) A). While the differences in monkey S are primarily restricted to low (4-12Hz) and medium frequencies (12-28Hz), there are also significant differences in the high frequency (28-90Hz) range for monkey H. Comparing the time-frequency discriminant maps of the LFP and MUA data the maximal ROC-values are higher in the LFP data max(LFP) = 0.75(0.73) to max(MUA) = 0.48(0.60) for monkeys S(H). This is interesting, because the time-resolved ROC-curves of Figure (5.2) reach higher ROC-values for the MUA data max(MUA) = 0.91(0.96) to max(LFP) = 0.84(0.87). On the other hand, the spread of the MUA separability is more expanded than for the LFP data. This may be the reason why the segregation in the MUA data reaches higher values in Figure (5.2).

The time-frequency analysis of the rivalrous data revealed that the differences are primarily restricted to low and medium frequencies (Fig. 5.6 B). The perception-related components in the LFP data (Fig. 5.2 B), shortly before the response time, are primarily restricted to low frequencies. Therefore, the early perception component (250 to 150 ms pre-response time) is evoked by differences in the low and medium frequency range. In addition, the LFP data of monkey H contain a significant difference in the high frequency range (70 to 90 Hz). In accordance with the results from (Fig. 5.3 B), the time-frequency discriminant maps of the MUA data show no significant difference shortly before the monkeys reported their percepts. There is a significant effect (300 to 200 ms pre-response time) in monkey H, but a similar phenomenon was not found in the other monkey. These results are confirmed when looking at the amplitude density spectrum averaged over signals from all electrodes and all sessions for the different classes (data not shown). From this, it can be seen that the appearance of frequency components in monkey S are restricted to the range 4 - 40 Hz. As a consequence pronounced differences between the two conditions appear only in this frequency range. In contrast, the amplitude density spectrum of monkey H contains frequency components in the range 40 - 70 Hz with slight differences between the two rivalrous conditions.

Similar to the previous time-frequency multi-channel discriminant analysis with short time windows of 128 ms, I performed a time-frequency multi-channel discriminant analysis with signal segments of 512 ms. First I estimated the mean values for each trial and removed them from the corresponding signal epochs. After the multiplication with a Hamming window the data have been transformed via FFT. I applied a cosines-tapered band-pass filter (8 Hz width, half height). In contrast to the previous calculations, I transformed the signals back into the time domain and performed the discriminant analysis in the time domain. The results of the time-resolved discriminant analysis are illustrated in Figure (5.7). As can be seen differences exist between this kind of analysis and the short-time Fourier analysis with 128 ms. It becomes clear that the frequency-resolved discriminant analysis with signal segments of 512 ms is much more elongated. The difference in the rivalrous condition (LFP monkey S) about 220 ms pre-response becomes significant, whereas the difference in the high frequency range (LFP monkey H) disappears.

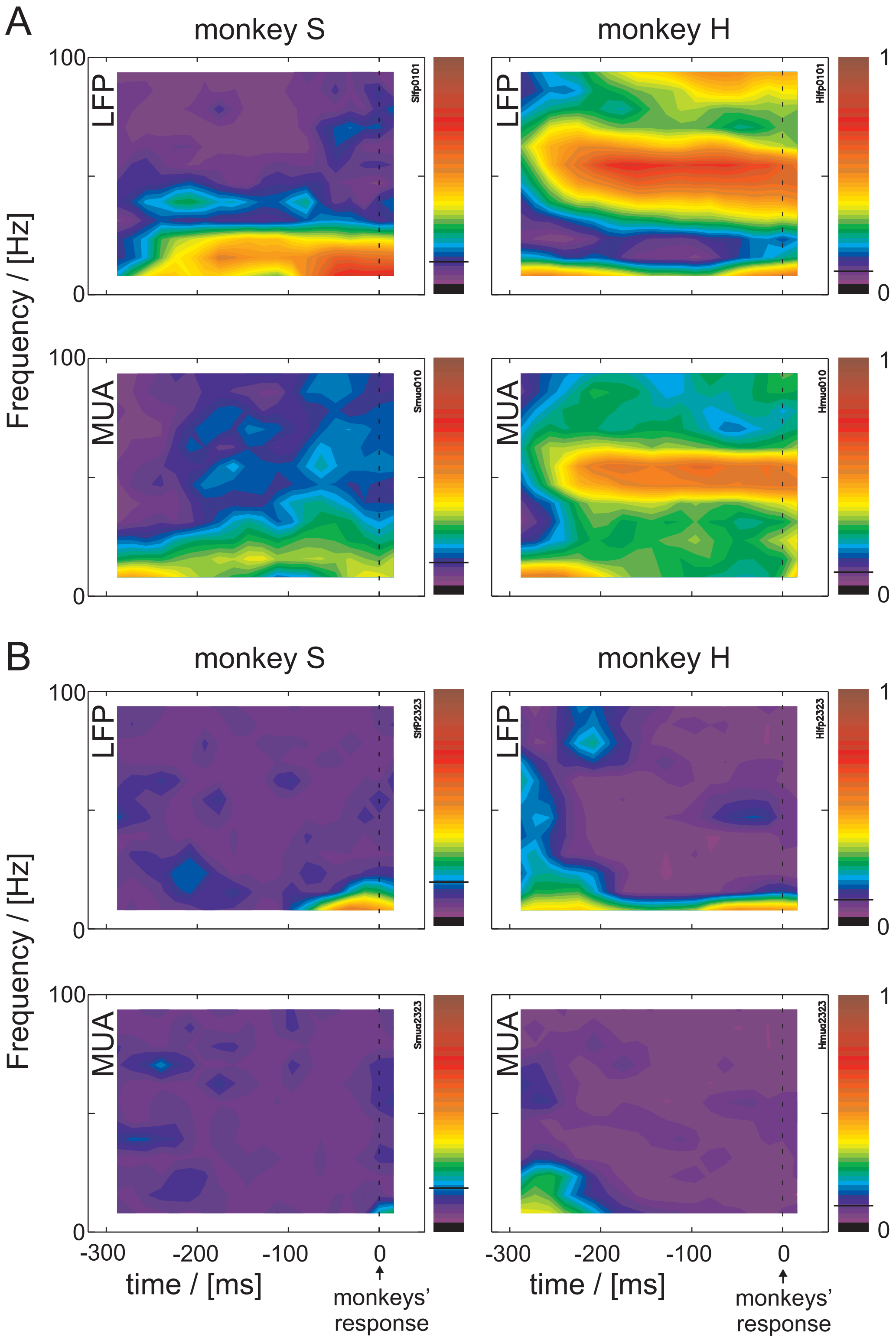

Figure 5.6: (**A**) Short-Time-Frequency discriminant maps (window length: 128 ms) of the response triggered multi-channel data (LFP and MUA) between the congruent conditions (I and II) for the two monkeys S and H. (**B**) Short-Time-Frequency discriminant maps of the response triggered multi-channel data (LFP and MUA) between the incongruent conditions (III and IV) for the two monkeys S and H. Color-coded: Separability, quantified by twice the area between the receiver operating characteristic (ROC) and the diagonal of the unit square. The corresponding significance level $p < 0.01$ is indicated by the horizontal lines in the color tables on the right side of the curves.

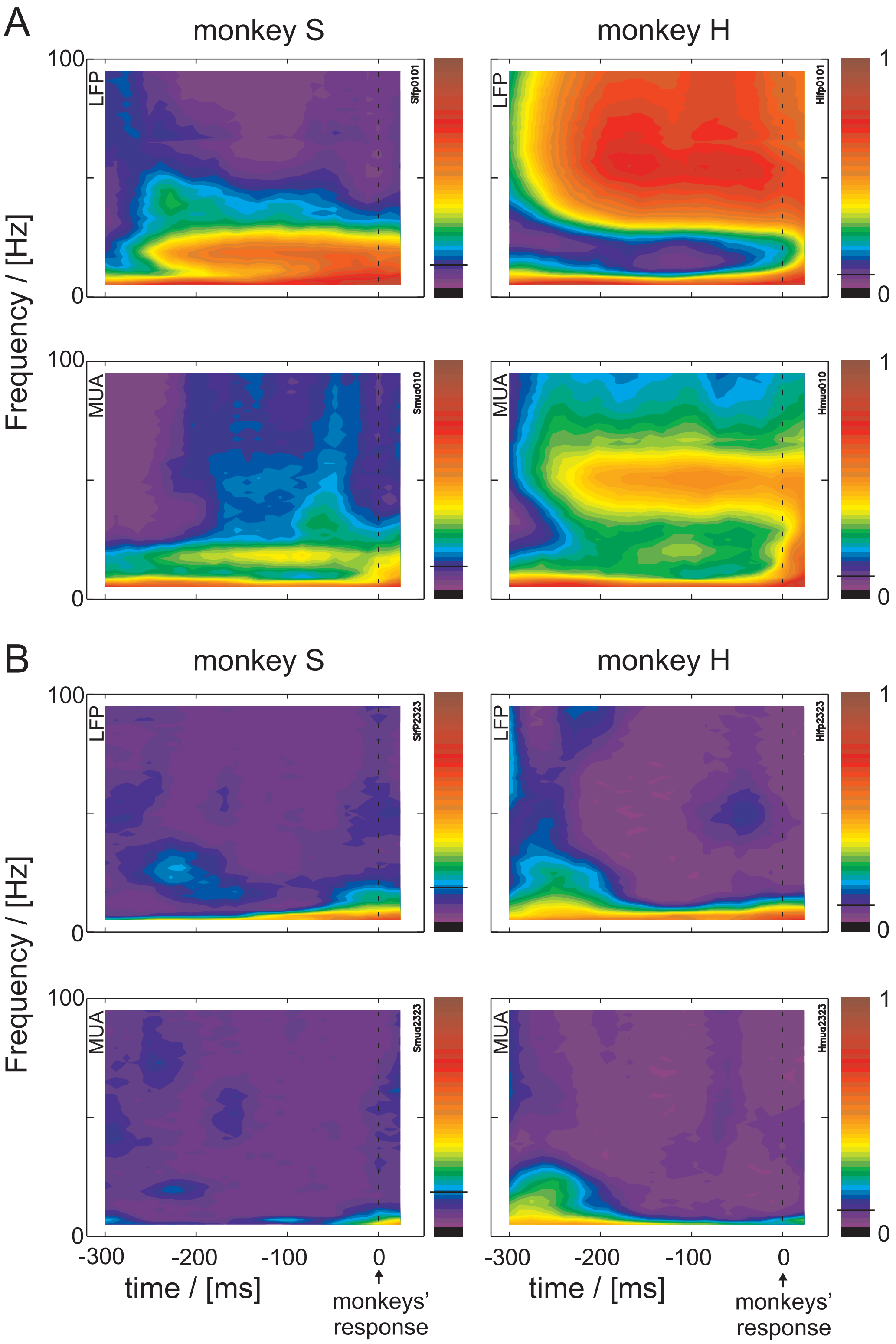

Figure 5.7: Time-frequency discriminant maps (window length: 512 ms) of the response triggered multi-channel data (LFP and MUA). Comparison has been done in the time domain after bandpass filtering in the frequency domain. (**A**) Discrimination between the congruent conditions (I and II) (**B**) Discrimination between the incongruent conditions (III and IV). Color-coded: Separability, quantified by twice the area between the receiver operating characteristic (ROC) and the diagonal of the unit square. The corresponding significance level $p < 0.01$ is indicated by the horizontal lines in the color tables.

5.3.3 Cross-Correlation Analysis

Further insight in the relevant features of the recorded neural signals may be obtained by investigating their global coupling structure. For this, the cross-correlation coefficient matrix has been computed trial-by-trial:

$$M = (\frac{< x_i - \bar{x_i}, x_j - \bar{x_j} >}{|x_i - \bar{x}||x_j - \bar{y}|})_{i,j=1\cdots p} . \quad (5.1)$$

The vectors $\bar{x_i}$ and $\bar{y_j}$ are the means of the sample populations per session: $\bar{x_i} = \frac{1}{n}\sum_{k=1}^{n} x_i^k$ and $\bar{x_j} = \frac{1}{n}\sum_{k=1}^{n} x_j^k$, respectively [146]. Due to the normalization all matrix entries are in the range $[-1, 1]$. I used signal segments of 128 ms length (n=64) for the estimation of the pairwise correlation in accordance to the work of Gail et al. [94]. For comparison, only the upper parts of the matrices were used because of their symmetry. The analysis has been done on the simultaneous recordings from monkey S (number of recording channels: 11) and from monkey H (number of recording channels: 13). According to this, discrimination has been carried out in a 55 (S) and 78 (H) dimensional space.

From the results in Figure (5.8) it can be seen that there are significant differences over a wide range (up to 300 ms pre-response) between the pairwise couplings in the congruent condition. In comparison to the results in Figure (5.2 B), all curves remain at lower values, indicating some sort of information loss. Interestingly, the values are not so low in the LFP discriminant analysis than in the MUA discriminant analysis. The transmitted information is reduced to 70 $\%$ for monkey H (MUA). It decreases to less than 50 $\%$ (0.1 bit) in monkey S near the monkeys response time.

Looking at the results corresponding to the incongruent condition (Fig. 5.8 B) in both monkeys (up to 100 ms pre-response time), significant differences appear in the LFP data. As before, comparing the two rivalrous data sets but looking at the linear pairwise coupling lead to an information loss. This information loss is strongest near the response-time. In addition, the temporal dynamic in Figure (5.3 B) is not visible after the restriction to the pairwise coupling. From these results, prediction from the coupling matrices is not recommended.

5.3.4 Parameter Dependency of the Discriminant Analysis

In the following, I analyzed the influence of the number of recorded signals and the window duration to the discrimination performance. Since the effects are similar for the different conditions and signal types, only the results for the LFP recordings of monkey S in the rivalrous condition are plotted. Figure (5.9 A) shows the time-resolved discriminant curves for window durations from 8 ms to 128 ms. From this depiction you can see that the segregation increases slightly with increasing window duration. Therefore, I excluded that the reported differences are a side effect of the chosen window duration. Furthermore, the statement that there are two perception related components is confirmed by this illustration. The stability of these effects is interesting, since the discrimination has been done in very different dimensions. For a window duration of 128 ms discrimination has to be done in a more than 700 dimensional space. In contrast, the discrimination of 8 ms signal segments results in a 44 dimensional problem.

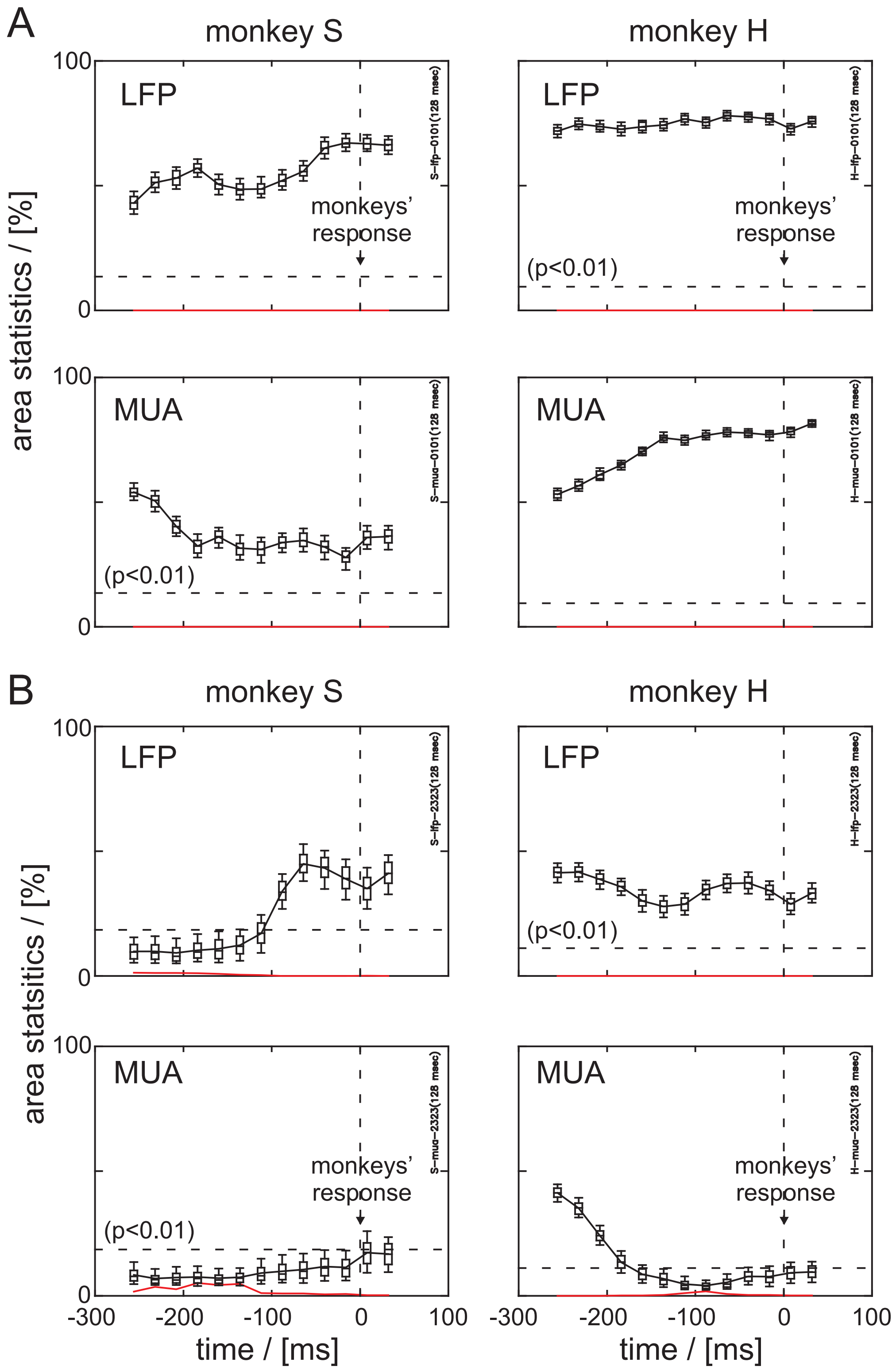

Figure 5.8: Time resolved discriminant analysis based on the correlation coefficient matrices (upper part) estimated by signal segments of 128 ms. (**A**) Comparison of the congruent conditions (I and II). (**B**) Comparison of the incongruent conditions (III and IV). Horizontal dashed line: significance level with $p < 0.01$. Red curves: One side dominance expressed by twice the area below the receiver operating characteristic.

Although the discrimination is similar for different window sizes of LFPs this is not the case for the pairwise correlation analysis. I repeated the analysis from the previous Section and varied the window duration for the estimation of the pairwise coupling between 8 ms and 128 ms. The results are shown in Figure (5.9 B). At first, up to a window size of about 50 ms no significant segregation appears. Second, shortly after the response time separability is slightly reduced with long signal segments ($>$ 100 ms). Third, the maximum changes its position (see also Section 4.4.2). From the comparison of Figure (5.9 A) and (5.9 B) you can see that the discrimination according to the coupling matrix is only significant near the response time. No pronounced segregation can be found 200 ms before the monkey reports his percept. During 300 to 150 ms pre-response time, the discrimination values are always under the significance level, independent from the window duration.

In the next study, I investigated the influence of the number of recording channels to the discrimination. For this, I chose randomly a fixed number of single recordings from each session. After that, I averaged the projected samples over the different recording sessions (see Section 5.2). According to the window size (48 ms), the dimensionality increases from 24 dimensions for one recording channel to 264 dimensions for all recording channels. Like before, 200 cross validations have been repeated for each signal segment. An illustration of the results can be found in Figure (5.9 C). As you can see, the segregation by the data from one cortical position is near the significance level in which the standard deviation (not shown) is about ± 0.006. On average, one recording channel transmitted up to 0.04 bit near the decision. Therefore, the prediction for the monkey's response is near chance. This means, up to 60 $\%$ may be correctly classified but about 40 $\%$ may be misclassified by the signals from one electrode. A better segregation is reached by increasing the number of simultaneously analyzed signal courses. Already with the signals from two electrodes, discrimination becomes much better. The lowest Bayes error belongs to the analysis in which the signals of all recording sites have been used simultaneously. Near the monkeys' response time, on average, 76 $\%$ can correctly be classified. Notice, the monotone increase with combining the signals from different recording channels indicates that the decision is not dominated by a small fraction of all recording channels. Furthermore, All these results confirm our assumption that the different recording sites transmit perception-related information and that the observed effects are not a local phenomenon of only a few electrodes or belong to single sessions. For this stimulus configuration, the simultaneous analysis of all available recording sites provides the lowest classification error. The differences, in terms of segregation, between a single recording channel analysis and the multi-channel approach can be seen in Figure (5.9 C).

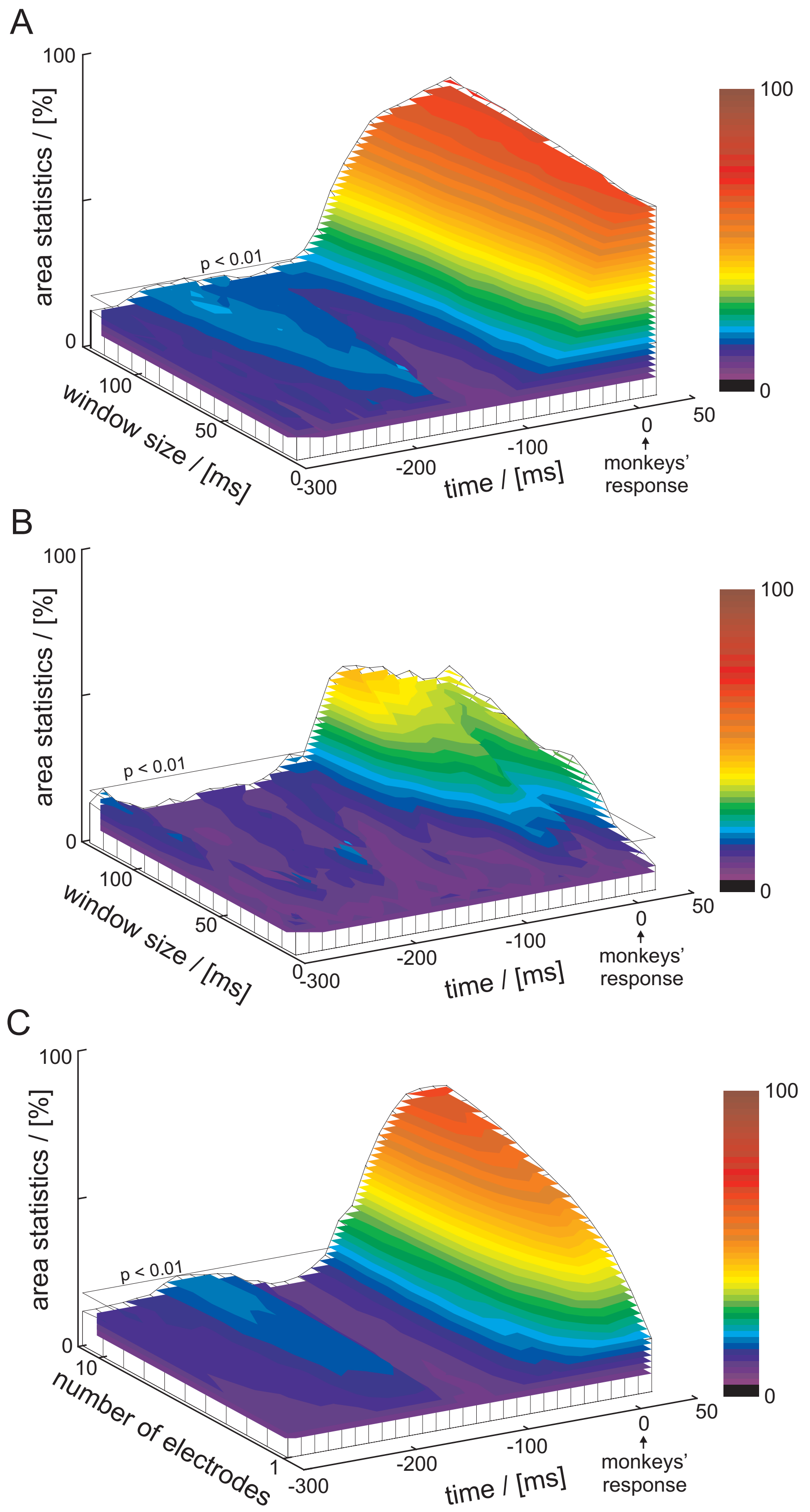

Figure 5.9: Data from monkey S (LFP) during binocular rivalrous stimulation. (**A**) Discrimination of binocular rivalry multi-channel amplitude data depends not on the window size. (**B**) The discrimination of the multi-channel correlation data depends on the window duration. (**C**) The significance of the discriminant analysis is effected by the number of simultaneous recording channels.

5.4 Stimulus Invariant Components

The nonlinear RBF method enables us to take the spatio-temporal dependencies more accurately into account, leading to a significant discrimination between two data sets. In the next Section I will go one step further. For this I combined the data from the rivalrous and the non-rivalrous condition. With regard to the monkeys' perceptions, two situations can be distinguished: perception of a horizontal or vertical grating. We want to know whether there exists any relation between the monkeys' perceptions that are independent from the underlying stimulus arrangement (congruent, incongruent). At the neuronal level, one can ask whether there are stimulus-invariant perception-related components in the different multi-channel recordings. A direct comparison of the rivalrous and the non rivalrous data by the discrimination approach would make no sense, since the signals are strongly influenced by the different visual inputs. In addition to the previously applied dimension reduction approach, I used a trick to solve this problem.

5.4.1 Perception-Related Co-Modulation

In contrast to the previous discrimination studies which used disjoint subsets of the same data for learning and testing, now I combine different sets of data. Recordings during congruent (incongruent) stimulation were used for the training of the RBF function. Testing was done with recordings during incongruent (congruent) stimulation. A pleasant side-effect of this procedure is that I can use the whole data set at once for learning. In the same way all trials can be used in the congruent (incongruent) condition for testing. As you can see in Figure (5.10) 300 to 100 ms pre-response) accordance between the relation in the rivalrous and non rivalrous condition is mostly not significant. There is a slight modulation in the data of one monkey, which disappeared when I used the data from the rivalrous condition for training. As a consequence, during this period of time the corresponding neural processes must be different. With respect to the time interval up to 80 ms pre-response, a significant perception-related co-modulation is indicated in the data of both monkeys.

Based on the multi-electrode LFP data, shortly before the response time, correct single-trial prediction from the congruent signals of the monkeys' reaction under binocular rivalry is on average 78(68) $\%$ for monkey S(H). I want to mention that the prediction according to a single-channel discriminant analysis is near chance. As in the incongruent condition, the effects in the MUA data are not compatible with those from the LFP recordings. There is a significant co-modulation in the data from monkey S, but this effect can be seen only in the situation where I used the rivalrous data for training. However, in both cases, the discrimination is above the significance level for monkey H, and it contradicts the one-side dominance. Therefore, this form of co-modulation is a side-effect and was rejected. Notice, in all other cases, the contradiction to the one-side dominance is below the significance level. In contrast to the LFP data, a reliable prediction was not possible from the MUA data.

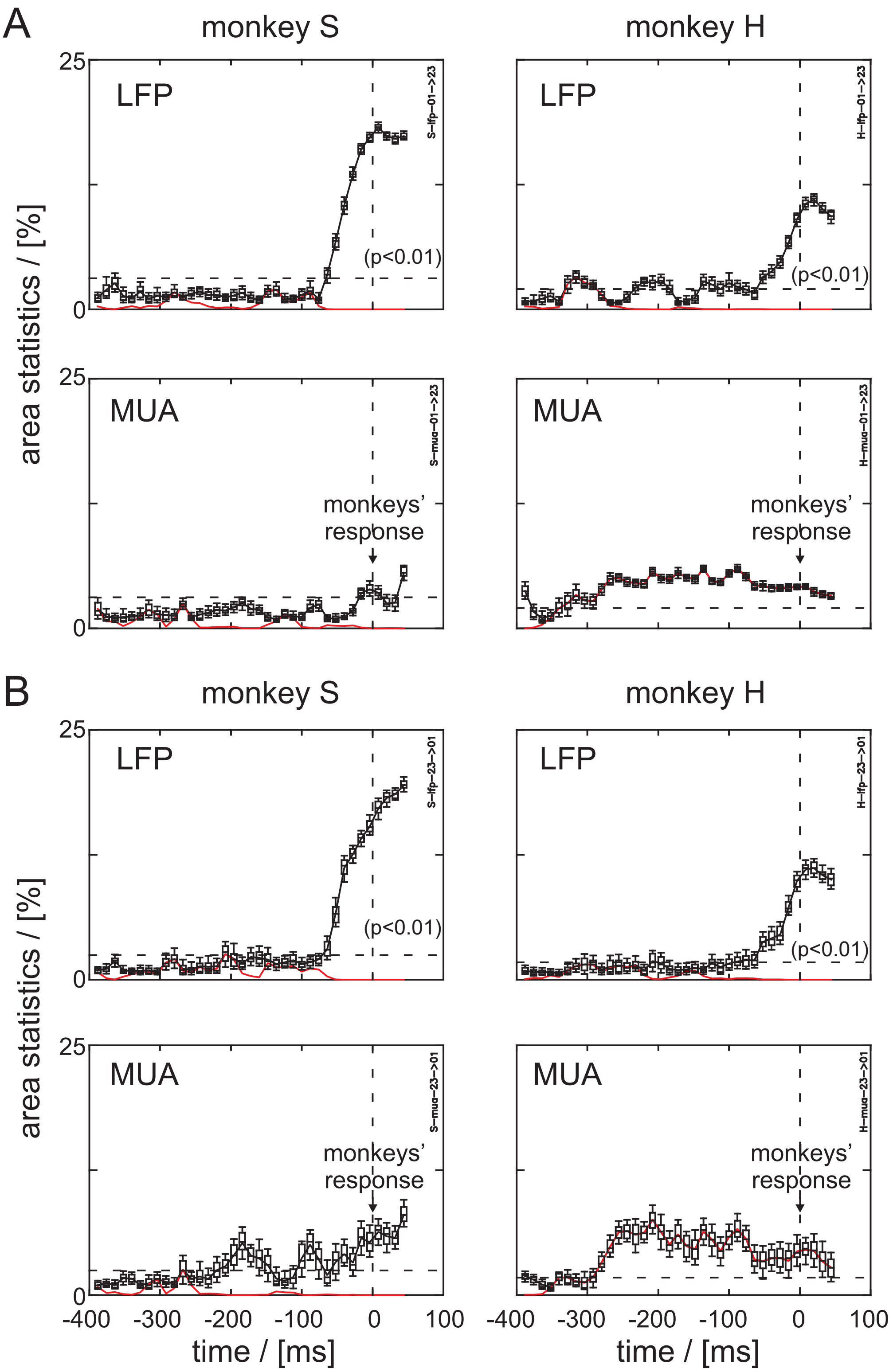

Figure 5.10: Comparison of the data from the congruent condition and the incongruent condition. (**A**) RBF approximation has been done using the data during non-rivalrous stimulation. Test statistics have been carried out using the rivalrous multi-channel recordings. (**B**) Reversed situation to (**A**). The RBF map was build by the recordings during rivalrous stimulation. Test statistics have been done using the recordings during non-rivalrous stimulation. The significance level of $p < 0.01$ is marked by the horizontal dashed line. Red curve: Contradiction to the one-side dominance quantified by twice the area under the receiver operating characteristic.

5.4.2 Multidimensional Scaling

In addition to the previous comparison, I mapped the multi-channel recordings near the response time to the two-dimensional space by a nonlinear dimension reduction approach called *multidimensional scaling* (MDS). MDS seeks to find a low-dimensional representation of objects such that the interpoint distances (after projection) are as close as possible to the given dissimilarities [262, chapt.9.4]. In order to judge whether one low-dimensional configuration fits the original configuration of points better than another, various error functions (also called STRESS functions) are considered. From the vast literature I concentrate on Euclidean (metric) MDS also known as Sammon's mapping [208]. To measure the fit between dissimilarities and distances, the error function is given by:

$$STRESS = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (\delta_{ij} - d_{ij})^2 \ , \tag{5.2}$$

where δ_{ij} corresponds to Euclidean distances between two points in the high-dimensional input space and d_{ij} represents the distance between two points in the low-dimensional $(d = 2)$ space. The indices i and j indicate recordings from different trials. the coefficient ω_{ij} is a given non-negative normalization weight. In order to avoid the situation in which larger distances have a stronger influence than smaller distances, I have chosen $\omega_{ij} = \frac{1}{\delta_{ij}^p}$. For the estimation of the global minimum, I used the distance smoothing approach proposed by [27]. This technique performs well in comparison to, e.g., simple gradient methods [165].

The results, divided for the different recording sessions and stimulus-perception situations, can be seen in Figure (5.11) and (5.12). The location and orientation of the data with vertical perception are similar during many sessions (monkey S). In the data with horizontal perception, you can see that the orientation of the projection is often different but the point spread is similar. The results from monkey H are even less clear and the similarities between the four stimulus-perception reactions are less pronounced. In the congruent condition with horizontal perception, the point spread is largest. This is not the case in the incongruent condition with the same perception. There is a common trend in favor for the data with vertical perception but not as clear as in monkey S. Comparing these results with the results from the RBF method one has to notice that MDS is an unsupervised method. Furthermore, the results are influenced by the metric. MDS tries to illustrate the spatial extension of high-dimensional features into two-dimensions, but there is no guarantee that this is appropriate. Even simple geometries in high dimensions show a complex pattern in two-dimensions.

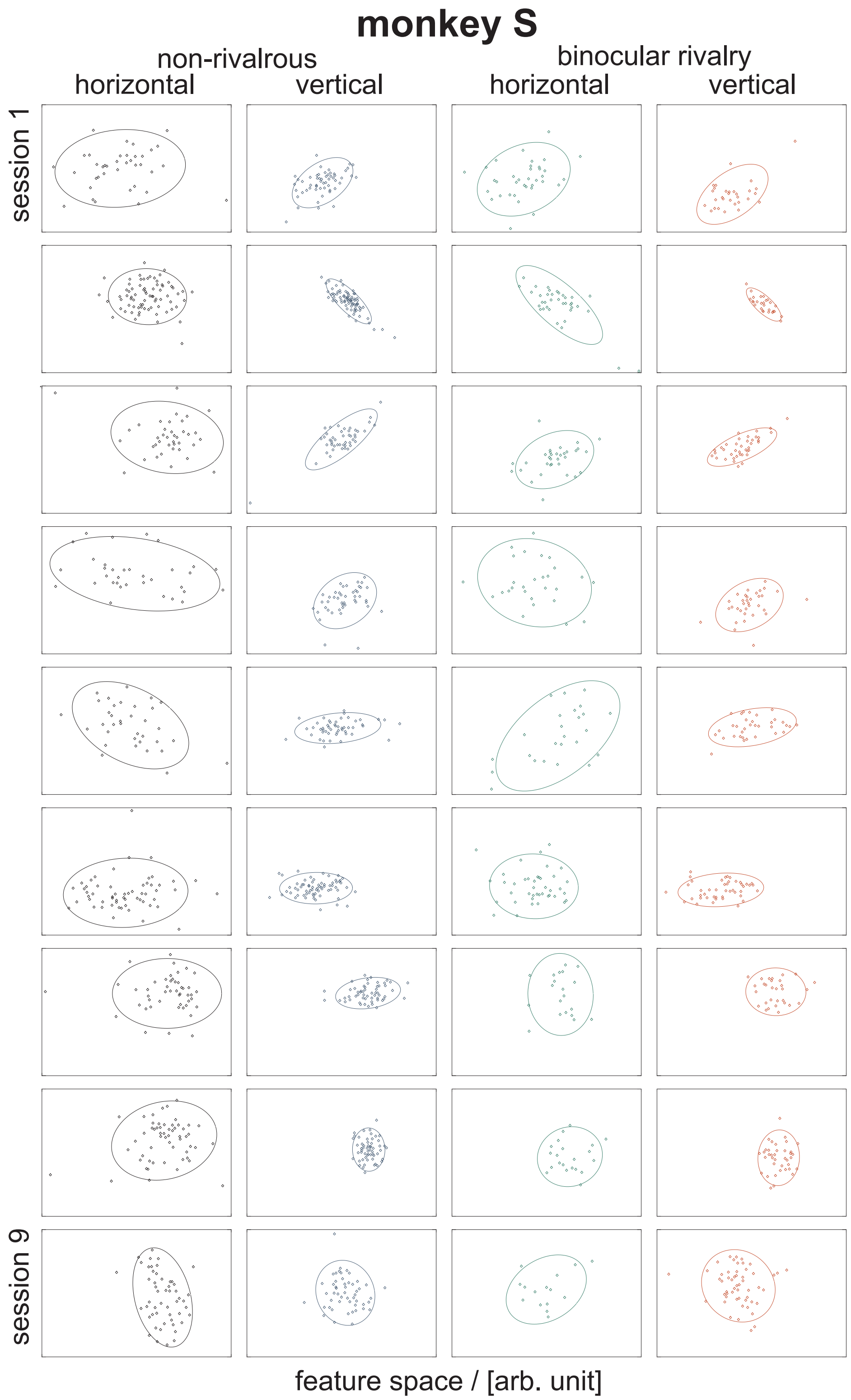

Figure 5.11: Projection to a two-dimensional feature space of multi-electrode local field potentials from monkey S near the response time (-24 to 0 ms pre-response time) by multidimensional scaling in combination with principal component analysis. Each point represents a single trial. Each row corresponds to a single session and each column to a specific stimulus condition. The ellipses illustrate the orientation of the principal components.

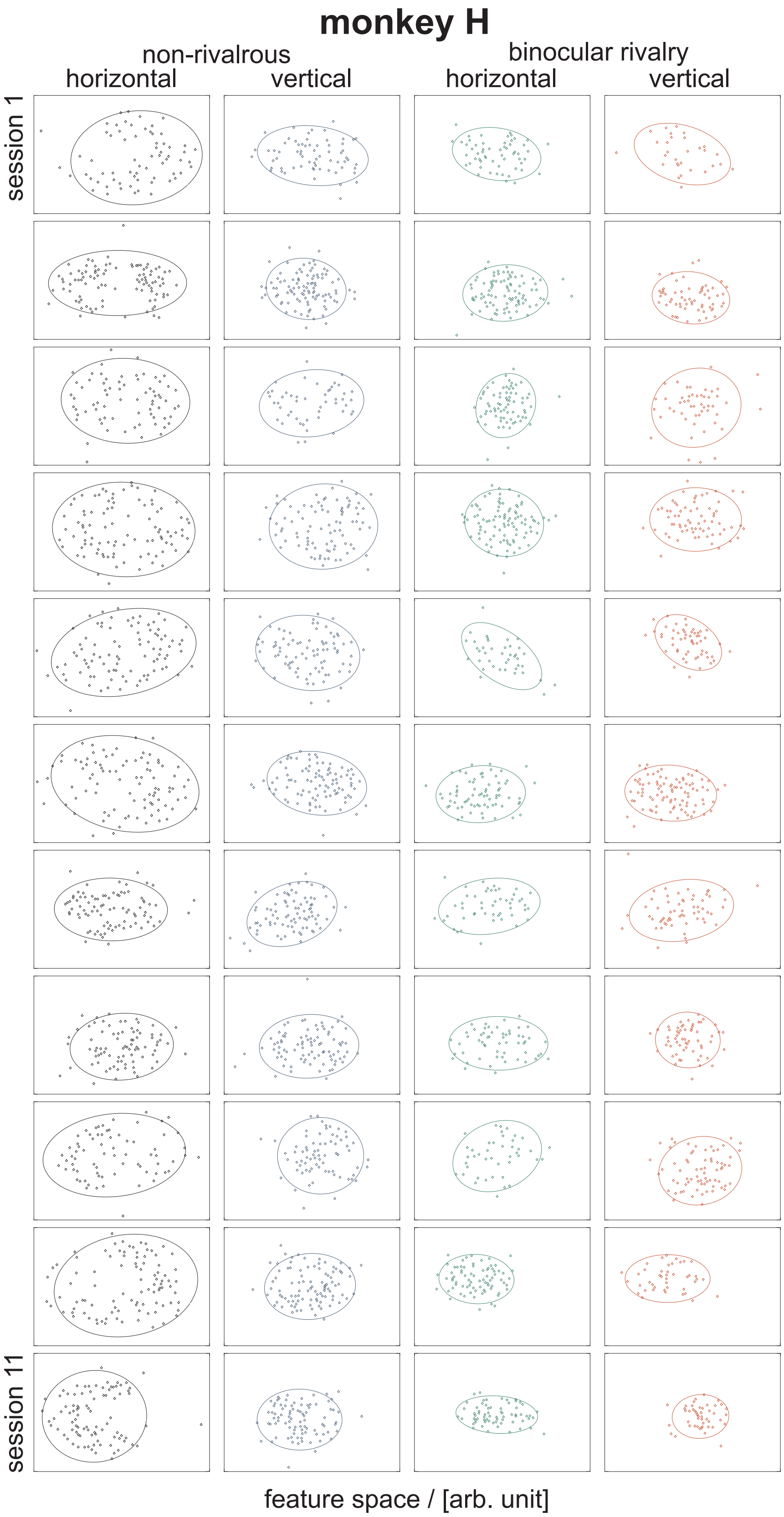

Figure 5.12: Corresponding results of the multidimensional scaling for local field potentials from monkey H. Each row stands for a single session and each column stands for a special stimulus condition (see Fig. 5.11).

5.5 Discussion

Main results. I have demonstrated that the RBF dimension reduction approach, applied to real data from multi-channel recordings of neural signals, works well with respect to sensitivity and robustness. Moreover, the method can bring further insight in the functional role of the primary visual cortex. One reason for the high sensitivity can be explained by the fact that this approach takes the spatio-temporal statistical dependencies of all recorded channels simultaneously into account. The benefits of the RBF method against classical single-channel approaches became clear from the discriminant analysis of spatio-temporal distributed processes. Although dimension reduction is affected by instationary recordings and although it does not deliver a model of cortical information processing, this nonparametric dimension reduction method opens promising ways for further research. In the next three Sections I will discuss various aspects with respect to the analyzed multi-channel recordings.

5.5.1 Binocular Rivalry in V1

In the previous Sections, I described several findings of the cortical dynamics in V1 during binocular stimulation. Some findings confirmed results from previous single electrode and pairwise correlation studies. For example, Gail et al. [94] found for the same set of data that the number of perception-related recording sites in the low-frequency LFP power increases near the time of the monkeys decisions (see also Fig. 5.6 and 5.7). They argue that this phenomenon may be explained by a feedback impact on V1 from other cortical areas already representing or carrying a prediction about the oncoming perceptual state. The notion of an influence from other cortical areas on V1, being responsible for the observed perception-related modulation around the response time, is possible, since many neural feedback connections terminate in layer 2/3 of V1 (in addition to infragranular layers) and are typically modulatory in nature [207]. In addition, studies have shown that such feedback projections can act as fast as the internal processing in area V1 [127]. An origin of this modulation may be higher cortical areas, e.g., from inferotemporal or prefrontal areas. Sheinberg and Logothetis have shown that binocular rivalry in monkeys may be partially explained by the cortical activity in the inferotemporal cortex [225]: they found that single cell spike rate modulations were almost always correlated to perceptual switches in their study.

Besides the perception-related modulation around the response time, the high temporal resolution of the multi-channel RBF projection approach revealed information about the temporal dynamic in the rivalrous conditions that has not been found by classical studies. From the temporal dynamic in Section 5.2.2 (see also Fig. 5.9), you can see that the discrimination reaches two local maxima. One maximum appears around the response time, which is in accordance with previous observations [94]. The other maximum appears 250 to 150 ms pre-response time, which has not been reported before. At the moment, it is not clear which are the main reasons for the significant distinction in the rivalrous condition about 250 to 150 ms pre-response time. What we know is that ocular dominance of one eye above the other eye can be excluded by the work of Gail et al. [94]. One explanation for the local maximum, may be differences in the lateral interaction, probably evoked by attention-driven modulations. Differences in the mean reaction time for the two perceptions may be as well contribute to the segregation about 200 ms pre-response time. In this context, Gail et al. have investigated the mean reaction time in the rivalrous condition with respect to the monkeys perception [94]. They found that the mean reaction time was different in both monkeys perceiving a vertical or horizontal grating. For monkey H, mean reaction time perceiving a vertical grating is significantly longer, whereas for monkey S, perceiving a horizontal grating took more time on average. The absolute difference between horizontal and vertical perception was about 90 ms (S) and 130 ms (H). Unfortunately, within the multi-channel discriminant analysis it was not possible to remove the temporal shift in the reaction time, because the number of trials was to small for a significant analysis.

The comparison of the two local maxima revealed that in contrast to the assumption of a perception-related feedback modulation there exists no relation between the signals under rivalrous and the signals under non-rivalrous conditions about 200 ms pre-response time (see Section 5.4.1). Furthermore, from the multi-channel correlation coefficient discriminant analysis (Fig. 5.8) a significant segregation of the recordings during rivalrous stimulation (200 ms pre-response time) was not obtained. From these findings, I conclude that the two local maxima belong to separate cortical processes but take an active role in aware perception.

5.5.2 Information Transmission by LFP and MUA Signals

The results of the discriminant analyses from the multi-channel local field potentials and the multi-unit activity showed pronounced distinctions. What are the reasons for the different behavior? Well, MUA and LFP represent different signal types. MUA comprises of superimposed spike densities, mainly originated at the axon hillock, whereas LFP reflects a weighted average of dendro-synaptic signal components. The catchment volume of the MUA signal is more local than the LFP signal (for further information about these signal types see Section 5.1). As a consequence both signal types decode different aspects of cortical information processing. In the following, I will concentrate myself on three discovered distinctions in more detail.

1. Within the congruent condition, the Shannon information estimated from the multi-channel MUA signals reached higher values compared to the multi-channel LFP signals (Fig. 5.2 B).
2. For the multi-channel LFP recordings the time-resolved discrimination between the non-rivalrous stimuli indicated a local minimum about 100 ms pre-response time. The discriminant analysis of the corresponding MUA signals, showed no corresponding minimum.
3. The comparison of the multi-channel MUA signals in the rivalrous condition showed less significant perception-related components. In contrast, from the multi-channel LFP recordings a consistent and significant perception-related distinction was obtained before the monkeys pushed the lever.

The first point, may be explained by the fact that the MUA signals, recorded in V1, decode essentially information about the visual input. As a piece of circumstantial evidence for this claim think of the difference in the orientation tuning between MUA and LFP. In most cases one obtains a much sharper orientation tuning profile from the MUA signals. The second point, may be explained by the larger catchment volume of local field potentials. As I described in Section 5.1, the multi-channel recordings have been obtained from the supragranular layers in the primary visual cortex.

Layer 2/3 of V1 contains cells that have long-range lateral connections and these intralaminar connections are clustered [270]. According to the larger catchment volume of the LFP signal, neurons from other areas more apart influence the recorded signals. The third point, may be explained in a similar fashion. As has been pointed out by Gail et al. [94] the difference in the rivalrous condition between horizontal and vertical perception shortly before the monkeys report their perception, may be triggered by a feedback-modulation from other cortical areas. It is possible that the LFP signal contains such modulatory (perhaps sub-threshold) signal components.

5.5.3 Single versus Multi-Channel Approaches

The topographical organisation of the primary visual cortex is subdivided in local information processing modules, which makes it possible to study functional properties of the cortical neural network with micro-electrodes arrays. Classical approaches, which investigate each recording channel separately, are not able to detect spatial interactions between these modules satisfactorily. The multivariate discriminant method, that has been developed in Chapter 2, offers a way to investigate the spatio-temporal cortical signal processes, detected by multi-channel recordings, in a more efficient and sophisticated manner. In terms of discrimination, the advantage of the multi-channel approach over classical single channel approaches becomes evident, e.g., from Figure (5.9 C). As you can see, the segregation from single channel LFP recordings (rivalrous condition) is near the significance level. On average, at best 60 $\%$ are correctly classified, what means at least 40 $\%$ are misclassified by the signals from one electrode (on average). A better segregation was reached, increasing the number of simultaneously analyzed electrodes. In the case when the signals from 10 electrodes have been analyzed simultaneously, the classification error decreased to 25 $\%$ near the response time (see also [150]). Under the assumption that the different electrodes transmitted the same information, a simple extrapolation, fitting the data by an exponential decay $exp(-sqrt(0.045518 * x))$ ($x =$ number of electrodes), reveals that from the signals of a 10×10 electrode array, the classification error goes down to 5 $\%$. Accordingly, in order to obtain a classification error below 1 $\%$, the simultaneous information from more than 300 electrodes would be necessarily. Applying the same restrictive assumption for an extrapolation of the Shannon information means that 30(250) electrodes are necessary to transmit more than 0.5(0.99) bit. For the corresponding mapping I used the following parametric equation: $1 - exp(-a * x^b)$.

In terms of discrimination, investigating all recording channels at once has some advantages. On the other hand, from the multi-channel analysis it is not possible for example, to determine which electrode transmits most of the information. I mentioned previously, that the concentration on the relevant features can increase the separability and reduce the curse of dimensionality (see also Section 6.2). Unfortunately, the computation of all electrode combinations is impractical. For example, if data have been recorded with 10 micro-electrodes simultaneously, more than 1000 constellations have to be tested, to find the most discriminative electrode combination.

In order to counteract this combinatorial explosion, at least three strategies are discussed in the literature (see also Chapter 6). The first one is to look at each recording site separately. One drawback of this strategy may be that in a nonlinear situation with statistical dependencies between the recording channels, the discriminant analysis could indicate only negative results. The second strategy is similar to the leave-one-out approach, which has been successfully applied in [146]. Suppose n electrodes are used in an experiment. Looking at the n configurations consisting of $n-1$ electrodes and comparing the results with the segregation achieved by all electrodes gives a hint about the influence of each single electrode. Theoretically, the transmitted information should be a monotonically increasing function of the number of electrodes. As I have shown in Chapter 3, the curse of dimensionality has a pronounced influence for all methods. Therefore, the estimated information can decrease if an electrode does not contribute to the discrimination. However, an increase in the information indicates a gain in the discrimination. In front of the single electrode comparison, the leave-one-out technique has the advantage of taking the multi-channel dependence into consideration. To find the relevant features, it can be applied recursively. The numerical effort is at most of $o(n^2)$. The third method, which I will discuss in more detail in the general discussion of Chapter 6, is to embed the feature selection process in the discriminant analysis. In this method the information about the relevance of single recording channels is coded in the weight vector.

6 General Discussion

'Understanding is mostly luck: But you can act without understanding.' (Vladimir Vapnik, DAGM2004)

Chapter 6 serves as a critical reflection of the previous presented results. I will examine how the projection method can be generalized to multi-class problems. Further, I will discuss which opportunities exist for a fast and reliable feature extraction. In addition, I will describe which signal types might be handled in an analogous manner. The Chapter ends with a view to future research in this area.

6.1 Multi-Class Problem

In Chapter 1 I have defined the two-class problem in which there are two multivariate data sets to be segregated. In many neurophysiological recording experiments (and in everyday life) people are confronted with a much higher number of classes. Think for example about the number of faces of your friends which you can distinguish or the here evaluated binocular experiment, in which stimuli of additional orientations could be investigated as well. In general, there are three strategies to adapt the projection methods to multi-class problems.

The first strategy can be described as *one-against-all*, asking how each class relates to the rest. In other words, what makes your girlfriend different to all other persons? The exchange of the classes leads to a discrimination vector of the same size as the number of classes. Each component of the discrimination vector can be computed by the methods in Chapter 2. Theoretically, everything remains the same except that the number of samples in the two classes might be very different. Numerically, it might be better to adapt the cross-validation technique. The random sampling procedure should be applied to each class, so that the combined test set consists of samples from all classes.

The second strategy is to look for differences in a *one-against-one* manner. If the number of classes is n then the distinction will be coded in a symmetric $n \times n$ matrix. The numerical effort of this strategy grows quadratic with the number of classes. The diagonal elements may be picked up as stationary tests dividing the data of each class into two separate sets, e.g., at the beginning and at the end of a session.

6.1.1 Multi-Class Projection

Besides these two strategies, it is possible to generalize the dimension reduction process so that the projection is trained by the training samples of all classes simultaneously. In order to preclude the possibility that the discrimination result is influenced by the distance metric which is predefined by the indicator values, the projection space has to be $n-1$ dimensional. A useful configuration would be to map the training samples for the different classes to the edges of an equilateral polygon. I will illustrate this in the 3-dimensional space, given samples from four different classes. Without loss of generality, these samples could be mapped to the edges of a tetrahedron according to: $[g_1(x), g_2(x), g_3(x)] = [1, 0, 0]$ if $x \in C_1$, $[g_1(x), g_2(x), g_3(x)] = [-1, 0, 0]$ if $x \in C_2$, $[g_1(x), g_2(x), g_3(x)] = [0, \sqrt{3}, 0]$ if $x \in C_3$ and $[g_1(x), g_2(x), g_3(x)] = [0, \sqrt{3/4}, 3/2]$ if $x \in C_4$. As you can see from this example, the numerical effort grows linearly with the number of classes. Comparing this problem with the two-class problem one has to adjust more than one (non-)linear function $g(x)$. With n different classes, $n-1$ non-linear functions will be obtained. Therefore, instead of one system of linear equations, $n-1$ equation systems have to be solved for the adjustment of the coefficients. With the additional expenditure in computation another problem occurs. The formula and the algorithms for the estimation of the Bayes error and the Shannon information have to be generalized as well. For the multi-class problem ($n > 2$), Shannon information can be expressed by: $MI = h(p(x)) - \sum_c P(c)h(p(x|c))$ with $\sum_c P(c) = 1$ [47, sect.2.4]. In accordance with the two-class problem, the prior weighted differential entropy of the separate class's likelihood has to be subtracted from the differential entropy of the unconditional probability density function (see also Section 1.3 and 1.8). With n distinct classes, Shannon information is bound by the logarithm of the number of classes $log(n)$. The transmitted information will be zero if the output values are conditionally independent from the stimulus set: $p(x|c) = p(x)$, $\forall c \in C$. If each of the stimuli evokes a unique set of responses, i.e., the responses are different for different stimuli, then Shannon's information equals the entropy of the prior probabilities $-\sum_c p(c) log p(c)$. Numerically, the *a priori* probabilities $P(c)$ can be easily estimated. For the estimation of the individual differential entropy terms, I propose to use the nearest neighbor method of Kozachenko and Leonenko, as was shown in Chapter 1 [128; 144]. Its performance is competitive to to the approach by Vasicek in one dimension. Also computationally interesting is the *spacing method* by Darbellay, especially, if the conditional pdfs occupy a subspace of lower dimension (see below) [52].

The generalization of the Bayes error measurement can be understood from a predictive perspective. Suppose a multi-channel signal x has been recorded, and you wish to assign x to one of n different classes. In this situation, Bayes decision rule assigns x to the class C_i for which the probability, given the observation x, is greatest over all classes: $p(c_i|x) > p(c_j|x)$, $j = 1, .., n$, $i \neq j$ [89]. Therefore, in the multi-class setting the Bayes error measurement can be expressed in the form $E = 1 - \int max_c[p(c|x) \cdot p(x)]dx = 1 - \int max_c[p(x|c) \cdot P(c)]dx$ (see Chapter 1). The Bayes error will be zero if there is no overlap between the realizations of the different classes, and it is bound by the value $1 - \frac{1}{n}$ if all distributions are identical. The numerical estimation of the class-conditional probability density functions $p(x|c_i)$ can be done in a similar way to the one-dimensional case. I propose to use a nonparametric kernel density estimation approach after Scott, in order to prevent underestimation of the true Bayes error [219] (Section 6.3). In two or higher dimensions, it might be appropriate to rotate and rescale the data before the densities are estimated. Numerical problems can be reduced performing some sort of (local linear) embedding, provided the data are of much lower dimension [236].

If the final goal is to use the projection method for prediction, then a test sample x should be classified according to the maximum over-all class likelihoods: $x \in C_i$ if $p(x|c_i)P(c_j) >$

$p(x|c_j)P(c_j)$, $j = 1, .., n$, $i \neq j$ with $p(x|c_k)$ $(k = 1, .., n)$ the kernel density estimations. For the one-dimensional projection in the two-class problem, often a simple threshold exists which separates the two data sets in an optimal way. In the multi-class setting with a multivariate indicator function, one has to check for a combined threshold. First, the conditional probabilities $p(x|c_k)$ at a given point have to be estimated. After that, $n - 1$ comparisons have to be carried out.

According to the computational difficulties and in order to get an idea how much information will be transmitted, it might be useful to map the high-dimensional multi-class data to real values. The simplest convention is: $S(x) = i$ if $x \in C_i$ with $i = 1, .., n$. Nevertheless, one can try to incorporate prior knowledge in the indicator values as has been done, for example, for single-unit spike trains encoding different contour orientations in the visual cortex [72]. Independent of the proposed topology, the data processing inequality guarantees that a lower bound to the Shannon information will be obtained in the borderline case.

6.1.2 Multi-Class Test Statistic

Additionally to the generalization of the projection method and the distance estimation there exists another difficulty, concerning the hypothesis test. In two- or higher-dimensional projection spaces, there exist no multi-class hypothesis tests to my knowledge. Already nonparametric two-class hypothesis tests are difficult to handle in high-dimensional spaces, and there is much research in this field. For example, there are efforts to generalize the receiver operating characteristic to higher dimensions. Unfortunately, most of this work is theoretically oriented, without any algorithmic implementation [189]. Other test statistics investigate the L_1 or L_2 distances between kernel-type estimators of the conditional probability density functions [7; 9]. Some research groups conceive a difference in the distributions as the occurrence of a change point, or they count the number of nearest neighbor coincidences [77; 119; 120]. All these techniques are computational expensive.

Hypothesis tests based on the empirical characteristic functions have been proposed by Csörgö [50]. Bahr has generalized this approach in his doctoral thesis [14]. The test statistic measures the average quadratic distance of the empirical characteristic functions defined by: $C_X(x) = 1/m \sum_{k=1}^{m} exp(ix^T X_k)$ with $X_k \in R^d$, $k = 1, ..., m$. In addition, the measurement can be combined with an affine, invariant transformation, based on the pooled covariance matrix.

In the one-dimensional setting, one can apply the nonparametric test after Kruskal and Wallis for the multi-class problem. The so called H-test assesses the null hypothesis H_0: The samples belong to the same random process and the different cumulative distribution functions are identical against the alternative hypothesis H_1: at least two distributions are not identical [205, pp.395]. Therefore, the rank ordered statistic searches for general alternatives. Applied to univariate samples, received by a one-dimensional arrangement of the indicator values, it might be appropriate to formulate the alternative hypothesis in terms of a trend hypothesis. For example, the trend hypothesis after Jonckheere looks at the rank of the cumulative distributions functions F_i, and tests: $H_0 : F_1 = F_2 = ... = F_n$ against $H_1 : F_1 \leq F_2 \leq ... \leq F_n$ [205]. The usefulness of this hypothesis test depends on the correct choice of the one-dimensional indicator.

6.1.3 Single Trial Prediction

The classification of an unknown sample by the multi-class projection approach has been explained in Section 6.1.1. I said that for the Bayes risk minimization, the a posteriori probabilities of all classes $p(C_i|x)$, $(i = 1, ..., n)$ have to be estimated. This intermediate step is time consuming. For the two-class problem, I described a simple and efficient strategy to determine regions and boundaries which minimize the classification error. For this, the shape of the receiver operating characteristic computed from the training sets have to be analyzed. In addition, I argued that it might be useful to reformulate the multi-class problem in terms of repeated two-class problems. Therefore the question comes up, whether these strategies can be used for a single trial prediction as well. Under the assumption that you want to find out if an unknown sample belongs to a special class, it might be appropriate to perform the one-against-all projection. With the shape of the receiver operating characteristic obtained from the test set, it is easy to classify a new recording. The affiliation will be confirmed if the sample is assigned to the chosen class. Unfortunately, the result can be distorted, if the probability of the sample is higher than that from the combined class. But what about the *one-against-one* approach? Assuming that the signals of the different classes can be separated by a simple threshold after the projection: $x \in C_i$ if $g(x) < \theta_{ij}$ and $x \in C_j$ if $g(x) > \theta_{ij}$, an often applied heuristic is to assign the unknown sample to the class for which $g(x)$ is maximal apart from the threshold. However, the numerical effort grows by the power of two with the number of classes. Putting all these facts together it has to mention that the *one-against-one* and the *one-against-all* approaches are practical but they cannot replace the simultaneous multi-class projection in every situation.

6.2 Feature Extraction

In Chapter 3 I demonstrated that the discrimination is not only affected by the noise but also by the dimensionality. Taking the multi-dimensional relations into account makes it possible to discriminate highly noisy measurements. Restricted to finite sample sizes, I have shown that responses which are statistically independent of the stimulus set weaken the discrimination properties of all projection methods. Also, adding a component that is a function of other components is useless. Theoretically, collecting more features cannot increase the Bayes error [57, chapt.32]. However, from a practical perspective, to make the classification error as small as possible, we have to choose the features, the training sample size, and the discrimination function $g(x)$, carefully. The choice of the sample size and the projection method have been investigated in great detail. Therefore, in terms of discrimination the following question comes up: What are the most relevant and useful features for a given data set? There are many potential benefits and motivations for feature extraction. Our objective of feature extraction is to improve prediction and enhance understanding of the underlying neural computations. Besides that, feature extraction is important to provide faster and more cost-effective analyses, to reduce storage space, and to improve robustness reducing the curse of dimensionality. There are different strategies to do this. Subsets of the high-dimensional space can be analyzed, or features can be combined to build a new feature. The pre-processing step can be done in an unsupervised manner, without incorporating the classifier results, or by taking the properties and weights of the projection method into account. In the next two Sections, I will discuss solutions for the so called feature (subset) selection problem, which is a special case of the more general feature extraction problem (see also Section 5.5.3). After that, I will emphasize different pre-processing techniques, so called filter strategies, and the importance of surrogate data sets.

6.2.1 Wrapper Strategy

In the literature feature selection aims at picking out some of the dimensions (features) from the high-dimensional data [26; 105; 139]. The problem of selecting a subset of relevant features in a potentially overwhelming quantity of data can be found in many branches of science - including computer vision, gene expression, text categorization, etc. But how to select a subspace whose Bayes error is smallest? With hundreds of features it is computationally intractable to search through all combinatorial possibilities. Even with 10 features out of 100, one has to analyze more than 10^{13} combinations (see also Section 5.2).

The wrapper methodology, recently popularized by Kohavi and John, offers a simple and powerful way to address the problem of feature selection [141]. It uses the prediction performance of a given discriminant function to assess the relative usefulness of subsets of variables. In practice, one needs to define: (i) how to search the space of all possible variable subsets, (ii) how to assess the prediction performance of a learning machine, and (iii) which predictor to use.

If the dimensionality is not too large (< 10), an exhaustive search can conceivably be performed, but the problem is known to be NP-hard. Various techniques have been proposed and I will list a few of them. In the best-first or *forward selection* search, features are added one at a time from the given set of features to the initially empty set [54]. In *backward elimination*, one starts with the set of all variables and progressively eliminates the least promising ones (see also Section 5.5). Comparing these two methods, the latter is said to be stable but slower in selection [240].

Forward selection performs well with few relevant features, but it ignores correlated information in many cases. Since backward elimination starts from the whole feature space, it might capture interacting features more easily. Narendra and Fukunaga introduce an efficient method that can find the *optimal* set of features ([89, chapt.10; 176]). Their *branch-and-bound* procedure avoids searching through all subsets in many cases by making use of the monotonicity of the Bayes error measurement with respect to the partial ordering of the subsets.

Apart from deterministic search strategies, random strategies have been used for feature selection. For example, genetic algorithms were found to be very efficient, e.g., [132; 269].

All these stepwise search techniques are comparable, simple, and avoid exhaustive enumeration, but they do not guarantee the selection of the appropriate subset [57, chapt.32]. The wrapper strategy utilizes the learning machine as a black box to score subsets of variables according to their predictive power. Performance assessments are usually done using re-sampling or cross-validation procedures. Besides the classification error, several approaches use information theoretical criteria for the feature subset selection, e.g., [245]. By using the learning machine as a black box, wrappers are remarkably universal. The modular construction offers a huge flexibility, and there are strategies to accelerate the computational burden [61]. On the other hand, the optimum feature set might depend on the classifier [89, pp.440]. This means that no result exists which guarantees that the optimal feature set for a linear classifier coincides with the optimal feature set of a nonlinear classifier.

6.2.2 Embedded Strategy

In contrast to the wrapper approach, which treats feature selection as a wrapper around the projection process, the embedded approach integrates the selection within the learning (induction) algorithm. In contrast to the wrapper strategy, embedded approaches are usually specific to a given learning machine [105]. Up to now, they are mainly used with decision trees [193; 194] in the regularization network and the support vector machine framework [187]. The basic idea is to use a classifier in which the feature vector elements appear with an associated multiplicative weight. The process of feature selection amounts to finding a classifier in which only a subset of weights are non-zero.

For the linear classifier $g(x) = w^T x + b$ Bradley and Mangasarian [30] propose to solve the following minimization problem with regard to the weight vector w and the constant offset b: $\min_{w,b}(1-\lambda)\sum_i (1 - y_i(w^t x_i + b))_+ + \lambda ||w||_1$ with $y_i \in \{-1, 1\}$, $\lambda \in [0, 1)$ and $x_+ = max(x, 0)$ (see also [178]).

The L_1-based regularizer also known as the *lasso penalty term*, has the advantage that it often leads to solutions where some elements of w are exactly zero [244]. The number of zero weights depends continuously on the regularization parameter and the relevance of each dimension is coded by the absolute values of the weight vector components w. Similar to the L_2-norm regularizer - originally used with SVM - this regularizer has a single local (and global) optimum and can be solved by a linear programming approach [90; 273]. Instead of the L_1-norm, Weston et al. [264] have generalized the penalty term to the L_0-norm, and they describe an iterative algorithm for the optimization problem. In [178] the different penalty terms have been combined to exploit the good classification properties of the L_2-norm and, at the same time, the good feature selection properties of the L_1 and L_0-norm (see also [187]).

Recently, it has been shown that embedded regularized strategies can act very fast even in very high-dimensional spaces with only a few relevant components [91]. Nevertheless, extended benchmark tests point out that embedded and wrapper approaches can be further improved by so called filter strategies [105].

6.2.3 Filter Strategy

The main interest of wrapper and embedded approaches lies in identifying discriminatory features, whereas the objective of filtering is not only enhanced prediction but also model description. Filter methods try to discover the relevance of a subset of features or feature transformations by evaluating some descriptive and informative criteria. Therefore, in reference to feature selection, this strategy has its limitations, as it fails to find a feature set that would jointly maximize the discriminative criterion (see Section 1.1). Selecting the variables which achieve the best reconstruction of the data is usually sub-optimal for building a predictor, particularly if the variables are redundant [143].

Filter methods apply an unsupervised pre-processing step which is independent of the inference engine (supervised classifier). The advantage of the filter model is that it does not need to re-run the algorithm for every induction algorithm when choosing to run on a reduced feature dataset. As a consequence, the filter approach is generally computationally efficient, and it is practical for data sets with very high dimensionality. There is a vast collection of filter algorithms which can be found in the literature, including classical approaches like PCA, Fourier analysis, wavelet analysis, clustering, etc. Generally, a filter strategy consists of two steps: an unsupervised and a supervised.

Filter strategies can serve as noise and dimensionality reduction that is related to data compression. For instance if the features are highly redundant which means that the data do not span the entire (original) high-dimensional space, filter strategies can help to defy the curse of dimensionality (see also Chapter 1). Dimension reduction by an unsupervised filter technique is often used for visualization, as you can see, for example, in Section 5.4.2 where I had computed the 2-dim MDS. In many cases, filter strategies will be used for data normalization in order to enhance the comparison. The signal decomposition provided by Fourier analysis can be used to rank the separate features according to their individual discriminative power (Section 5.3.2). Such a feature ranking is also interesting in relation to information transmission. Finally, the individual evaluation of the features for ranking them may not be optimal for the discrimination, e.g. [62; 232]. In certain circumstances, the filter strategy is not adapted to take the mutual dependencies into consideration. Taking all these facts together, filter, embedded and wrapper strategies are not exclusive. It has been found that in many situations an appropriate combination leads to much better results [87]. Overall, deep knowledge and understanding of the properties of the investigated systems are extremely helpful in choosing a filter or selecting a feature.

6.2.4 Surrogate Test

The term *surrogate data* was coined by Theiler et al. [241] as a general procedure to test whether data are consistent with some class of models. Recently, surrogate data analysis has been used to test certain properties of neural spike dynamics [104], but the most common application is testing for linear models in the field of non-linear dynamics, trying to assess a functional relationship between some properties of a system to one of the system's features [216]. The essence of surrogate analysis is the construction of a (surrogate) data set derived from the original data. This is typically achieved by randomizing a feature which is under investigation, while all other features of the data are preserved. Therefore, surrogate data testing needs at least two ingredients: (i) a procedure to generate data (surrogates) from the class of models to be tested which are consistent with this null hypothesis, and (ii) a statistic to perform the testing. If the statistic computed from the original data is significantly different from the distribution of surrogate statistics then the null hypothesis will be rejected, indicating that the difference is related to the feature which is absent in the surrogates. If the hypothesis is tested positive (no rejection), it might be that the generated surrogates do not fulfill the claims, and the observed results from the original data should be taken with care. For the computation of the significance level various classical procedures can be used.

With respect to multi-channel data, I developed surrogate data to answer the following question: Given the cortical signals and the RBF classifier, is it possible to obtain the same segregation neglecting the class labels? Answering this question is important for two reasons. First, without this test, a skeptical researcher might argue that the observed differences in Chapter 5 can be explained by a stimulus and reaction independent stochastic process, or due to the low number of samples embedded in a high-dimensional space. Second, if the observed effects in the original data arrangement are not significantly different, then there is no reason to believe, e.g., that the monkey's reaction has anything to do with the stimulus presentation or its reaction. In doing so, I support the results in Chapter 5 in two ways. 1.) I quantified the segregation with distinct statistical measurements and proved that the significance of these effects has something to do with the underlying physiological processes. 2.) Compared to the surrogate data, I can exclude that the spatio-temporal effects are an artifact of the projection process or something else.

To finish this Section, I have to annotate that the proper choice of the surrogate data in more elaborated problems is very complex. Moreover, the randomization might produce signals which are physiologically unrealistic. In contrast to the feature selection or filter strategies, surrogate data analysis is more related to hypothesis testing. In this context surrogate data supplement the feature extraction process.

6.3 Projection Approach

One of the results from Chapter 3 was that no method achieves the lowest classification error in every benchmark test. It might be that the simplest method reaches the best performance. This fact is remarkable, because some of the more elaborated methods can be understood as a generalization of a simpler approach. The reason for such a contradictory behavior can be easily explained. In cases where the performance of a simpler method outperforms the generalized method some sort of overfitting occurs. In these cases the complexity of the more elaborated discriminant function, expressed for example by the number of free parameters, exceeds the data complexity.

Taking this result into consideration, the question is which complexity have the cortical multi-channel recordings and which method is *suitable* for the analysis of these recordings. From the study of the cortical-like signals in Chapter 4 I proposed that radial basis functions with a multiquadric kernel should be preferred for amplitude-continuous multi-channel recordings.

This recommendation is confirmed by the analysis in Kremper and Eckhorn [151] in which the performance of the six projection methods was investigated on multi-channel LFP and MUA recordings (see also [149] as well as Section 5.1 and 5.2). In most cases, the RBF estimation of the transmitted information outperforms all other methods significantly ($p < 0.01$). Comparing the various projection methods, Kremper and Eckhorn found that after the RBF projection, the loss of information was lowest for both signal types.

Although, the other five projection methods show a lower performance the temporal variation of the discrimination is similar for all methods. Therefore, it can be concluded that the reported differences are not a side effect of the RBF method what supports the correctness of the results in Chapter 5. In addition, the different performance of the six projection methods, give us further insight into the underlying processes, since each projection method is based on another model assumption. Taking this into account, the various projection approaches can be regarded as some sort of hypothesis test. For example, in [151], RDA outperforms LCC and LFD, which indicates that the multivariate distributions must be different at least in their second order moments (data not shown).

One might ask: What are the reasons for the favourable performance of radial basis functions? Three points might be responsible. First, RBF is a non-linear method, and this guarantees a more flexible adjustment to the cortical multi-channel recordings. This viewpoint is supported by the MDS depictions in Section 5.4.2. Second, the multiquadric kernel enables a better generalization than, e.g., Gaussian kernels. Its simple form enables an automatic tuning to the properties of the underlying distributions. A radial basis function approximation in which the influence of the kernel increases with increasing distance is more related to a local linear discrimination than a radial basis function approximation with local operating kernels. In addition, a continuous discrimination function is often sufficient in high-dimensional spaces. Haasdonk and Bahlmann [106] stated in a recent publication that distance-based kernels are favourable in contrast to standard kernels with various data sets, including neural cortical recordings. Third, in contrast to the least square support vector machine approach, in which a multiple of the identity matrix is used to regularize the within class variance, RBF uses a regularization which is data dependent, in analogy to SVM (see also Section 2.7), but it is not so time consuming.

The methods which I have investigated represent different model assumptions. Besides these projection models, many other projection approaches exist. Particularly, two statistical pattern recognition approaches became very popular in the last eight years. The first group of methods includes so called support vector machines (SVM) [251; 252]. In an analogous manner to radial basis functions, support vector machines can be deduced from regularization theory that offers a general framework for many other state-of-the-art pattern recognition methods [42; 99]. In this context, radial basis functions can be understood as a quadratic optimization problem, and the solution can be obtained by solving a linear system of equations. The SVM minimization problem has a unique solution, too. The weight vector for SVM can be obtained by convex quadratic programming. In the non-linear case, SVM maps the pattern vectors to a high-dimensional (not necessarily finite dimensional) feature space where a separating hyperplane is constructed. One key element of SVM is Mercer's theorem, which I have also used with KFD (see Chapter 2). Recent improvements by Vapnik have concentrated on transductive empirical inference taking the properties of the unlabeled test set into account [96; 105]. I refer the interested reader to the vast literature, since an extensive description lies beyond this scope, e.g., [214]. In comparison to RBF, SVM seems to be superior in many situations [215]. On the other hand, SVM seems to be more sensitive to noise and outliers than RBF. This can be understood by facing the different margin maximization strategies of SVM and RBF. A geometrical interpretation of SVM is to maximize the minimal margin in the two-class problem. As I have mentioned in Chapter 2, RBF is related to least square SVM in which the average margin will be maximized. Furthermore, in contrast to RBF, SVM is very time consuming, and there is no guarantee for better performance.

The second, even more successful group of pattern recognition methods consists of so called ensemble techniques or combined predictors [59; 60]. All current machine learning models have some constraints or local minima in certain application domains given limited training data. Ensemble methods try to overcome these detriments. Instead of developing a highly specialized classifier, many ensemble techniques try to combine so called weak classifiers which are easy to implement and very fast. There are different approaches to build an ensemble. An ensemble can be formed by multiple architectures, same architecture trained with different algorithms, or even different classifiers. The ensemble can be developed by training with equal or different data. The simplest approach to combine multiple classifiers is by averaging or majority voting of outputs from individual classifiers. Averaging the predictions helps to reduce the variance and often increases the reliability of the predictions. It has been shown theoretically that the performance of the ensemble can not be worse than any single model used separately if the predictions of individual classifiers are unbiased and uncorrelated [188]. In practice, ensemble methods have been shown to perform better than any of the base classifiers in almost all cases [17; 227]. Within this group of pattern recognition methods, two outstanding strategies, *Adaboost* and *Random Forests*, have been developed in recent years. *AdaBoost* creates new and improved weak classifiers by iteratively manipulating the training data set [82]. In doing so, Adaboost and its boosting variants put higher weights on data points which were incorrectly classified in a previous iteration step [211]. Adaboost is related to SVM and LS-SVM in that it performs some sort of margin maximization. *Random forests* was designed for use with tree predictor combination [32]. Random forests is related to bagging, which stands for Bootstrap Aggregation and combining, and random subspace methods [31]. In each iteration, the training examples are bootstrapped, to generate a different training set for the base classifier. One advantage of random forest is that it can handle missing data. Beyond that, it is insensitive to noise. Although much effort has been devoted to combining methods, several issues remain or have not been completely solved. These include the choice of individual classifiers included in the ensemble, the size of the ensemble, and the general optimal way to combine individual classifiers [271]. With respect to our RBF approach, as Freund and Schapire pointed out [82]: *'Naturally, one needs to consider whether the improvement in error is worth the additional computation time.'* This aspect is very important because in general a significant improvement by another method is not guaranteed.

Besides the application of other projection methods, it is also possible to generalize the radial basis function approach itself. In the following I will list a few strategies. For example, the choice and the position of the centres can be modified. Leonardis and Bischof proposed to use the minimum description length principle to adapt the complexity of RBF networks [157]. Other groups used SVM to estimate reasonable prototypes for the radial basis function [217]. Sometimes the number of training samples exceeded the number of centres. In these cases the weight vector can be computed by regularization in combination with its pseudo-inverse [182]. The shape parameter of the basis function may be optimized by information maximization [246]. Instead of the L2-distance, other data dependent metrics may be used, which is also interesting in terms of feature selection [99]. Morlini [174] developed a strategy to incorporate unlabeled or partially classified data in the RBF learning process. Orr et al. investigated the combination of decision trees with RBF [182] and Espinosa et al. studied various methods in combination with multilayer networks [121]. Overall, it is not clear if these modifications will lead to an improvement in information estimation or if these methods will require additional tests. According to my explanations, I propose to start the Bayes error or Shannon information estimation with methods like LCC or RBF.

6.4 Multi-Channel Population Data

In Chapter 5, I demonstrated the usefulness of the projection approach to multi-channel local field potentials and multi-unit activity. Furthermore, I applied my projection approach to cortical signals from three other experiments. In the first experiment, I analyzed local field potentials (LFP) recorded in V1 with five parallel μ-electrodes from an awake monkey during visual stimulation by (1.) a continuous sinusoidal grating compared to (2.) a grating in which an object was defined by a shifted rectangle [92]. The task of the monkey was to fixate throughout the visual stimulation. I quantified the segregation by the area under the ROC using the recordings of all electrodes simultaneously [146]. Finally, I estimated the contributions from single electrode recordings with different combinations of four out of five electrode signals. In accordance with the results from Chapter 5, I found that the simultaneous consideration of all parallel recordings produces the highest discrimination, though the influence of single recordings was significantly different. The distinction between the global and the local analysis has been further confirmed when I compared the results due to the simultaneous inclusion of all channels with that of the mean over the single-channel analyses. I also tested the reliability of each recording site separately and compared these results with those of standard approaches. In all cases, the multi-channel approach confirms classical findings but encloses further effects.

In the second experiment, I analyzed local field potentials recorded with one electrode in V1 from the behaving monkey. The monkey had to solve a simple matching-to-sample paradigm: the discrimination between two stimuli by moving a lever. Stimulus 1 consisted of randomly arranged Gabor patches. Stimulus 2 differed from stimulus 1 in that it contained a chain of oriented Gabor patches surrounded by Gabor patches incidentally positioned and oriented. The stimuli were prepared such that in the classical receptive field in both cases the same local element were positioned. In order to exclude effects evoked by different response latencies the trials were sorted such that both data sets had the same response latency distribution. Compared to the findings by Heinze [115, pp.48], I took the temporal correlations into account and found an even higher (significant) difference, about 200 ms after stimulus-onset.

In a comparative study, I tested three signal types generated by a multi-layer neural network, performing some sort of figure-ground segregation. The network consisted of coupled integrate-and-fire neurons with internal white noise [147]. In addition to artificial multi-channel LFP and MUA signals, I applied the RBF and ADC method to the raw membrane potentials recorded from multiple sites (sampling rate 1 kHz). Discrimination by RBF was higher than by ADC for all signal types. Comparing the three signal types with respect to their stimulus dependent information, LFP contains the most information. Interestingly, discrimination by the membrane potential was even higher than by MUA, indicating some subthreshold effects for this model.

Putting all these results in a nutshell, I am convinced that the projection approach is well suited to study multi-channel intra-cortical recordings. According to the studies of Kestler et al. [136], Natschläger and Maass [177], Paninski [183], as well as Eichhorn et al. [72], amongst others, the RBF method might be successfully applied to any other types of neurophsiological multi-channel data, including EEG, MEG, and envelope conversions of multiple- and single-unit spike trains.

6.5 Outlook

Finally, let me give a short look into the future at further work and improvements. For example, serious effects in combination with the misclassification error have to be explored in more detail. How to handle imbalanced data sets or missing values could be also important. Another point which has to be investigated concerns the data normalization and the feature extraction. Feature extraction and filter strategies will be even more important with increasing number of recording sites, e.g., 50 or 100 electrodes. Most of the feature extraction approaches in Chapter 5 are heuristically motivated and have to be combined in a stringent fashion. Furthermore, the alignment of the signals has to be reconsidered. At the moment, I have concentrated on stimulus-onset and response-triggered data sets. If the processing operates on different time scales, significant differences might be detected by the projection approach, though the main differences rely on differences in the time scale. For the investigation of spatial extended processes other non-linear correlation measures should be studied, e.g., those based on information theory. In this context, it would be interesting to determine if there are signal components which will be used in common. My work was mainly concentrated in discrimination, but not on classification. Therefore, RBF might be extended to online learning and monitoring.

A Appendix

A.1

$$U''(t) + (a+b)U'(t) + abU(t) = abV(t-\delta) + \xi(t), \quad (U(0) = 0)$$

$$\theta'(t) = \frac{1}{\gamma}(\theta_o - \theta(t)) + \frac{c}{\gamma}S(U(t) - \theta(t)), \quad (\theta(0) = 0)$$

$$S(x) = \frac{1}{1 + e^{-\alpha x}}$$

Parameter setting:

a $[1/ms]$	b $[1/ms]$	c $[ms]$	α $[mV]$	γ $[ms]$	δ $[mS]$	θ_o $[mV]$
1/16.8	1/4.2	180	1	50	4	4

$$V(t) = \begin{cases} 80\,mV & 32 < t < 33;\ 52 < t < 53;\ 108 < t < 109\,ms \\ 0\,mV & else \end{cases}$$

A.2

$$U_i''(t) + (a+b)U_i'(t) + abU_i(t) = abV(t) + \sum_{j=1}^{2} w_{ij}(t)S(U_j(t-\delta) - \theta_j(t-\delta)) + \xi_i(t)$$

$$\theta_i'(t) = \frac{1}{\gamma}(\theta_o - \theta_i(t)) + \frac{c}{\gamma}S(U_i(t) - \theta_i(t)), \quad i = 1, 2$$

$$S(x) = \frac{1}{1 + e^{-\alpha x}}$$

Parameter setting for **Class 1**:

a $[1/ms]$	b $[1/ms]$	c $[ms]$	α $[mV]$	γ $[ms]$	δ $[mS]$	θ_o $[mV]$
1/4.8	1/4.6	180	1	50	2	0.1

$$V(t) = \begin{cases} 16\,mV & 0 < t < 100\,ms \\ 0\,mV & else \end{cases}$$

$$\mathbf{W} = \begin{bmatrix} w_{11} = 0 & w_{12} = 25 \\ w_{21} = -25 & w_{22} = 0 \end{bmatrix}$$

Parameter setting for **Class 2** (same as before except for):

$$V(t) = \begin{cases} 16.8\,mV & 0 < t < 100\,ms \\ 0\,mV & else \end{cases}$$

A.3

Equations (see Appendix A.2).

Parameter setting for **Class 1**:

a $[1/ms]$	b $[1/ms]$	c $[ms]$	α $[mV]$	γ $[ms]$	δ $[mS]$	θ_o $[mV]$
$1/4.8$	$1/4.6$	180	1	50	2	0.1

$$V(t) = \begin{cases} 10\,mV & 0 < t < 100\,ms \\ 0\,mV & else \end{cases}$$

$$\mathbf{W} = \begin{bmatrix} w_{11} = 0 & w_{12} = 28 \\ w_{21} = -28 & w_{22} = 0 \end{bmatrix}$$

Parameter setting for **Class 2** (same as before except for):

$$\delta = 3ms$$

$$\mathbf{W} = \begin{bmatrix} w_{11} = 0 & w_{12} = 32 \\ w_{21} = 32 & w_{22} = -32 \end{bmatrix}$$

A.4

Stability analysis of a single neuron with local delayed feedback according to [108, chap.11]:

$$U''(t) + (a+b)\,U'(t) + a\,b\,(U(t) - U_o) = a\,b\,w \cdot S(U(t-\delta) - \theta(t-\delta))$$

$$\theta'(t) + \frac{1}{\gamma}(\theta(t) - \theta_o) = c \cdot S(U(t) - \theta(t)))$$

Equilibrium:

$$U^* = U_o + w \cdot S\Big(U^* - \frac{c}{w}(U^* - U_o) - \theta_o\Big)$$

$$\theta^* = \theta_o + c \cdot S\Big(U_o - \frac{w}{c}\theta_o + (\frac{w}{c} - 1)\theta^*\Big)$$

Characteristic equation:

$$e^{\lambda\,t}\begin{pmatrix} \lambda +\ a & -1 & 0 \\ -a\,b\,w \cdot S'\,exp(-\lambda\,\delta) & \lambda +\ b & a\,b\,w \cdot S'\,exp(-\lambda\,\delta) \\ -\frac{c \cdot S'}{\gamma} & 0 & \frac{1+\lambda\gamma+c \cdot S'}{\gamma} \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = e^{\lambda\,t}\ A\,c\ =\ 0$$

$$S' = \alpha \cdot (1 - S(U^* - \theta^*))\ S(U^* - \theta^*)$$

Determinant:

$$det(A) = (\lambda + a)(\lambda + b)(\lambda + \frac{1 + c \cdot S'}{\gamma}) - (\lambda + \frac{1}{\gamma})\ a\,b\,w \cdot S' e^{-\lambda\delta} = 0\ ,\ \lambda = u + i\,v$$

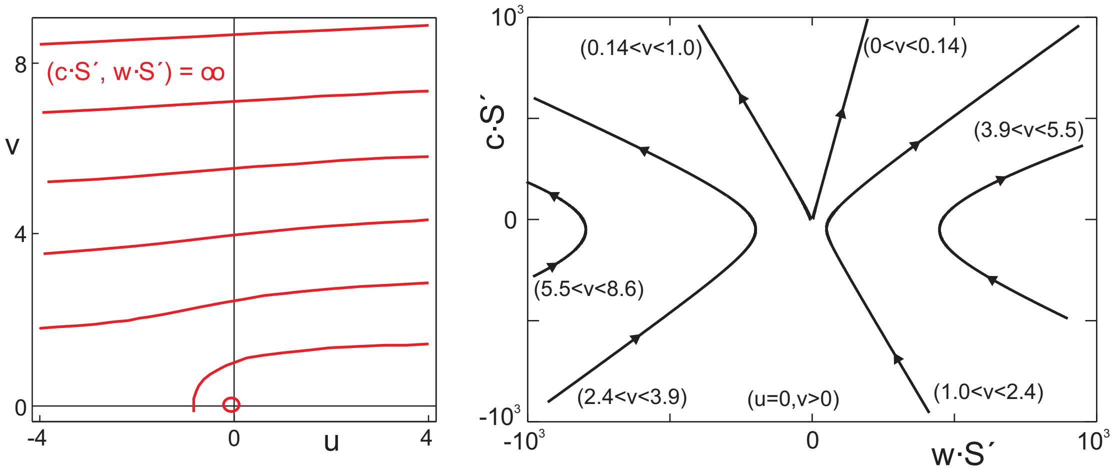

Figure A.1: Singularities and stability boundaries of $c \cdot S'$ over $w \cdot S'$. Parameter setting: $a = 1/4.8\ ms^{-1}$, $b = 1/4.6\ ms^{-1}$, $\gamma = 50\ ms$, $\delta = 2\ ms$.

Bibliography

[1] Abramson, I. (1982) On bandwidth variation in kernel estimates – a square root law. The Annals of Statistics, 10(4):1217–1223. 18

[2] Adini, Y., Sagi, D. and Tsodyks, M. (1997) Excitatory-inhibitory network in the visual cortex: Psychophysical evidence. Proc. Natl. Acad. Sci. USA, 94:10426–10431. 58

[3] Adrian, E.D. The basis of sensation; the action of the sense organs. Christophers: London, 1928. 51

[4] Ahrens, J.H. and Dieter, U. (1985) Sequential random sampling. ACM Transactions on Mathematical Software, 11(2):157–169. 12

[5] Akaike, H. (1954) An approximation to the density function. Ann. Inst. Statist. Math., 6:127–132. 15

[6] Allen, D.M. (1974) The relationship between variable selection and data augmentation and a method of prediction. Technometrics, 16:125–127. 12

[7] Allen, D.L. (1997) Hypothesis testing using an L_1-distance bootstrap. The American Statistican, 51(2):145–150. 101

[8] Amari, S.I. (1977) Dynamics of pattern formation in lateral-inhibition type neural fields. Biological Cybernetics, 27:77–87. 50

[9] Anderson, N.H., Hall, P. and Titterington, D.M. (1994) Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. Journal of Multivariate Analysis, 50(1):41–54. 101

[10] Anderson, J.S., Lampl, I., Gillespie, D.C. and Ferster, D. (2000) The contribution of noise to contrast invariance of orientation tuning in cat visual cortex. Science, 290(5498):1968–1972. 53

[11] Antos, A., Devroye, L. and Györfi, L. (1999) Lower bounds for Bayes error estimates. IEEE Trans. on Pattern Analysis and Machine Intelligence, 21(7):643–645. 10

[12] Arbib, M.A. The handbook of brain theory and neural networks (2nd ed.). The MIT Press, Cambridge, Ma, 2002. 49

[13] Arrenberg, J. (1996) A note on the nonparametric test based on the L_1-version of the Cramér-von Mises statistic. Statistical Papers, 37:95–104. 24

[14] Bahr, R. (1996) Ein neuer Test für das mehrdimensionale Zwei-Stichproben-Problem bei allgemeiner Alternative. Dissertationsschrift, Universität Hannover, 1996. 101

[15] Barrett, R., Berry, M., Chan, T.F., Demmel, J., Donato, J.M., Dongarra, J., Eijkhout, V., Pozo, R., Romine, C. and Vand der Vorst, H. Templates for the solution of linear systems: building blocks for iterative methods. SIAM, Philadelphia, 1994. 34

[16] Baudat, G. and Anouar, F. (2000) Generalized discriminant analysis using a kernel approach. Neural Computation, 12(10):2385–2404. 30

[17] Bauer, E., and Kohavi, R. (1999) An empirical comparison of voting classification algorithms: bagging, boosting and variants. Machine Learning, 36(1/2):105–139. 108

[18] Beatson, R.K., Cherrie, J.B. and Mouat, C.T. (1999) Fast fitting of radial basis functions: methods based on preconditioned GMRES iteration. Advances in Computational Mathematics, 11:253–270. 34

[19] Beirlant, J., Dudewicz, E.J., Györfi, L. and van der Meulen, E.C. (1997) Nonparametric entropy estimation: an overview. International Journal of the Mathematical Statistics Sciences, 6:17-39. 19

[20] Belkin, M. and Niyogi, P. (2003) Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation, 15(6):1373–1396. 9, 10

[21] Bellman, R. Adaptive Control Processes: A Guided Tour. Princeton University Press, 1961. 7

[22] Berlinet, A. and Devroye, L. (1994) A comparison of kernel density estimates. Publications de l'Institut de Statistique de l'Université de Paris, 38(3):3–59. 15, 16

[23] Bhattacharyya, A. (1943) On a measure of divergence between two statistical populations defined by their probability distributions. Bulletin of the Calcutta Mathematical Society, 35:99–109. 4

[24] Bialek, W., Rieke, F., van Steveninck, R.R. and Warland, D. (1991) Reading a neural code. Science, 252:1854–1857. 5

[25] Biem, A., Katagiri, S. and Juang, B.-H. (1997) Pattern recognition using discriminative feature extraction. IEEE Trans. on Signal Processing, 45(2):500–504 . 10

[26] Blum, A.L., and Langley, P. (1997). Selection of relevant features and examples in machine learning. Artificial Intelligence, 97(1-2):245-271. 103

[27] Borg, I. and Groenen, P. Modern multidimensional scaling: Theory and applications. Springer-Verlag, New York, 1997. 10, 91

[28] Borg-Graham, L.J., Monier, C. and Frégnac, Y. (1998) Visual input evokes transient and strong shunting inhibition in visual cortical neurons. Nature, 393:369–373. 53

[29] Borst, A. and Theunissen, F.E. (1999) Information theory and neural coding. Nature Neuroscience, 2(11):947–957. 5

[30] Bradley, P.S. and Mangasarian, O.L. (1998) Feature selection via concave minimization and support vector machines. In Machine Learning Proceedings of the Fifteenth International Conference (ICML '98), J. Shavlik, Ed. Morgan Kaufmann, San Francisco, California, pp.82–90. 104

[31] Breiman, L. (1996) Bagging predictors. Machine Learning, 24(2):123-140. 108

[32] Breiman, L. (2001) Random forests. Machine Learning, 45(1):5-32. 108

[33] Broomhead, D.S. and Lowe, D. (1988) Multivariable functional interpolation and adaptive networks. Complex Systems, 2:321–355. 37

[34] Buhmann, M. Radial basis functions: theory and implementations. Cambridge University Press, 2003. 33

[35] Carreira-Perpinan, M.A. (1997) A review of dimension reduction techniques. Technical report CS-96-09, Dept. of Computer Science, University of Sheffield. 8, 10

[36] Celebrini, S. and Newsome, W.T. (1994) Neuronal and psychophysical sensitivity to motion signals in extrastriate area MST of the macaque monkey. Journal of Neuroscience, 14:4109-4124. 7

[37] Cencov, N.N. (1962) Evaluation of an unknown distribution density from observations. Soviet Mathematics Dokl., 3:1559–1562. 14

[38] Chacron, M.R., Longtin, A. and Maler, L. (2003) The effects of spontaneous activity, background noise and the stimulus ensemble on information transfer in neurons. Network: Comput. Neural Syst., 14:803–824. 53

[39] Chakrabarti, S., Roy, S. and Soundalgekar, M.V. (2003) Fast and accurate text classification via multiple linear discriminant projections. The International Journal on Very Large Data Bases, 12(2):170-185. 28

[40] Chao, R., Cuevas, A. and González-Manteiga, A. (1994) A comparative study of several smoothing methods in density estimation. Computational Statistics and Data Analysis, 17:153–176. 16

[41] Chen, L.-F., Liao, H.-J.M., Ko, M.-T., Lin, J.-C. and Yu, G.-J. (2000) A new LDA-based face recognition system which can solve the small sample size problem. Pattern Recognition, 33:1713–1726. 27

[42] Chen, Z. and Haykin, S. (2002) On different facets of regularization theory. Neural Computation, 14(12):2791-2846. 107

[43] Chernoff, H. (1952) A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations. Annals of Mathematical Statistics, 23:493–507. 4

[44] Cooke, T. (2002) Two variations on Fisher's linear discriminant for pattern recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 24(2):268–273. 28

[45] Cortes, C. and Vapnik, V. (1995) Support vector networks. Machine learning, 20:273–297. 31

[46] Cover, T.M. and Hart, P.E. (1967) Nearest neighbor pattern classification. IEEE Trans. on Information Theory, 13(1):21–27. 4, 32

[47] Cover, T.M. and Thomas, J.A. Elements of Information Theory. Wiley Series in Telecommunications. John Wiley & Sons, New York, 1991. 5, 6, 10, 11, 19, 20, 100

[48] Cowen, L.J. and Priebe, C.E. (1997) Randomized nonlinear projections uncover high-dimensional structure. Advances in Applied Mathematics, 19(3):319–331. 32

[49] Cox, T.F. and Cox, M.A.A. Multidimensional sacling. Chapman and Hall, second edition, 2001. 10

[50] Csörgö, S. (1986) Testing for normality in arbitrary dimension. Annals of Statistics, 14(2):708–723. 101

[51] Darbellay, G.A. (1998) Predictability: An information-theoretic perspective. In Signal Analysis and Prediction, A. Procházka, J. Uhlir, P.J.W. Rayner and N.G. Kingsbury (eds), Birkhäuser, Boston, pp.249–262. 5

[52] Darbellay, G.A. and Vajda, I. (1999) Estimation of the information by an adaptive partitioning of the observation space. IEEE Trans. on Information Theory, 45(4):1315–1321. 100

[53] Destexhe, A. and Paré, D. (1999) Impact of network activity on the integrative properties of neocortical pyramidal neurons in vivo. J. Neurophysiol., 81(4):1531-1547. 53

[54] Devijver, P.A. and Kittler, J. Pattern recognition: A Statistical Approach. Prentice-Hall International, Englewood Cliffs, NJ., 1982. 103

[55] Devroye, L. and Györfi, L. Nonparametric density estimation. The L1 view. John Wiley and Sons, New York, 1985. 14, 15

[56] Devroye, L. (1996) Random variate generation in one line of code. In: 1996 Winter Simulation Conference Proceedings, (edited by J.M. Charnes, D.J. Morrice, D.T. Brunner and J.J. Swain), ACM, San Diego, pp.265–272. 18, 20

[57] Devroye, L., Györfi, L. and Lugosi, G. A Probabilistic Theory of Pattern Recognition. Springer-Verlag, New York, 1996. 2, 4, 10, 102, 103

[58] Devroye, L. (1997) Universal smoothing factor selection in density estimation: theory and practice (with comments). Test, 6(2):223–320. 16, 37

[59] Dietterich, T.G. (2000) Ensemble methods in machine learning. In J. Kittler and F. Roli (Ed.) First International Workshop on Multiple Classifier Systems. Lecture Notes in Computer Science. Springer Verlag, New York, pp.1–15. 107

[60] Dietterich, T.G. (2002) Ensemble Learning. In The handbook of brain theory and neural networks. 2nd edition, M.A. Arbib, Ed., The MIT Press, Cambridge, Ma., pp.405–408. 107

[61] Dong, M. and Kothari, R. (2003) Feature subset selection using a new definition of classifiability. Pattern Recognition Letters, 24(9-10):1215-1225. 9, 103

[62] Duch, W., Wieczorek, T., Biesiada, J. and Blachnik, M. (2004) Comparison of feature ranking methods based on information entropy. Proc. of International Joint Conference on Neural Networks, IEEE Press, Budapest, pp.1415–1420. 105

[63] Duda, R.O., Hart, P.E. and Stork, D.G. Pattern Classification. John Wiley & Sons, second edition, New York, 2000. 2, 10, 26, 27, 30

[64] Duin, P.W.R. (2000) Classifiers in almost empty spaces. In Proc. 15th Int. Conference on Pattern Recognition, Barcelona, A. Sanfeliu, J.J. Villanueva, M. Vanrell, R. Alquezar, A.K. Jain, J. Kittler (eds.), IEEE Computer Society Press, Los Alamitos, vol.2, pp.1–7. 27, 39

[65] Eckhorn, R. and Pöpel, B. (1975) Rigorous and extended application of information theory to the afferent visual system of the cat. II. Experimental results. Biological Cybernetics, 17(1):71–77. 5

[66] Eckhorn, R., Grüsser, O.-J., Kröller, J., Pellnitz, K. and Pöpel, P. (1976) Efficiency of different neural codes: information transfer calculations for three different neuronal systems. Biological Cybernetics, 22(1):49–60. 5

[67] Eckhorn, R. and Thomas, U. (1993) A new method for the insertion of multiple microprobes into neural and muscular tissue, including fiber electrodes, fine wires, needles and microsensors. Journal of Neuroscience Methods, 49(3):175–179. 71

[68] Efromovich, S. (1996) Adaptive orthogonal series density estimation for small samples. Computational Statistics & Data Analysis, 22(6):599–617. 14

[69] Egan, J.P. Signal Detection Theory and ROC Analysis. Series in Cognition and Perception. Academic Press, New York, 1975. 23

[70] Eger, M. and Eckhorn, R. (2002) Quantification of sensory information transmission using timeseries decorrelation techniques. BioSystems, 67(1-3):55–65. 9

[71] Eggermont, P.B. and LaRiccia, V.N. (1999) Best asymptotic normality of the kernel density entropy estimator for smooth densities. IEEE Trans. on Information Theory, 45(4):1321–1326. 20

[72] Eichhorn, J., Tolias, A.S., Zien, A., Kuss, M., Rasmussen, C.E., Weston, J., Logothetis, N.K. and Schölkopf, B. (2004) Prediction on spike data using kernel algorithms. Advances in Neural Information Processing Systems, 16:1367–1374. 101, 109

[73] Elston, G.N. (2003) Cortex, cognition, and the cell: new insights into the pyramidal neuron and prefrontal function. Cerebral Cortex, 13(11):1124–1138. 49

[74] Epanechnikov, V.A. (1969) Nonparametric estimation of a multivariate probability density. Theory of Probability and its Applications, 14:153–158. 15

[75] Erdogmus, D. and Principe, J.C. (2001) Information transfer through classifiers and its relation to probability of error. In Proceedings of the International Joint Conference on Neural Networks, Washington D.C., vol.1, pp.50–54. 11

[76] Erdogmus, D. and Principe, J.C. (2004) Lower and upper bounds for misclassification probability based on Renyis information. The Journal of VLSI Signal Processing, 37(2-3):305–317. 6

[77] Ferger, D. (2000) Optimal tests for the general two-sample problem. Journal of Multivariate Analysis, 74(1):1–35. 101

[78] Fischer, G.W., Carmon, Z., Ariel, D., Zauberman, G. and L'Ecuyer, P. (1999) Good parameters and implementations for combined multiple recursive random number generators. Operations Research, 47(1):159–164. 12, 16, 37

[79] Fisher, R.A. (1936) The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7:179–188. 27

[80] Fix, E. and Hodges, J.L.Jr. (1951) Discriminatory analysis, nonparametric discrimination, consistency properties. Report no. 4, project no.21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas. 13

[81] Fralick, S. and Scott, R.W. (1971) Nonparametric Bayes-risk estimation. IEEE Trans. on Information Theory, 17(4):440–444. 13

[82] Freund, Y. and Schapire, R.E. (1996). Experiments with a new boosting algorithm. In Proc. Thirteenth International Conference on Machine Learning. Morgan Kauffman, San Francisco, pp.148–156. 108

[83] Friedman, J., Baskett, F. and Shustek, L. (1975) An algorithm for finding nearest neighbor. IEEE Trans. on Computers, 24:1000–1006. 32

[84] Friedman, J. (1989) Regularized discriminant analysis. Journal of the American Statitical Association, 84(405):165–175. 28, 29

[85] Friedman, J. (1994) Flexible metric nearest neighbor classification. Technical Report 113, Dept. of Statistics, Stanford University. 32

[86] Fodor, I.K. (2002) A survey of dimension reduction techniques. LLNL, Technical report, UCRL-ID-148494. 10

[87] Fürnkranz, J. (2002) Round robin classification. Journal of Machine Learning Research, 2(4):721–747. 105

[88] Fukunaga, K. and Hummels, D.M. (1987) Bayes error estimation using Parzen and k-NN procedures. IEEE Trans. on Pattern Analysis and Machine Intelligence, 9(5):634–643. 13, 33

[89] Fukunaga, K. Introduction to Statistical Pattern Recognition. Second edition, Academic Press, New York, 1990. 1, 2, 7, 13, 26, 100, 103, 104

[90] Fung, G. and Mangasarian, O.L. (2002) A feature selection Newton method for support vector machine classification. Data Mining Institute, Computer Sciences Department, University of Wisconsin, no.2-3. 104

[91] Fung, G. and Mangasarian, O.L. (2003) The disputed federalist papers: SVM Feature selection via concave minimization. Proceedings of the Conference of Diversity in Computing. Atlanta, Georgia, USA, pp.42–46. 104

[92] Gail, A., Brinksmeyer, H.J., Eckhorn, R. (2000) Contour decouples gamma activity across texture representation in monkey striate cortex. Cerebral Cortex, 10(9):840-850. 109

[93] Gail, A., Brinksmeyer, H.J., Eckhorn, R. (2003) Simultaneous mapping of binocular and moncular receptive fields in awake monkeys for calibrating eye alignment in a dichoptical setup. Journal of Neuroscience Methods, 126(1):41–56. 71

[94] Gail, A., Brinksmeyer, H.J., and Eckhorn, R. (2004) Perception-related modulations of local field potential power and coherence in primary visual cortex of awake monkey during binocular rivalry. Cerebral Cortex, 14(3):300–313. 73, 79, 85, 94, 95, 96

[95] Gallager, R.G. Information theory and reliable communication. John Wiley & Sons, New York, 1968. 5

[96] Gammerman, A., Vovk, V. and Vapnik, V. (1996) Transduction in nonparametric pattern recognition. Technical report, CSD-TR-96-21, Department of Computer Science, Royal Holloway, University of London. 107

[97] Gentle, J.E. Random number generation and Monte Carlo methods. Springer-Verlag, New York, 1998. 16, 18, 20

[98] Gerstner, W. and Kistler, W.M. Spiking neuron models. Cambridge University Press, 2002. 49, 51, 53

[99] Girosi, F., Jones, M. and Poggio, T. (1995) Regularization theory and neural networks architectures. Neural Computation, 7(2):219–269. 107, 108

[100] Glad, I.K., Hjort, N.L. and Ushakov, N.G. (2003) Correction of density estimators which are not densities. The Scandinavian Journal of Statistics, 30(2):415–427. 13, 14

[101] Goria, M.N., Leonenko, N.N., Mergel, V.V., Inverardi, P.L.N. (2005) A new class of random vector entropy estimators and its applications in testing statistical hypotheses. Journal of Nonparametric Statistics, 17(3):277–297. 20

[102] Goutte, C. (1997) Note on free lunches and cross-validation. Neural Computation, 9(6):1245–1249. 12

[103] Green, D.M. and Swets, J.A. Signal Detection Theory and Psychophysics. Wiley: New York, 1966. 23

[104] Grün, S., Diesmann, M. and Aertsen, A. (2002) Unitary events in multiple single-neuron spiking activity: I. detection and significance. Neural Computation, 14(1):43–80. 105

[105] Guyon, I. and Elisseeff, A. (2003) An introduction to variable and feature selection. Journal of Machine Learning Research, 3(7-8):1157–1182. 103, 104, 107

[106] Haasdonk, B., Bahlmann, C. (2004) Learning with distance substitution kernels. Pattern Recognition - Proc. of the 26th DAGM Symposium, Tübingen, Springer, Berlin, 2004. 107

[107] Härdle, W. Applied Nonparametric Regression. Econometric Society Monographs No. 19, Cambridge University Press, 1990. 12

[108] Hairer, E., Noersett, S.P. and Wanner, G. Solving ordinary differential equations I. Nonstiff Problems. Springer-Verlag, New York, 2.ed., 1993. 54, 113

[109] Hairer, E., and Wanner, G. Solving ordinary differential equations II. Stiff and differential-algebraic problems. Springer-Verlag, New York, 2.ed., 1996. 56

[110] Hall, P. and Wand, M.P. (1988) Minimizing L1 distance in nonparametric density estimation. Journal of Multivariate Analysis, 26:59–88. 15

[111] Hall, P. and Morton, S.C. (1993) On the estimation of entropy. Ann. Inst. Stat. Math., 45(1):69–88. 20

[112] Hastie, T. and Tibshirani, R. (1996) Discriminant adaptive nearest neighbor classification. IEEE Trans. on Pattern Analysis and Machine Intelligence, 18(6):607–616. 32

[113] Hastie, T. Tibshirani, R. and Friedman, J.H. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, New York, 2001. 2, 3, 29, 32, 33, 45

[114] Haykin, S. Neural Networks: A Comprehensive Foundation. Prentice Hall; 2nd edition, New York, 1998. 33

[115] Heinze, S. (2001) Kooperative neuronale Gruppenbildung im primären visuellen Cortex bei der Mustererkennung. Fortschritt-Berichte VDI, Reihe 17, Nr.214. 109

[116] Hellman, M.E. and Raviv, J. (1970) Probability of error, equivocation and the Chernoff bound. IEEE Trans. on Information Theory, 16(4):368–372. 6

[117] Hellman, M.E. (1970) The nearest neighbor classification rule with a reject option. IEEE Trans. on System, Man, and Cybernetics, 6(3):179–185. 4

[118] Hengartner, N.W. (1997) Asymptotic unbiased density estimator. Submitted to Journal of the Royal Statistical Society, Series B. 16

[119] Henze, N. (1988) A multivariate two-sample test based on the number of nearest neighbor type coincidences. The Annals of Statistics, 16(2):772–783. 101

[120] Henze, N. and Penrose, M.D. (1999) On the multivariate runs test. The Annals of Statistics, 27(1):290–298. 101

[121] Hernández-Espinosa, C., Fernández-Redondo, M. and Ortiz-Gómez, M. (2003) Ensemble methods for multilayer feedforward networks. Proceedings of the European Symposium on Artificial Neural Networks. Bruges(Belgium), pp.23-25;261–266. 108

[122] Hochbruck, M., Lubich, C. and Selhofer, H. (1998) Exponential integrators for large systems of differential equations. SIAM Journal on Scientific Computing, 19(5):1552–1574. 54

[123] Hodgkin, A.L. and Huxley, A.F. (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. Journal of Physiology, 117(4):500–544. 53

[124] Hong, Z.-Q. and Yang, J.-Y. (1991) Optimal discriminant plane for a small number of samples and design method of classifier on the plane. Pattern Recognition, 24(4):317–324. 28

[125] Howlett, R.J. and Jain, L.C. Radial Basis Function Networks 1: Recent developments in theory and applications. Sudies in Fuzziness and Soft Computing, vol.66, Physica-Verlag, Heidelberg, 2001. 33

[126] Hughes, G.F. (1968) On the mean accuracy of statistical pattern recognizers. IEEE Trans. on Information Theory, 14(1):55–63. 39

[127] Hupé, J.M., James, A.C., Girard, P., Lomber, S.G., Payne, B.R. and Bullier, J. (2001) Feedback connections act on the early part of the responses in monkey visual cortex. J. Neurophysiol., 85(1):134-145. 94

[128] Inverardi, P.L.N. (2003) MSE-Comparison of some different estimators of entropy. Communications in Statistics – Simulation and Computation, 32(1):17–30. 20, 100

[129] Jackson, J.C. and Redish, A.D. (2003) Detecting dynamical changes within a simulated neural ensemble using a measure of representational quality. Network: Comput. Neural Syst., 14(4):629-645. 7

[130] Jain, A.K. and Chandrasekaran, B. (1982) Dimensionality and sample size considerations in pattern recognition practice. In Handbook of Statistics, vol.2 P.R. Krishnaiah and L.N. Kanal eds. North-Holland, Amsterdam, pp.835–855. 32

[131] Jain, A.K., Duin, R.P.W. and Mao, J. (2000) Statistical pattern recognition: a review. IEEE Trans. on Pattern Recognition and Machine Intelligence, 22(1):4–37. 10, 26

[132] Jain, A. and Zongker, D. (1997) Feature selection: Evaluation, application, and small sample performance. IEEE Trans. Pattern Analysis and Machine Intelligence, 19(2):153-158. 103

[133] Jolliffe, I.T. Principal Component Analysis. Springer-Verlag, New York, 1986. 10

[134] Jones, E.G. (1986) Connectivity of the primate sensory-motor cortex. Cerebral Cortex, 5:113–183. 58

[135] Kandel, E.R., Schwartz, J.H., and Jesse, T.M. Principles of Neural Science. 4th edition, McGraw-Hill, New York, 2000. 49

[136] Kestler, H.A., Schwenker, F. and Palm, G. (2001) RBF network classification of ECGs as a potential marker for sudden cardiac death. Ulmer Informatik-Berichte, no.2001–03. 33, 109

[137] Kestler, H.A. and Schwenker, F. (2001) Classification of High-Resolution ECG Signals. In R.J. Howlett and L.C. Jain, editors, Radial Basis Function Neural Networks 2: New Advances in Design. Physica-Verlag, pp.167–214. 33

[138] Kiss, T. and Érdi, P. (2002) Mesoscopic Neurodynamics. BioSystems, Michael Conrad's special issue, 64(1-3):119–126. 49

[139] Kittler, J. (1986). Feature Selection and Extraction. In Handbook of Pattern Recognition and Image Processing, Eds. Tzay Y. Young, King-Sun Fu. Academic Press, New York, chapter 3, pp.59–83. 103

[140] Kohavi, R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th international Joint Conference on Artifical Intelligence, Morgan Kaufmann, San Mateo, vol.2, pp.1137–1143. 12

[141] Kohavi, R. and John, G.H. (1997) Wrappers for feature selection. Artificial Intelligence, 97(1-2):273–324. 103

[142] Kohonen, T.K. (1990) The self-organizing map. Proceedings of IEEE, 78(9):1464–1480. 10

[143] Koller, D. and Sahami, M. (1996) Toward optimal feature selection. In 13th International Conference on Machine Learning, pp.284–292. 104

[144] Kozachenko, L.F. and Leonenko, N.N. (1987) A statistical estimate for the entropy of random vector. Problems Information Transmission (russian), 23(2):9–16. 20, 100

[145] Kraft, S. and Schmid, F. (2000) Nonparametric tests based on area-statistics. Discussion papers in Statistics and Econometrics, no.2/00, University of Cologne. 24, 74

[146] Kremper, A., Schanze, T. and Eckhorn, R. (2002) Classification of neural signals by a generalized correlation classifier based on radial basis functions. Journal of Neuroscience Methods, 116(2):179–187. 33, 85, 97, 109

[147] Kremper, A. and Eckhorn, R. (2002) Comparison of two methods for dimension reduction applied to neurophysiological data. In Statistical modelling and inference for complex data structures. Louvain-la-Neuve, Belgium. 69, 109

[148] Kremper, A. and Eckhorn, R. (2003) Reduction of high dimensional brain signals by radial basis functions for extracting differences in the small-sample case. Proceedings of the Göttingen Neurobiology Conferene, Thieme-Verlag, pp.1077. 69

[149] Kremper, A. and Eckhorn, R. (2005) Comparison of projection methods for dimension reduction and the determination of lower bounds for transmitted sensory information. Proceedings of the 30th Göttingen Neurobiology Conferene, no.186B. 106

[150] Kremper, A., Gail, A. and Eckhorn, R. (2005) Discrimination and prediction of perceptual states from multiple-electrode recordings in monkey striate cortex. Proceedings of the 30th Göttingen Neurobiology Conferene, no.185B. 96

[151] Kremper, A. and Eckhorn, R. (2005) Lower bounds for the transmitted information in sensory multi-electrode recordings by different projection methods. In preparation. 106

[152] Kurita, T. and Taguchi, T. (2002) A modification of kernel-based Fisher discriminant analysis for face detection. Proceedings of the Fifth IEEE Intern. Conf. on Automatic Face and Gesture Recognition, Washington, D.C., pp.300–305. 31

[153] Lachenbruch, P.A. (1967) An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. Biometrics, 23(4):639–645. 12

[154] Lai, P.L. and Fyfe, C. (2000) Kernel and nonlinear canonical correlation analysis. International Journal of Neural Systems, 10(5):365–377. 31

[155] Landgrebe, D. (1999) Information extraction principles and methods for multispectral and hyperspectral image data. In Information Processing for Remote Sensing, edited by C. H. Chen. World Scientific Publishing, New York, chapter I. 8

[156] Laubach M., Wessberg, J. and Nicolelis, M.A.L. (2000) Cortical ensemble activity increasingly predicts behaviour outcomes during learning of a motor task. Nature, 405(6786):567-571. 7

[157] Leonardis, A. and Bischof, H. (1998) An efficient MDL-based construction of RBF networks. Neural Networks, 11(5):963–973. 108

[158] Liley, D.T.J., Cadusch, P.J. and Dafilis, M.P. (2002) A spatially continuous mean field theory of electrocortical activity. Network: Computation in Neural Systems, 13(1):67–113. 50

[159] Loftsgaarden, D.O. and Quesenberry, C.P. (1965) A nonparametric estimate of a multivariate density function. Annals of Mathematical Statistics, 36(3):1049–1051. 13

[160] Longtin, A., Moss, F. and Bulsara, A. (1991) Time interval sequences in bistable systems and noise induced transmission of neural information. Phys. Rev. Lett., 67:656–659. 53

[161] Manwani, A. and Koch, C. (1999) Detecting and estimating signals in noisy cable structure, I: neuronal noise sources. Neural Computation, 11(8):1797–1829. 53

[162] Markovich, N.M. (2002) Transformed estimates of densities of heavy-tailed distributions and classification. Automation and Remote Control, 63(4):627–640. 15

[163] Martin, J.K. and Hirschberg, D.S. (1996a) Small sample statistics for classification error rates, I: Error rate measurements. Technical report, No.96-21, University of California, Irvine. 12, 48

[164] Martin, J.K. and Hirschberg, D.S. (1996b) Small sample statistics for classification error rates, II: Confidence intervals and significance tests. Technical report, No.96-22, University of California, ICS Dept., Irvine. 12

[165] Mathar, R. Multidimensionale Skalierung : mathematische Grundlagen und algorithmische Aspekte. Teubner-Verlag, Stuttgart, 1997. 91

[166] Mazurek, M.E. and Shadlen, M.N. (2002) Limits to the temporal fidelity of cortical spike rate signals. Nature Neuroscience, 5(5):463–471. 51

[167] McDonough, R.N. and Whalen, A.D. Detection of signals in noise. 2nd ed. Academic Press, San Diego, 1995. 59

[168] Mercer, J. (1909) Functions of positive and negative type, and their connection with the theory of integral equations. Philosophical Transactions of the Royal Society London, A, 209:415–446. 30

[169] Michhelli, C.A. (1986) Interpolation of scattered data: Distance matrices and conditionally positive definite functions. Constructive Approximation, 2(1):11–22. 34

[170] Mika, S., Rätsch, G., Weston, J., Schölkopf, B., and Müller, K.-R. (1999) Fisher discriminant analysis with kernels. In Neural Networks for Signal Processing, volume IX, IEEE Press, New York, pp.41-48. 3, 30, 31, 32

[171] Mika, S., Rätsch, G. Müller, K.-R. (2001) A mathematical programming approach to the kernel Fisher algorithm. Proc. Conf. Neural Information Processing Systems, T.K. Leen, T.G. Dietterich, and V. Tresp, eds., vol.13, MIT Press, pp.591–597. 32

[172] Mika, S., Rätsch, G., Weston, J., Schölkopf, B., Smola, A.J. and Müller, K.-R. (2003) Constructing descriptive and discriminative nonlinear features: Rayleigh coefficients in kernel feature spaces. IEEE Trans. on Pattern Analysis and Machine Intelligence, 25(5):623–633. 30

[173] Miller, G.A. (1955) Note on the bias on information estimates. In H. Quastler (Ed.), Information theory in Psychology; Problems and Methods II-B. Free Press, Glencoe, IL, pp.95-100. 7

[174] Morlini, I. (1999) Radial basis function networks with partially classified data. Ecological Modelling, 120(2-3):109–118. 108

[175] Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K. and Schölkopf, B. (2001) An introduction to kernel-based learning algorithms. IEEE Trans. on Neural Networks, 12(2):181–202. 30

[176] Narendra, P.M. and Fukunaga, K. (1977). A branch and bound algorithm for feature subset selection. IEEE Trans. on Computers, 26(9):917–922. 103

[177] Natschläger, T. and Maass, W. (2004) Information dynamics and emergent computation in recurrent circuits of spiking neurons. In S. Thrun, L. Saul, and B. Schölkopf, editors, Proc. of NIPS 2003, Advances in Neural Information Processing Systems. MIT Press, Cambridge, volume 16, pp.1255–1262. 109

[178] Neumann, J., Schnörr, C. and Steidl, G. (2004) SVM-based feature selection by direct objective minimization. In Pattern recognition eds. Rasmussen, Bülthoff, Giese, Schölkopf, Springer-Verlag, Berlin. 104

[179] Nicholls, J.G., Martin, A.R. and Wallace, B.G. From neuron to brain. Sinauer Associates, Inc., Third edition, 1992. 49

[180] Oeser, E. Geschichte der Hirnforschung: Von der Antike bis zur Gegenwart. Primus Verlag, 2002. 49

[181] Optican, L.M. and Richmond, B.J. (1987) Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. III. information theoretic analysis. Journal of Neurophysiology, 57(1):162–178. 5

[182] Orr, M., Hallam, J., Takezawa, K., Murray, A., Ninomiya, S., Oide, M. and Leonard, T. (2000) Combining regression trees and radial basis function networks. International Journal of Neural Systems, 10(6):453–465. 108

[183] Paninski, L. (2003). Estimation of entropy and mutual information. Neural Computation, 15(6):1191–1254. 10, 109

[184] Park, B.U., Jeong, S.-O., Jones, M.C. and Kang, K.-H. (2003) Adaptive variable location kernel density estimators with good performance at boundaries. Nonparametric Statistics, 15(1):61–75. 19

[185] Parzen, E. (1963) On the estimation of a probability density function and the mode. Annals of Mathematical Statistics, 33(3):1065–1076. 15

[186] Perkel, D.H. and Bullock, T.H. (1968) Neural coding. Neurosciences Research Program Bulletin, 6:221-348. 51

[187] Perkins, S., Lacker, K. and Theiler, J. (2003) Grafting: Fast, incremental feature selection by gradient descent in function space. Journal of Machine Learning Research, 3:1333–1356. 104

[188] Perrone, M.P. and Cooper, L.N. (1993) When networks disagree: Ensemble methods for hybrid neural networks. In Neural Networks for Speech and Image Processing, R. J. Mammone, Eds. Chapman & Hall, London, pp.126-142. 108

[189] Polonik, W. (1999) Concentration and goodness-of-fit in higher dimensions: (asymptotically) distribution-free methods. The Annals of Statistics, 27(4):1210–1229. 101

[190] Powell, M.J.D The theory of radial basis function approximation in 1990. In: Light, W. editor. Advances in numerical analysis II: wavelets, subdivision algorithms and radial functions. Oxford University Press, pp.105–210. 33, 34

[191] Principe, J., Xu, D. and Fisher, J. (2000) Information theoretic learning. In S. Haykin (ed.) Unsupervised adaptive filtering. John Wiley, New York, pp.265–319. 7

[192] Provost, F. and Fawcett, T. (2001) Robust classification for imprecise environments. Machine Learning, 42(3):203-231. 23

[193] Quinlan J.R. (1986) Induction of decision trees. Machine Learning, 1(1):81-106. 104

[194] Quinlan, J.R. Programs for Machine Learning. Morgan Kaufmann, San Marteo, CA, 1993. 104

[195] Raudys, S.J. and Jain, A.K. (1991) Small sample size effects in statistical pattern recognition: Recommendations for practitioners. IEEE Trans. on Pattern Analysis and Machine Intelligence, 13(3):252–264. 11, 25

[196] Raudys, S. and Duin, R.P.W. (1998) Expected classification error of the Fisher linear classifier with pseudo-inverse covariance matrix. Pattern Recognition Letters, 19(5-6):385–392. 27

[197] Renyi, A. Probability Theory. Elsevier Publishing Company, Inc., New York, 1970. 19

[198] Rieke, F., Warland, D., van Steveninck, R.R. and Bialek, W. Spikes: Exploring the Neural Code. MIT Press, Cambridge, MA., 1997. 5

[199] Rolls, E.T., Treves, A., Tovee, M.J. and Panzeri, S. (1997) Information in the neuronal representation of individual stimuli in the primate visual cortex. Journal of Computational Neuroscience, 4(4):309–333. 9

[200] Rosenblatt, M. (1956) Remarks on some nonparametric estimates of a density function. Annals of Mathematical Statistics, 27:832–837. 15

[201] Roth, V. and Steinhage, V. (2000) Nonlinear discriminant analysis using kernel functions. In S.A. Solla, T.K. Leen and K.-R. Müller, eds, Proc. Conf. Advances in Neural Information Processing Systems. MIT Press, Cambridge, MA, vol.12, pp.568–574. 30

[202] Rotter, S. and Diesmann, M. (1999) Exact digital simulation of time-invariant linear systems with applications to neuronal modeling. Biological Cybernetics, 81(5-6):381–402. 58

[203] Roweis, S.T. and Saul, L.K. (2000) Nonlinear dimensionality reduction by locally linear embedding. Science, 290(5500):2323–2326. 10

[204] Rudolph, M. and Destexhe, A. (2003) Characterization of subthreshold voltage fluctuations in neuronal membranes. Neural Computation, 15(11):2577–2618. 53

[205] Sachs, L. Angewandte Statistik. Springer-Verlag, Heidelberg, 2002. 101

[206] Sain, S.R. and Scott, D.W. (2002). Zero-Bias bandwidths for locally adaptive kernel density estimation. Scandinavian Journal of Statistics, 29:441–460. 18

[207] Salin, P.A. and Bullier, J. (1995) Corticocortical connections in the visual system: structure and function. Physiol. Rev., 75(1):107–154. 94

[208] Sammon, J.W. (1969) A non-linear mapping for data structure analysis. IEEE Trans. on Computers, 18(5):401–409. 91

[209] Saunders, C., Gammerman, A. and Vovk, V. (1998) Ridge regression learning algorithm in dual variables. Proc. 15th Internation Conference on Machine Learning. Morgan Kaufmann Publishers, San Francisco, pp.515–521. 31

[210] Schaffer, C. (1993) Selecting a classification method by cross-validation. Machine Learning, 13(1):135–143. 12

[211] Schapire, R.E. (2002) The boosting approach to machine learning: An overview. MSRI workshop on nonlinear estimation and classification. Berkeley, CA. 108

[212] Schmid, F. and Trede, M.A. (1995) A distribution free test for the two sample problem for general alternatives. Computational Statistics & Data Analysis, 20(4):409–419. 23, 24, 74

[213] Schölkopf, B., Smola, A.J. and Müller, K.-R. (1998) Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, 10(5):1299–1319. 31

[214] Schölkopf, B. and Smola, A.J. Learning with Kernels. MIT Press, 2002. 30, 31, 107

[215] Schölkopf, B., Sung, K., Burges, C., Girosi, F., Niyogi, P., Poggio, T. and Vapnik, V. (1996) Comparing support vector machines with gaussian kernels to radial basis function classifiers. Technical Report, A.I. Memo No.1599, MIT, Cambridge, MA. 107

[216] Schreiber, T. (1999) Interdisciplinary application of nonlinear time series methods. Physics Reports, 308(1):1–64. 105

[217] Schwenker, F., Kestler, H.A. and Palm, G. (2001) Three learning phases for radial-basis-function networks. Neural Networks, 14(4-5):439–458. 34, 108

[218] Scott, D.W. (1979) On optimal and data-based histograms. Biometrika, 66(3):605–610. 13

[219] Scott, D.W. Multivariate density estimation: Theory, Practice, and Visualization. John wiley & Sons, New York, 1992. 12, 13, 15, 100

[220] Shannon, C.E. (1948) A mathematical theory of communication. Bell System Technical Journal, 27:379–423;623–656. 5

[221] Shannon, C.E. and Weaver, W. A mathematical theory of communication. Board of Trustees of the university of Illinois, Urbana, 1949. 5

[222] Shashua, A. (1999) On the relationship between the support vector machine for classification and sparsified Fisher's Linear Discriminant. Neural Processing Letters, 9(2):129–139. 28

[223] Shawe-Taylor, J. and Cristianini, N. Kernel Methods for Pattern Analysis. Cambridge University Press, 2004. 31, 45

[224] Sheather, S.J. and Jones, M.C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society, Ser. B, 53(3):683–690. 16

[225] Sheinberg, D.L. and Logothetis, N. K. (1997) The role of temporal cortical areas in perceptual organization. Proc. Natl. Acad. Sci. USA, 94(7):3408–3413. 94

[226] Silverman, B.W. Density estimation for statistics and data analysis. Monographs on Statistics and Applied Probability. Chapman and Hall, London, 1986. 15, 18

[227] Skurichina, M. and Duin, R.P.W. (2002) Bagging, boosting and the random subspace method for linear classifiers. Pattern Analysis & Applications, 5(2):121–135. 48, 108

[228] Stein, R. B. (1967). The frequency of nerve action potential generated by applied currents. Proc. R. Soc. London, Ser. B, 167(6):64–86. 50

[229] Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society, Ser. B, 36(2):111–147. 12

[230] Stone, M. (1977) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. Journal of the Royal Statistical Society, Ser. B, 39(1):44–47. 12

[231] Stone, M. (1978) Cross-validation: A review. Mathematics, Operation Research and Statistics, 9(1):127–139. 12

[232] Stoppiglia, H., Dreyfus, G., Dubois, R. and Oussar, Y. (2003) Ranking a random feature for variable and feature selection. The Journal of Machine Learning Research, 3:1399–1414. 105

[233] Strehmel, K. and Weiner, R. Numerik gewöhnlicher Differentialgleichungen. B.G. Teubner, Stuttgart, 1995. 58

[234] Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, B. and Vanderwalle, J. Least square support vector machines. World Scientific Publishing, 2002. 32

[235] Szentagothai, J. (1978) The Ferrier Lecture, 1977. The neuron network of the cerebral cortex: a functional interpretation. Proc. R. Soc. Lond. B: Biol. Sci., 201(1144):219-248. 49

[236] Tenenbaum, J.B., de Silva, V. and Langford, J.C. (2000) A global geometric framework for nonlinear dimensionality reduction. Science, 290(5500):2319–2323. 10, 100

[237] Terrell, G.R. (1990) The maximal smoothing principle in density estimation. Journal of the American Statistical Association, 85:470–477. 15

[238] Terrell, G.R. and Scott, D.W. (1985) Oversmoothed nonparametric density estimates. Journal of the American Statistical Association, 80:209–214. 15

[239] Terrell, G.R. and Scott, D.W. (1992) Variable kernel density estimation. The Annals of Statistics, 20:1236–1265. 19

[240] Thawonmas, R. and Abe, S. (1997) A novel approach to feature selection based on analysis of class regions. IEEE Trans. on Systems, Man, and Cybernetics, Part B, 27(2):196-207. 103

[241] Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., Farmer, J.D. (1992) Testing for nonlinearity in time series: the method of surrogate data. Physica D, 58(1-4):77-94. 105

[242] Thomson, A.M. and Bannister, A.P. (1998) Postsynaptic pyramidal target selection by descending layer III pyramidal axons: dual intracellular recordings and biocytin filling in sclices of rat neocortex. Neuroscience, 84(3):669–683. 50, 52

[243] Thomson, A.M. and Bannister, A.P. (2003) Interlaminar Connections in the Neocortex. Cerebral Cortex, 13(1):5–14. 58

[244] Tibshirani R.J. (1994) Regression shrinkage and selection via the Lasso. Technical report, Dept. of Statistics, University of Toronto. 104

[245] Torkkola, K. (2003) Feature extraction by non-parametric mutual information maximization. Journal of Machine Learning Research, 3:1415–1438. 6, 103

[246] Torkkola, K. and Campbell, W.M. (2000) Mutual information in learning feature transformations. Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann, pp.1015–1022. 3, 6, 108

[247] Treves, A. and Panzeri, S. (1995) The upward bias in measures of information derived from limited data samples. Neural Computation, 7(2):399-407. 7

[248] Truccolo, W.A., Ding, M. and Bressler, S.L. (2001) Variability and interdependence of local field potentials: effects of gain modulation and nonstationarity. Neurocomputing, 38-40:983–992. 9

[249] Tubbs, J.D., Coberly, W.A. and Young, D.M. (1982) Linear dimension reduction and Bayes classification with unknown population parameters. Pattern Recognition, 15(3):167–172. 4

[250] Van Gestel, T., Suykens, J.A., Lanckriet, G., Lambrechts, A., De Moor, B. and Vandewalle, J. (2001) Bayesian framework for least-square support vector machine classifiers, Gaussian processes and kernel Fisher discriminant analysis. Neural Computation, 14(5):1115–1147. 32

[251] Vapnik, V.N. The Nature of Statistical Learning Theory. Springer-Verlag, New York, 1995. 30, 107

[252] Vapnik, V.N. Statistical Learning Theory. John Wiley & Sons, New York, 1998. 31, 107

[253] Vasconcelos, N. (2003) Feature selection by maximum marginal diversity: optimality and implications for visual recognition. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Madison, Wisconsin, volume 1, pp.762–769. 6

[254] Vasicek, O. (1976) A test for normality based on sample entropy. J. Royal Statistical Society, Ser. B, 38:54–59. 20

[255] Verdugo Lazo, A.C.G. and Rathie, P.N. (1978) On the entropy of continuous probability distributions. IEEE Trans. on Information Theory, 24(1):120–122. 20

[256] Verleysen, M., Francois, D., Simon, G. and Wertz, V. (2003) On the effects of dimensionality on data analysis with neural networks. In Artificial Neural Nets Problem solving methods. Lecture Notes in Computer Science 2687 edited by J. Mira, J.R. Alvarez eds., Springer-Verlag, pp.105–112. 8

[257] Veropoulos, K., Campbell, C. and Cristianini, N. (1999) Controlling the sensitivity of support vector machines. Proceedings of the International Joint Conference on Artifical Intelligence, Stockholm, pp.55–60. 23

[258] Verwer, J.G., Spee, E.J., Blom, J.G. and Hundsdorfer, W. (1999) A second order Rosenbrock method applied to photochemical dispersion problems. SIAM Journal on Scientific Computing, 20(4):1456–1480. 54

[259] Victor, J.D. (2002) Binless strategies for estimation of information from neural data. Physical Review E, 66(1):1–15. 20

[260] Vlassis, N., Motomura, Y. and Kröse, B. (2001) Supervised dimension reduction of intrinsically low-dimensional data. Neural Computation, 14(1):191-215. 9

[261] Wand, M.P. and Jones, M.C. Kernel smoothing. Chapman and Hall, London, 1995. 15

[262] Webb, A.R. Statistical Pattern Recognition. 2.ed., John Wiley & Sons, England, 2003. 2, 27, 32, 91

[263] Weiss, S.M. and Kulikowski, C.A. Computer systems that learn. Morgan Kaufmann, San Mateo, CA, 1991. 11

[264] Weston, J., Elisseeff, A., Schöelkopf, B., and Tipping, M. (2003) Use of the zero norm with linear models and kernel methods. Journal of Machine Learning Research, 3:1439–1461. 104

[265] White, E.L. Cortical Circuits. Birkhauser, Boston, MA, 1989. 49

[266] Wieczorkowski, R. and Grzegorzewski, P. (1999) Entropy estimators – improvements and comparisons. Communications in Statistics – Simulation and Computation, 28(2):541–567. 20

[267] Wilson, H.R., and Cowan, J.D. (1972) Excitatory and inhibitory interactions in localized populations of model neurons. Biophysical Journal, 12(1):1–24. 50, 58, 65

[268] Wolpert, D.H. (1995) The relationship between PAC, the statistical physics framework, the Bayesian framework, and the VC framework. In D.H. Wolpert, ed., The Mathematics of Generalization. Addison Wesley, pp.117–214. 12

[269] Yang, J. and Honavar, V. (1997) Feature subset selection using a genetic algorithm. In Proceedings of the International Conference on Genetic Programming. Stanford, CA., pp.380–385. 103

[270] Yoshioka, T., Blasdel, G.G., Levitt, J.B. and Lund, J.S. (1996) Relation between patterns of intrinsic lateral connectivity, ocular dominance, and cytochrome oxidase-reactive regions in macaque monkey striate cortex. Cerebral Cortex, 6(2):297–310. 96

[271] Zhang, G.P. (2000) Neural networks for classification: a survey. IEEE Trans. on Systems, Man, and Cybernetics -part C: Applications and Reviews, 30(4):451–462. 108

[272] Zhang, K., Ginzburg, I., McNaughton, B.L. and Sejnowski, T.J. (1998) Interpreting neuronal population activity by reconstruction: unified framework with application to hippocampal place cells. J. Neurophysiol., 79(2):1017-1044. 7

[273] Zhu, J., Rosset, S., Hastie, T. and Tibshirani, R. (2003) 1-norm support vector machines. In S. Thurn, L. Saul and B. Schölkopf, editors, Advances in Neural Information Processing Systems. MIT Press, Cambridge MA, USA. 104

Abstract

Estimating the transmitted sensory information from neural signals is often made difficult by the low number of stimulus-response pairs obtained during an experiment with awake animals or humans and the high dimensionality of the measured multi-channel response space. Classical approaches, analyzing the signals from each recording site separately or an averaging of the different time series, make use of the spatio-temporal correlations in an unsatisfactory way. Therefore, I have developed a method to take the spatio-temporal statistical dependencies of all recorded channels simultaneously into account. Instead of estimating the underlying probability density functions of stimuli and response signals, I investigated different projection methods generating simpler, lower-dimensional representations of the data. In addition to two linear approaches, a simple correlation classifier and Fisher's discriminant method, I implemented methods based on nearest neighbors search strategies, regularization theory, kernel machines, and radial basis functions. None of the methods requires more than the solution of a linear system of equations.

At first, I did extensive studies regarding the distance metric, the approximation kernels and the regularization parameters. To quantify the reduction of information after dimension reduction I estimated the information loss by k-fold cross validation in combination with Monte-Carlo sampling for various sets of artificial data. Emphasis was placed on the relationship between the training size and the dimensionality. One result from these studies was that increasing the number of irrelevant signal components has a negative effect on the transmitted information for all methods. In general, the performance of the various methods was related to the type of data, in particular to their signal correlations. As expected, none of the methods performed optimally under all conditions (e.g., in certain situations linear prediction can be quite accurate).

The performance for more realistic data was examined by discriminating signals from small simulated neural networks. The network signals were generated based on a mean field model, superimposed with internal uncorrelated colored noise processes. A similar dependence on the signal-to-noise ratio, was found for all methods. The estimated information decreases monotonely with increasing signal-to-noise ratio. Interestingly, the performance of the radial basis function approach outperformed all other approaches within the neural network signals, significantly.

The possibilities and benefits of this new approach were tested on multi-channel micro-electrode population recordings from monkey primary visual cortex during a matching-to-sample experiment. The dimension reduction approach was used to predict different perceptual states, evoked by ambiguous visual stimulation. Stimuli consisted of perpendicularly oriented gratings presented singly to each eye resulting in rivaling percepts. The following discriminant analysis was based on local field potentials (LFP) and multi-unit activity (MUA).

In contrast to previous prediction studies, which used disjoint subsets of the same data for learning and discrimination, I also combine different sets of data. Responses to congruent stimuli (identical in both eyes) were used for training of the algorithm but testing was done with responses to incongruent (rivalrous) stimuli. Based on simultaneous recordings from 11 channels trained on local field potentials in the congruent condition and tested by the rivalrous condition, Bayes error decreased in one monkey up to 23 % percent near response time. In contrast prediction was near chance, with single-channel analysis of the same data.

Putting all these results in a nutshell, I am convinced that this algorithm is well suited to study other multi-channel intra-cortical recordings. The multi-channel approach enables a better discrimination or at least the same discrimination compared to single-electrode approaches, because it allows to take the spatio-temporal dependency into account. In contrast to classical approaches, this approach makes it possible to quantify the information augmentation due to the increase in the electrode number without any assumption about the statistical dependence. This approach is not restricted in dimensionality and might be used to analyze other electro-physiological multi-channel recordings. The generalization to multi-class problems is straight forward and the extraction of relevant features describing the underlying dynamical processes can be done in an efficient manner.

Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit der Entwicklung und Anwendung von Dimensionsreduktionsverfahren zur Untersuchung von neuronalen Multikanal-Signalen. Insgesamt wurden sechs Verfahren aus dem Bereich des überwachten maschinellen Lernens auf ihre Anwendbarkeit bei kortikalen Multikanal-Daten hin untersucht. Neben zwei linearen Methoden, einem *Korrelationsklassifikator* und der *Fisher-Diskriminanz-Methode*, wurden vier nichtlineare Verfahren aus dem Bereich der *Nächsten-Nachbarn*, der *Kern-Maschinen*, der *radialen Basisfunktionen* und der *Regularisierungstheorie* implementiert.

Die Motivation für diesen Ansatz lag in dem zunehmenden Einsatz von Multikanal-Mikro-Elektroden, deren Daten mit klassischen Einzelkanal- oder paarweisen Korrelationsverfahren nur unzureichend analysiert werden können. Schwerpunkt dieser Arbeit war es daher herauszufinden, inwieweit die einzelnen Verfahren in der Lage sind die raum-zeitlichen, statistischen Abhängigkeiten gleichzeitig aufgenommener neuronaler Signale zu berücksichtigen und die übertragene sensorische Information aus den neuronalen Signalen verlässlich abzuschätzen. Ein weiteres Ziel war es eine niedrigdimensionale und damit interpretierbare Repräsentation der hochdimensionalen Multikanal-Signale zu ermöglichen. All diese Fragestellungen wurden im Rahmen des *Zwei-Klassen-Problems* studiert, das heißt, unter der Voraussetzung, dass die Klassenzugehörigkeit der einzelnen Messdaten bekannt ist.

Zunächst wurden umfangreiche Tests zur Wahl der Abstandsmetrik, der Approximationskerne und der Regularisierungsparameter durchgeführt. Um eine Aussage über die Zuverlässigkeit der gewonnenen Ergebnisse zu erhalten wurden die Ergebnisse der sechs Verfahren zusätzlich mit statistischen Verfahren aus der Signal-Erkennungstheorie und einer mehrfachen *Kreuz-Validierung* kombiniert. Das Verhalten der einzelnen Methoden wurde vor allem im Hinblick auf ihre Anwendbarkeit in hochdimensionalen Räumen bei gleichzeitig geringem Stichprobenumfang hin untersucht. Desweiteren wurde getestet, inwieweit die Ergebnisse der einzelnen Ansätze durch Rauschen oder Ausreißer beeinflusst werden.

Das Verhalten der einzelnen Verfahren wurde in einem ersten Schritt an künstlich erzeugten Daten erprobt. Zur Quantifizierung wurde der theoretische *Bayes-Fehler* sowie die *TransInformation nach Shannon* bestimmt und anschließend mit den Ergebnissen nach Anwendung der einzelnen Projektionsverfahren verglichen. Dabei zeigte sich, dass die einzelnen Methoden eine sehr unterschiedliche Abhängigkeit von der Dimensionalität der Daten und der jeweiligen Signalkorrelation besitzen. Die Verdopplung des Stichprobenumfanges hatte teilweise einen deutlichen Einfluss auf den berechneten Bayes-Fehler. Während je nach Datenzusammensetzung die einzelnen Verfahren einen unterschiedlichen Informationsverlust aufwiesen, zeigte sich, dass das auf den radialen Basisfunktionen basierende Verfahren insgesamt ein sehr robustes Verhalten besitzt.

Hinsichtlich der Anwendung auf neuronale Signale wurden die einzelnen Projektionsverfahren zusätzlich zur Trennung simulierter neuronaler Netzwerkdaten eingesetzt. Die Netzwerkdaten wurden mit einem kontinuierlichen Feldansatz mit internen farbigen, unkorrelierten Rauschprozessen erzeugt. Untersucht wurde, inwieweit Rückkopplungsmechanismen sowie die Überlagerung verschiedener Rauschamplituden die Ergebnisse der einzelnen Verfahren beeinflussen. Interessanterweise zeigte sich, dass in allen untersuchten Fällen die Schätzung der übertragenen Information aufgrund des radialen Basisfunktionsansatzes am größten war und die Werte der anderen Verfahren signifikant übertraf.

Das Potential sowie die Zuverlässigkeit dieses Ansatzes für die Analyse neuronaler Multikanal-Signale wurde schließlich an verschiedenen kortikal aufgenommenen Daten getestet. Die Daten wurde mithilfe von Mikro-Elektroden aus dem primären visuellen Kortex wacher Rhesusaffen abgeleitet. Bei den Signalen handelte es sich um lokale Gruppen-Impulswahrscheinlichkeitsdichten (multiple unit-activity, MUA) und lokale Feldpotentiale (LFP). Dabei wurde unter anderem untersucht, inwieweit verschiedene perzeptuelle Zustände, hervorgerufen durch mehrdeutige visuelle Reize, anhand der neuronalen Signale unterschieden werden können. Die visuellen Reize bestanden aus senkrecht bzw. waagerecht orientierten zweidimensionalen Gabor-Wavelets, die jeweils separat dem linken und rechten Auge dargeboten wurden und im Falle einer nicht deckungsgleichen Stimulation zu einer rivalisierenden Wahrnehmung führten. Mithilfe des neu entwickelten Multikanal-Analyse-Ansatzes war es möglich einen eindeutigen Zusammenhang zwischen der Wahrnehmung und der visuellen Stimulation herzustellen. Basierend auf den mit 11 Mikro-Elektroden gleichzeitig aufgezeichneten lokalen Feldpotentialen, konnte bei einem Versuchstier unter Anwendung des radialen Basisfunktionsansatzes eine Vorhersage der Signale bei nicht deckungsgleicher Stimulation anhand der Signale bei deckungsgleicher Stimulation mit einem Bayes-Fehler von 23 % erzielt werden. Demgegenüber erbrachte eine Einzelkanal-Analyse derselben Daten keine signifikanten Unterschiede.

Aufgrund der erzielten Resultate ist daher davon auszugehen, dass dieser neu entwickelte Ansatz dazu geeignet ist, bei ähnlichen Fragestellungen herauszufinden, inwieweit die neuronale Aktivität mit bestimmten visuellen Reiz- oder Wahrnehmungssituationen in Verbindung steht. Ein wesentlicher Vorteil dieses Ansatzes gegenüber Einzelkanal- oder paarweisen Analysemethoden ist die Berücksichtigung der raum-zeitlichen Kopplungsstruktur von Multikanal-Signalen sowie die kompakte niedrigdimensionale Repräsentation der hochdimensionalen Daten. Eine prinzipielle Beschränkung in der Dimensionalität der Daten besteht nicht. Überdies ist davon auszugehen, dass dieser Dimensionsreduktionsansatz auch bei anderen elektrophysiologischen Multikanal-Aufnahmen und anderen sensorischen Signalen gewinnbringend eingesetzt werden kann. Zusätzlich ist eine Anwendung des Algorithmus auf Mehr-Klassen-Probleme und zur Merkmalsextraktion möglich, um zum Beispiel nähere Informationen über die kooperativen dynamischen Prozesse in der neuronalen Vernetzung zu gewinnen oder aber, um in intakten und durch Krankheit veränderten Hirnprozessen charakteristische Merkmale und Zusammenhänge aufzudecken.

Danksagung

Ein Sprichwort sagt: "Wenn einer eine Reise tut, dann kann er was erleben". Auf meiner Reise in die faszinierende Welt der neuronalen Prozesse habe ich einiges erlebt und dabei viele Menschen kennengelernt. Einigen davon möchte ich an dieser Stelle meinen ganz herzlichen Dank aussprechen.

Zunächst möchte ich mich bei meinem Betreuer Prof. Dr. Reinhard Eckhorn bedanken. In zahlreichen Gesprächen hatte ich Gelegenheit von seinem umfangreichen Wissensschatz zu profitieren und einen tieferen Einblick in die Funktionsweise des Gehirns zu erhalten. Weiterhin danke ich den Professoren Prof. Dr. Bernhard Schmitt, Prof. Dr. Martin Buhmann und Prof. Dr. Stephan Dahlke für Ihre Unterstützung bei mathematischen Fragen. Diese Arbeit entstand dank der Unterstützung von Seiten der Konrad-Adenauer-Stiftung e.V., für die ich mich ebenfalls ganz herzlich bedanke.

Den Kollegen Sabine Heinze, Marcus Eger und Alexander Gail danke ich für die Bereitstellung von Messdaten, sowie Basim Al-Shaikhli, Frank Michler und Tim Zwickel für die anregenden Diskussionen zur Simulation neuronaler Netzwerke. Andreas Gabriel und Andreas Bruns danke ich für die zahlreichen Anregungen und Ihre Unterstützung bei Fragen der Datenanalyse. Frau Sigrid Thomas danke ich für Ihre Hilfe bei allen formalen Angelegenheiten.

Meiner Frau und meiner Familie danke ich für Ihre Unterstützung und Geduld.

Widmen möchte ich diese Arbeit

Herrn Hans Jörg Brinksmeyer

und

Frau Dr. Daniela Tandecki

Wissenschaftlicher Werdegang

Persönliche Daten

Name: Helmut Alexander Kremper
Anschrift: Cappeler Str. 2a, 35085 Ebsdorfergrund
geboren am: 14.10.1970 in Lahnstein

Schulbildung

08/83 - 05/90	Wilhelm-Hofmann-Gymnasium, St. Goarshausen Abschluß: Abitur

Studium

09/91 - 07/94	Studium der Physik an der Rheinischen Friedrich-Wilhelms-Universität Bonn
10/93	Erwerb des Vordiploms in Physik
08/94 - 12/98	Studium der Physik an der Philipps-Universität Marburg
12/98	Erwerb des Diploms in Physik Titel der Diplomarbeit: *Biologienahe Modellierung der elektrischen Stimulation neuronaler Strukturen*

Hochschultätigkeit

01/99 - 09/00	Mitarbeit an einem Konzept zur Errichtung eines Sonderforschungsbereiches: *Dimensionsreduktion von komplexen Systemen* an den Universitäten Marburg und Giessen
10/00 - 01/05	Mitarbeit am Fachbereich Physik in der Arbeitsgruppe Neurophysik der Philipps-Universität Marburg

Derzeitiges Arbeitsverhältnis

02/05 - heute	Systementwickler bei Battenberg Robotic